Groupement des Acousticiens de Langue Française

10èmes JOURNEES D'ETUDES SUR LA PAROLE

DU GROUPE

COMMUNICATION PARLEE

Textes des exposés

GRENOBLE - 30 Mai - 1er Juin 1979

Groupement des Acousticiens de Langue Française

10èmes JOURNEES D'ETUDES SUR LA PAROLE

DU GROUPE

COMMUNICATION PARLEE

Textes des exposés

GRENOBLE - 30 Mai - 1er Juin 1979

Les thèmes de travail des 10èmes Journées étaient les suivants :

I. LA SYNTHESE DE LA PAROLE

(a) Méthodes et applications

Président : P. LORAND

Animateurs : J. LE ROUX, X. RODET

(b) Applications de la synthèse, modèles articulatoires

Président : P. SIMON

Animateurs : S. MAEDA, R. DESCOUT

(c) Synthèse à partir de textes

Président : P. MARTIN

Animateurs : F. EMERARD, J.S. LIENARD

II. LA FORMALISATION DU LEXIQUE, DE LA PHONOLOGIE EN VUE DE L'APPLICATION A LA RECONNAISSANCE ET A LA SYNTHESE DE LA PAROLE

Président : G. MERCIER

Animateurs : G. PERENNOU, M. ROSSI.

TABLE DES MATIERES

THEME I (a)

SYNTHESE DE LA PAROLE

Méthodes

10èmes JOURNÉES D'ÉTUDE SUR LA PAROLE

GRENOBLE - 30 MAI - 1ER JUIN 1979

L'ICOPHONE LOGICIEL : UN SYNTHETISEUR PAR FORMES D'ONDES

Vito ASTA - Jean-Sylvain LIENARD

Laboratoire d'Informatique pour la Mécanique et
les Sciences de l'Ingénieur (L.I.M.S.I. - C.N.R.S.)
B.P. 30 - 91406 ORSAY Cedex

RESUME

L'ICOLOG (Icophone Logiciel) est un synthétiseur de parole constituant une version purement logiciel du système ICOPHONE V.

Chaque période de l'onde finale est construite point par point par sommation de formes d'ondes partielles conservées en mémoire, qui correspondent aux oscillateurs de l'ICOPHONE V et à leur mise en action. L'organe de sortie est alors un simple convertisseur numérique-analogique ; on peut utiliser tel quel le dictionnaire de diphonèmes de l'ICOPHONE V et modifier la mélodie du signal synthétisé.

Une version sur microprocesseur en temps réel de ce programme est actuellement en cours de réalisation.

"ICOPHONE LOGICIEL", an all-software speech waveform synthesizer

Vito ASTA - Jean-Sylvain LIENARD

L.I.M.S.I., B.P. 30, 91406 ORSAY Cedex

SUMMARY

ICOLOG (Icophone Logiciel) is a speech synthesizer designed as an all-software version of the ICOPHONE V system.

The same methodology was chosen, i.e. diphone synthesis with events coded into 44 1-bit words, but the oscillators bank is realized with digital techniques. Numerous are the advantages of this solution, essentially because of enhanced flexibility and pitch-capability.

The compatibility between the two systems permits to use the whole diphone dictionary without any modification.

ICOLOG first complete version is an almost entirely FORTRAN written program, running on an IBM 370/168 computer connected to a digital-to-analog converter via a System 7 minicomputer. The table look-up method, widely adopted in digital music synthesizers and programs, was found to be the most economical and versatile method for the oscillator bank implementation.

The synthesized signal multiplication by an envelope of adequate shape makes it possible to vary the pitch, reconstituting it in the time-domain, thus avoiding frequency - modulation of the oscillators (which would cause unwanted shifts of the formant - frequencies). Oscillator waveshapes are stored already enveloped, so that no multiplication is required at run-time. Modified Kaiser windows proved to be effective for the necessity to leave the diphone patterns undistorted with a minimum window-length.

A microprocessor-based version of the system is actually being realized, where a hardware version of the table look-up oscillators is used. The extreme simplicity of the ICOPHONE synthesis procedure makes it possible to implement the system in real-time on an 8-bit microprocessor, with very little hardware added.

10èmes JOURNÉES D'ÉTUDE SUR LA PAROLE
GRENOBLE - 30 MAI - 1ER JUIN 1979

L'ICOPHONE LOGICIEL : UN SYNTHETISEUR PAR FORMES D'ONDES

Vito ASTA - Jean-Sylvain LIENARD

L.I.M.S.I., B.P. 30, 91406 ORSAY Cedex

INTRODUCTION

L'activité du L.I.M.S.I. en matière de parole a eu comme point de départ la commande numérique du synthétiseur ICOPHONE II mis au point, sous sa forme à commande optique, au Laboratoire d'Acoustique de l'Université Paris VI, (LEIPP, E. et coll., 1967). En 1969 l'ICOPHONE III, connecté à l'IBM 1130 du L.I.M.S.I., était déjà capable de transformer un texte écrit en une parole intelligible, mais sans intonation. L'ICOPHONE IV tentait de remédier à ce défaut, en offrant la possibilité de multiplier toutes les fréquences d'oscillateurs par un même facteur, mais il faut bien reconnaître que la voix "anamorphosée"de cette manière avait un timbre étrange, sauf si l'on compensait par programme l'évolution de la mélodie de façon à conserver une même échelle formantique. L'ICOPHONE V, appareil autonome, toujours en pleine santé, n'avait pas de prétentions mélodiques. L'ICOPHONE VI, un mutant, adoptait la synthèse à formants, mais exigeait la mise au point d'un nouveau dictionnaire de diphonèmes. Avec l'ICOPHONE LOGICIEL nous revenons aux sources, en utilisant la technique de reconstruction point par point de la forme d'onde.

Cette technique, utilisée initialement en musique synthétique (MATHEWS, M.V., 1969), a déjà été appliquée dans le domaine de la parole : le vocoder CIPHON (BOURGENOT, J.S. et DECHAUX, C., 1975), le synthétiseur SARA (RODET, X., 1977) en sont deux exemples. Elle semble promise à un grand avenir, car elle va dans le sens de l'évolution technologique actuelle : les microprocesseurs, les convertisseurs, les méthodes de traitement numérique du signal permettent d'envisager des réalisations souples, puissantes, bon marché, facilement reproductibles. A ces avantages s'ajoutent, dans le cas de l'ICOPHONE LOGICIEL, la possibilité de conserver le dictionnaire de diphonèmes et la logique de commande de l'ICOPHONE V dont l'intelligibilité est éprouvée, et la possibilité de faire varier la fréquence fondamentale sans modifier l'échelle formantique.

DESCRIPTION DU SYSTEME

Le but initial du présent travail était d'écrire un programme réalisant une version purement logicielle du système de synthèse ICOPHONE V (TEIL, D., 1975). Le programme suit donc la même philosophie de synthèse de ce dernier (synthèse par diphonèmes, dont les événements sont codés par 44 mots de 1 bit), en utilisant aussi certains de ses programmes, mais se différencie de celui-ci dans la phase de synthèse du signal proprement dite : dans l'ICOPHONE V, elle est réalisée par un banc de 44 oscillateurs analogiques sinusoïdaux, de fréquences fixes, commandés en tout ou rien (amplitude codée à 1 bit) ; dans ICOLOG les oscillateurs sont eux-mêmes réalisés numériquement. Cela introduit de nombreux avantages, surtout en ce qui concerne la flexibilité du système : on peut souligner en particulier

- la possibilité d'effectuer des variations temporaires pour expérimenter directement différentes solutions possibles, ou tester l'influence de certains paramètres dans la qualité globale obtenue : ce qui fait de ce programme un outil intéressant pour des recherches ultérieures ;
- la possibilité d'introduire des caractéristiques particulières à ce système, irréalisables avec les systèmes précédents : en particulier l'introduction du pitch, le codage des amplitudes des oscillateurs à plusieurs bits, la programmabilité des formes d'onde et des fréquences des oscillateurs.

Le but à long terme de ce travail a été, dès le début, la mise en oeuvre du programme sur microprocesseur ; en effet, bien que la méthode de synthèse ICOPHONE ne fournisse pas une voix très esthétique, cette technique présente le grand avantage d'être très simple : grâce à cette caractéristique, il est possible maintenant, comme on le détaillera plus loin, de transférer ce système de synthèse sur un simple microprocesseur à 8 bits avec un matériel très réduit pour obtenir un module de synthèse fonctionnant en temps réel, de petites dimensions et de très faible prix, avec la même qualité de voix que celle de l'ICOPHONE V. Cette considération s'est traduite en pratique de la façon suivante :

- le programme devait être intrinsèquement compatible avec la traduction en langage assembleur c'est-à-dire ne pas faire appel à des sous-programmes de bibliothèques et ne pas faire de calculs en virgule flottante ;
- le programme devait arriver à la synthèse du signal par des opérations simples et rapides ; en particulier, on a exclu a priori toute opération de multiplication et de division, même en virgule fixe, sauf dans les phases pré-

liminaires de préparation des formes d'onde des oscillateurs, qui dans le microprocesseur seront stockées une fois pour toutes dans une mémoire morte.

On dispose actuellement d'une première version complète de ICOLOG, c'est-à-dire un programme écrit presque entièrement en FORTRAN, exécutable sur ordinateur IBM 370, qui accepte un texte (en code phonétique avec marqueurs prosodiques) entré par télétype, calcule les paramètres caractéristiques des formes sonagraphiques de l'intonation et de la prosodie, synthétise les échantillons du signal de parole correspondant, les stocke dans une mémoire-tampon, et envoie enfin la suite des échantillons au convertisseur numérique-analogique, par l'intermédiaire d'un miniordinateur IBM System 7.

LES OSCILLATEURS NUMERIQUES

Plusieurs méthodes ont été proposées dans la littérature pour la génération numérique de la forme d'onde d'un oscillateur sinusoïdal (RABINER, L. et GOLD, B., 1975) ; la plus économique et en même temps la plus souple est sans doute la technique de génération par consultation de tableau ("table look-up"), aujourd'hui employée surtout dans les systèmes digitaux de synthèse musicale (MATHEWS, M.V., 1969 ; ASTA, V., 1978).

Suivant cette technique, la forme d'onde de l'oscillateur (qui, à différence des autres méthodes, ne doit pas être forcément sinusoïdale) est calculée une fois pour toutes et stockée dans une mémoire-tampon ; ensuite cette mémoire est lue circulairement, avec un pas variable dépendant de la fréquence recherchée. Il est évident qu'on peut réaliser plusieurs ocillateurs avec le même tableau de mémoire, et que la forme d'onde mémorisée peut être absolument arbitraire.

Une telle solution est extrêmement intéressante aussi bien pour une réalisation purement logicielle, où elle permet une diminution importante des temps de calcul, que pour une réalisation matérielle, où elle permet d'obtenir un banc d'oscillateurs avec un nombre de circuits très réduit : deux registres seulement sont nécessaires par oscillateur (un pour mémoriser le pas de lecture correspondant à la fréquence voulue, un autre pour mémoriser la phase actuelle), un additionneur et une mémoire pour la forme d'onde (éventuellement une mémoire morte : 2 à 4 Kmots constituent une taille raisonnable).

Cette technique a donc été adoptée pour l'ICOLOG, au niveau logiciel pour la version actuelle, et au niveau matériel pour la version sur microprocesseur,

en substitution des 44 oscillateurs sinusoïdaux de l'ICOPHONE V. Le nombre des oscillateurs a été réduit à 40 (cela n'introduit aucune perte appréciable de qualité dans le résultat global) ; par contre, la forme d'onde de chaque oscillateur n'est pas sinusoïdale mais contient plusieurs fréquences pures autour de la fréquence nominale : ce qui enrichit le spectre, en améliorant légèrement le naturel de la voix obtenue.

LE "PITCH"

La possibilité de faire varier la fréquence fondamentale du signal synthétisé constitue la première innovation importante par rapport à l'ICOPHONE V. En effet, dans ce dernier système, les 44 oscillateurs sinusoïdaux ont des fréquences fixes qui vont de 100 à 4400 Hz, avec une différence constante de 100 Hz entre chaque oscillateur et le suivant. Il en résulte une fréquence fondamentale constante de 100 Hz. Pour changer cette valeur, la méthode la plus simple consiste évidemment à multiplier les fréquences de tous les oscillateurs par un facteur constant ; malheureusement, cette méthode est inacceptable en pratique, car elle introduit des déformations souvent très graves dans la reproduction des voyelles, à cause du décalage introduit dans les formants.

La méthode choisie a été alors la suivante : on multiplie le signal par une fenêtre temporelle fixe, reproduisant les impulsions du signal glottique, se répétant avec une période correspondant à la fréquence fondamentale recherchée. De cette façon, le pitch est reconstitué dans le domaine temporel, et on peut éviter de moduler en fréquence les oscillateurs. Pour éviter toute multiplication en cours de synthèse, les formes d'onde des oscillateurs sont stockées en mémoire déjà enveloppées par la fenêtre, sur toute la longueur de celle-ci. Il est donc nécessaire de disposer de 40 tables de formes d'onde : une pour chaque oscillateur, mais d'un autre côté, chaque table étant utilisée pour synthétiser une seule fréquence, le pas de lecture est constant et égal à 1. Cela permet de limiter au maximum le nombre d'échantillons à mémoriser pour chaque table ; l'encombrement total de mémoire, comme on le détaillera ensuite ne dépasse pas les 4 Kmots de 16 bits. Soit NC la période fondamentale à reproduire, exprimée en nombre d'échantillons, et N la longueur en échantillons de la fenêtre employée ; tous les NC échantillons on déclenchera une impulsion (c'est-à-dire un groupe d'oscillateurs enveloppés), qui s'additionnera éventuellement à la précédente si elle n'est pas encore finie (selon que NC est plus grand ou plus petit que N) (voir fig. 1 et 2).

Si, pour limiter les calculs, on ne veut jamais avoir plus de deux impulsions superposées, il faut que

$$NC \geqslant \frac{N}{2}$$

c'est-à-dire

$$F_z \leqslant \frac{2\,F_c}{N}$$

où F_z est la fréquence fondamentale et F_c la fréquence d'échantillonnage fixée à 10 K Hz.

L'intérêt de minimiser N est donc double : soit pour avoir un encombrement de mémoire moindre avec les formes d'onde des oscillateurs, soit pour avoir plus de possibilités de variation du pitch, ce qui est important surtout pour des applications à la voix chantée.

L'expérience a rapidement montré que le choix de la forme et de la longueur de la fenêtre était très critique pour la qualité de voix obtenue ; toutes les enveloppes testées dans une première phase, et surtout, paradoxalement, celles qui approchaient le mieux la forme du signal glottique, ont donné des résultats très mauvais : le fait de "forcer" la périodicité du signal dans le domaine du temps, en le multipliant par l'enveloppe, introduit des raies dans tout le spectre dues à la convolution en fréquence entre le signal et l'enveloppe, qui masquent ainsi les "formes" fréquentielles des diphonèmes. L'enveloppe finalement choisie donnant la qualité de voix recherchée avec une longueur de 100 échantillons, est actuellement une fenêtre de Kaiser modifiée avec paramètre $m = 4.7$. Les propriétés optimales de la fenêtre de Kaiser normale, exprimée par la formule

$$W_K(n) = \frac{I_o\left[m\sqrt{1-\left(\frac{2n}{N-1}-1\right)^2}\right]}{I_o(m)} \qquad 0 \leqslant n \leqslant N-1$$

où $I_o(-)$ est la fonction de Bessel modifiée du premier type et d'ordre zéro (KAISER, J.F., 1974) et de la fenêtre de Kaiser modifiée, exprimée par la formule

$$W_{Km}(n) = \frac{W_K(n) - W_K(0)}{1 - W_K(0)}$$

(GECKINLI, N.C., et YAVUZ, D., 1978) sont bien connues ; ces caractéristiques nous ont permis de sélectionner une enveloppe introduisant des produits d'intermodulation (dûs aux lobes latéraux du spectre de la fenêtre) atténués de plus de 32 dB par rapport au spectre d'origine, et une dispersion en fréquence

(due à la largeur du lobe principal) inférieure à $\pm$ 175 Hz, tout en gardant une allure satisfaisante du point de vue temporel et une longueur réduite. La mémoire occupée par les oscillateurs est donc actuellement de 4000 mots de 16 bits et la plus haute fréquence fondamentale qu'il est possible d'obtenir est de 200 Hz. Les résultats obtenus pour le pitch sont de bonne qualité ; le programme ICOLOG a, dans sa version actuelle, la possibilité de "faire chanter" le système, en attribuant une note musicale à chaque syllabe du texte à synthétiser.

LA MISE EN OEUVRE DU MICROPROCESSEUR

Cette phase de travail est actuellement en cours de réalisation. Comme on l'a déjà souligné, l'extrême simplicité du type de synthèse adopté permet d'obtenir les performances voulues en temps réel avec un microprocesseur à 8 bits en technologie MOS : le 8080 de INTEL a été choisi d'une part parce qu'il est disponible sur le marché et bien connu des constructeurs de systèmes basés sur microprocesseurs, d'autre part parce qu'il a un vaste catalogue de circuits périphériques.

Dans ce premier prototype, on est revenu, provisoirement, à la solution à fréquence fondamentale constante ; néanmoins, on a gardé l'enveloppe pour les oscillateurs, d'une part par souci de compatibilité avec les versions suivantes, d'autre part pour minimiser les discontinuités dans la dérivée première du signal aux instants de changement de configuration sonagraphique : quand un oscillateur est déclenché ou s'arrête, le contour global du signal total passe toujours par un zéro.

Toute la procédure de synthèse est effectuée par programme à l'exception de l'accumulation des contributions partielles des oscillateurs pour le calcul de l'échantillon du signal résultant, ce qui est réalisé par des oscillateurs digitaux tels qu'on les a décrits précédemment. Le logiciel du microprocesseur comprendra aussi le programme de traduction phonétique de PROUTS, B. (1979), ce qui permettra d'avoir un système complet de synthèse du français à partir du texte écrit sur microprocesseur.

Tout le matériel entourant l'unité centrale composée du microprocesseur et des mémoires de programme et de données, tient sur une carte mesurant 30 x 17 cm, comprenant environ 45 circuits intégrés. Pour des raisons de standardisation et de simplicité de reproduction, la logique d'entrée-sortie de la carte

a été conçue pour fonctionner sur bus universel INTEL équipant le système de développement INTELLEC MDS 800.

Un schéma simplifié à blocs du système compris dans la carte est présenté sur la figure 3. Tous les éléments, à l'exception de la mémoire (ROM), sont réalisés par des circuits en technologie Low-Power-Shottky TTL. La mémoire, qui contient les formes d'onde, est organisée de la façon suivante : d'abord les premiers échantillons des 40 oscillateurs, puis les seconds échantillons des oscillateurs et ainsi de suite. De cette façon avec un seul compteur qui adresse circulairement toute la mémoire on peut réaliser les 40 oscillateurs. La durée totale d'un cycle complet du système est de 2608 nsec, divisé en 16 microcycles de 163 nsec ; pendant ce temps, un échantillon d'un des oscillateurs est sélectionné et additionné (s'il le faut) au résultat partiel précédent, pour former l'échantillon du signal final.

L'accumulation des oscillateurs est faite par l'additionneur et le premier registre (REG. 1) ; dans un registre à décalage (SHIFT REG.) circulent les 40 bits de la configuration sonagraphique actuelle, qui habilitent ou non le registre à accumuler les échantillons de chaque oscillateur. Tous les 40 cycles un échantillon à 12 bits est envoyé au convertisseur par le deuxième registre (REG. 2), pour une fréquence d'échantillonnage résultante de 9586 Hz. Le microprocesseur lui-même se limite donc, en phase de synthèse, à calculer la prochaine configuration sonagraphique et à la transmettre au registre de décalage à chaque fin de fenêtre, soit environ toutes les 10 msec.

Le programme de synthèse occupe environ 14 K octets de ROM et 1 K octet de RAM, soit 5 K octets pour le programme, 9 K octets pour le dictionnaire de diphonèmes et la table d'adressage du dictionnaire, et 1 K octet pour le stockage temporaire des données. Le programme complet de traduction phonétique et de synthèse occupe donc au total environ 28 K octets de ROM et 2 K octets de RAM.

CONCLUSION

L'ICOPHONE LOGICIEL constitue une réalisation moderne, plus souple et meilleur marché, de l'ICOPHONE V. Dans une phase ultérieure nous mettrons en oeuvre la commande mélodique. Mais cette méthode de synthèse sera également étudiée sous l'aspect de la qualité de la synthèse, car elle permet de maîtriser tous les paramètres du signal vocal considéré comme une suite d'impulsions glottiques, en particulier la phase des composantes, qu'elles proviennent de

la source ou du conduit vocal.

REFÉRENCES

ASTA, V., 1977, Tecniche di sintesi per la musica elettronica : la situazione attuale a i recenti sviluppi all' IRCAM. 6° Convegno dell' Associazione Italiana di Acoustica, Ivrea.

BOURGENOT, J.S., DECHAUX, C., 1975, Codage de la parole à faible débit : le vocoder CIPHON, Revue Technique Thomson-CSF, 7, n° 4.

GECKINLI, N.C., YAVUZ, D., 1978, Some novel windows and a concise tutorial comparison of window families, IEEE Trans. ASSP-26, n° 6, pp. 501-507.

KAISER, J.F., 1974, Non recursive digital filter design using the I_o-sinh window function, Proc. IEEE International Symposium on Circuits & Systems, San-Francisco, pp. 20-23.

LEIPP, E., CASTELLENGO, M., SAPALY, J., LIENARD, J.S., 1968, Structure physique et contenu sémantique de la parole, La Revue d'Acoustique.

MATHEWS, M.V., 1969, The technology of computer music, M.I.T. Press, Boston.

PROUTS, B., 1979, Traduction phonétique de texte écrit en français, 10èmes Journées d'Etude sur la Parole, Grenoble.

RABINER, L.R., GOLD, B., 1975, Theory and application of digital signal processing, Englewood Cliffs, Prentice Hall.

RODET, X., 1977, Analyse du signal vocal dans sa représentation amplitude-temps ; synthèse de la parole par règles, Thèse d'Etat, Université de Paris VI.

TEIL, D., 1975, Conception et réalisation d'un terminal à réponse vocale, Thèse de Docteur-Ingénieur, Université de Paris VI.

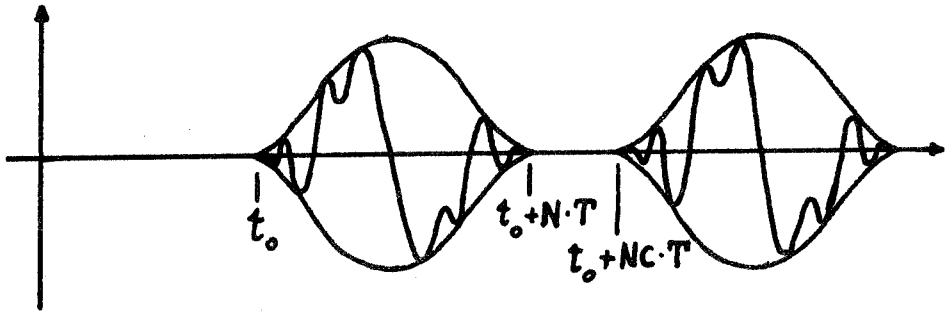


Figure 1. NC > N . Dans la période de temps $t_o + N.T \div t_o + NC.T$, le signal est identiquement nul.

NC > N . In the time interval $t_o + N.T \div t_o + NC.T$, the signal is identically equal to zero.

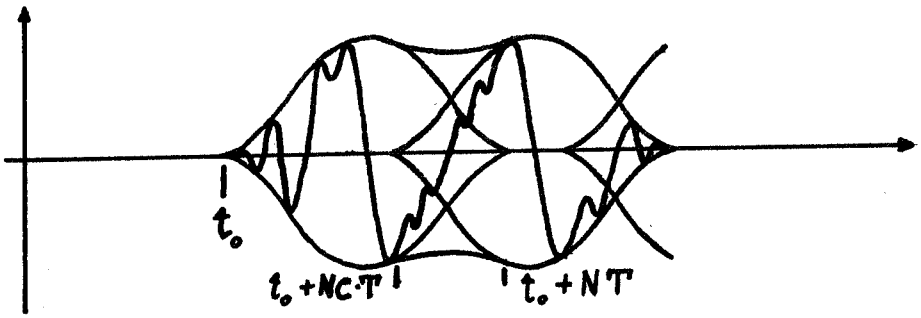


Figure 2. NC < N . Dans la période de temps $t_o + NC.T \div t_o + N.T$, les contributions des deux impulsions sont additionnées.

NC < N . In the time interval $t_o + NC.T \div t_o + N.T$, the contributions of both pulses are added.

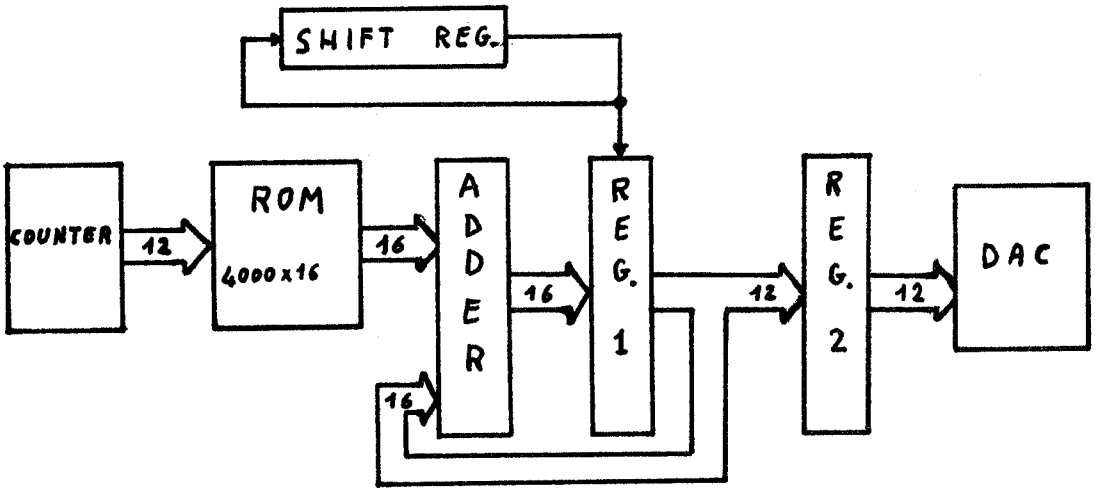


Figure 3.

10èmes JOURNÉES D'ÉTUDE SUR LA PAROLE

GRENOBLE - 30 MAI - 1ER JUIN 1979

QUELQUES EXPERIENCES DE SYNTHESE DE VOYELLES UTILISANT LES PASSAGES PAR ZERO

Camille BELLISSANT

I.M.A.G.
B.P. 53 X
F 38041-GRENOBLE-Cédex

RESUME

Cet article décrit une expérimentation dont le but est de tenter de reconstituer des sons vocaliques à partir de paramètres extraits en vue d'une reconnaissance de parole. Ces paramètres sont des mesures d'amplitudes crête à crête et de passages par zéro du signal et de sa dérivée première dans trois bandes de fréquence. Un système expérimental a été développé, permettant l'ajustement des diverses caractéristiques du signal à synthétiser. Ces caractéristiques sont la durée de l'intervalle d'extraction des paramètres, la désignation des paramètres servant à l'approximation des formants, l'indication de largeur des formants et l'amplitude et la fréquence du fondamental.

SOME EXPERIMENTS IN VOWELS SYNTHESIS USING ZERO-CROSSINGS MEASUREMENTS

Camille BELLISSANT
I.M.A.G.
B.P. 53 X
F 38041-GRENOBLE-Cédex

SUMMARY

This paper describes some experiments whose aim is to attempt to reconstruct vowel-like sounds from parameters extracted for the purpose of speech recognition. These parameters are measurements of peak to peak amplitudes, numbers of zero-crossings of the speech signal and of its first derivative within fixed time intervals and after filtering in three frequency bands (150-900 Hz, 900-2200 Hz, 2200-5000 Hz). An experimental system has been developed, which allows to settle the various characteristics of the signal to be synthetized. These characteristics are the duration of the time intervals in which zero-crossings measurements are done, the designation of parameters used in formants adjustment, the indication of formants bandwidth and the amplitude and frequency of the pitch.

10èmes JOURNÉES D'ÉTUDE SUR LA PAROLE
GRENOBLE - 30 MAI - 1er JUIN 1979

QUELQUES EXPERIENCES DE SYNTHESE DE VOYELLES UTILISANT LES PASSAGES PAR ZERO

Camille BELLISSANT

INTRODUCTION

On s'est intéressé, dans cette étude, à une reconstruction de signal de parole synthétique à partir de paramètres extraits en vue d'une reconnaissance. Dans une première étape, on s'est limité aux seules voyelles orales. Le processus de reconnaissance développé à l'I.M.A.G. (BELLISSANT, C., 1978) fait intervenir un prétraitement électronique dont le but est d'extraire pendant des intervalles de temps de durée fixe Δt certains paramètres du signal de parole. Ce signal est tout d'abord filtré dans trois bandes de fréquence (150-900 Hz, 900-2200 Hz, 2200-5000 Hz), et dans chacune de ces trois bandes, 3 paramètres sont extraits du signal filtré : l'amplitude crête à crête du signal $(A_{max} - A_{min})_{\Delta t}$, le nombre de passages par zéro du signal $Z_{\Delta t}$ et le nombre d'extrema relatifs du signal ou encore le nombre de passages par zéro de la dérivée première du signal $D_{\Delta t}$. C'est donc 9 paramètres qui sont extraits par unité de temps. En reconnaissance, la durée Δt couramment utilisée est de 10 millisecondes; l'appareillage permet de fixer sa valeur entre 0,5 et 15 ms.

METHODE UTILISEE

La méthode est basée sur l'hypothèse que les mesures de passages par zéro sont dans certains cas des approximations correctes des formants (FLANAGAN, J.L., 1972). Ceci est en général vérifié pour ce qui concerne le premier formant des voyelles orales dont une approximation peut être donnée par la mesure Z1 du nombre de passages par zéro du signal dans la première bande (pour une voix masculine). La mesure D1 du nombre d'extrema relatifs double fréquemment pour les voyelles orales la mesure Z1, ou bien lui est légèrement supérieure. Pour les formants d'ordre supérieur, la correspondance est beaucoup moins nette. Par exemple, si pour la voyelle [e] le second formant est correctement approché par la mesure Z2, c'est la mesure Z3 qui approchera le mieux ce même second formant dans la voyelle [y]. Il n'est donc pas possible de fixer une doctrine uniforme pour la détermination du paramètre Z_i ou D_i qui sera utilisé dans l'approximation d'un formant donné pour une voyelle donnée. Aussi, le système expérimental qui a été développé autorise-t-il toute combinaison dans la détermination des composantes du signal à synthétiser. Le fonctionnement de ce système est entièrement interactif et l'utilisateur a la faculté de changer chaque valeur de paramètre ou de coefficient après chaque écoute de synthèse.

DETERMINATION DES CARACTERISTIQUES

La première indication que le système réclame à l'utilisateur est la durée Δt de l'intervalle de temps durant lequel se fait l'extraction des 9 paramètres A_i, Z_i, D_i (i=1, 2, 3) par le prétraitement. Si Δt = 10ms, la quantité d'information est de 900 octets par second de parole (les paramètres sont codés sur 8 bits). Si

Δt = 1ms, c'est 9000 octets qui représentent alors une seconde de son. L'expérience a corroboré ce que l'intuition pouvait laisser prévoir : Plus la durée Δt est courte, plus grande est la précision de la mesure des paramètres pour une même séquence de parole, et en conséquence, meilleure est la qualité de la synthèse obtenue.
Une fois connue la durée Δt, le système construit une période de signal correspondant à un fondamental de 100 Hz. En effet le premier filtre de bande du prétraitement a une fréquence de coupure de 150 Hz, ceci pour éliminer de la mesure des passages par zéro l'influence de la fréquence laryngienne. En synthèse, il faut donc reconstituer ce fondamental manquant. C'est une approximation triangulaire suggérée par FLANAGAN, J.L. (1972) qui a été mise en oeuvre.

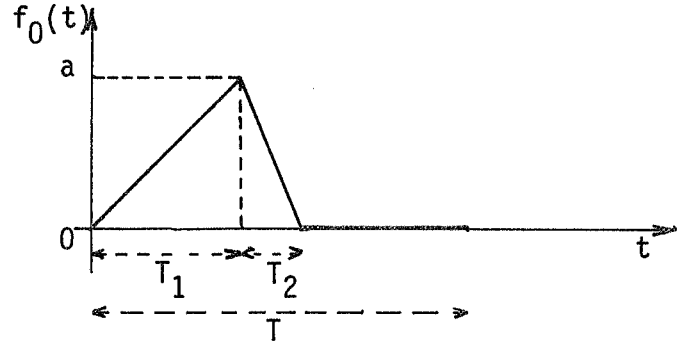


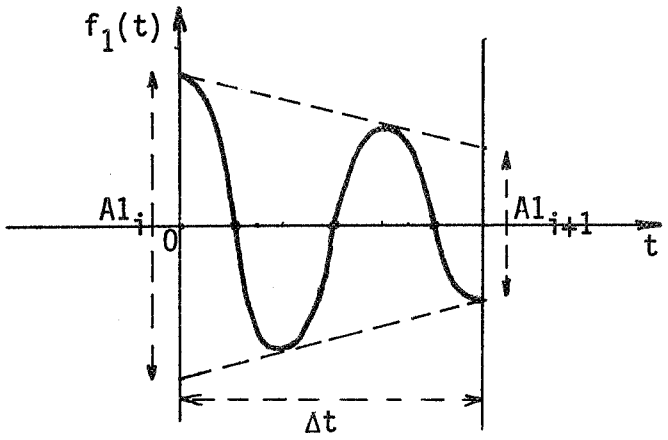
Fig.1 Approximation triangulaire du fondamental.

Triangular approximation to the pitch.

C'est un fondamental fixe de 100 Hz qui est engendré par cette approximation triangulaire. Pour une valeur de la période T de 10 ms, les temps d'ouverture T_1 et de fermeture T_2 qui ont été retenus sont respectivement $T_1 = 0.4T$ et $T_2 = 0.16T$. Ces valeurs ont été obtenues empiriquement après de multiples écoutes.
Après le calcul d'une période de fondamental, le système entreprend la construction de 3 sinusoïdes en demandant à l'utilisateur de lui indiquer pour chacune d'elle quels paramètres extraits par le prétraitement il désire utiliser. Pour chaque sinusoïde, l'utilisateur a trois possibilités : ou bien une mesure du nombre de passages par zéro du signal (par exemple Z1) ou bien une mesure du nombre d'extréma relatifs (par exemple D1) ou encore une moyenne géométrique des deux (l'utilisateur tape alors F1 pour indiquer F1 = partie entière ($\sqrt{Z1 \times D1}$)). Ainsi, si l'utilisateur a répondu

Z1 F2 D3

le système calculera 3 sinusoïdes dont les fréquences pour chaque intervalle Δt seront fournies respectivement par les mesures Z1, $\sqrt{Z2 \times D2}$ et D3. Les amplitudes de ces sinusoïdes sont calculées proportionnellement aux mesures d'amplitudes crête à crête A1, A2, A3 fournies par le prétraitement. Les mesures Z_i, D_i et F_i étant des nombres entiers, le raccordement des sinusoïdes à la frontière entre deux intervalles Δt ne pose pas de problème particulier, la seule donnée à retenir d'un intervalle sur l'autre étant la valeur atteinte par la sinusoïde amortie en fin d'intervalle, la dérivée étant toujours nulle à la frontière entre deux intervalles.



Sinusoïde engendrée pour une valeur Z1 = 3

Fig.2

Example of generated waveform for a value Z1=3

Lorsque ces 3 sinusoïdes f_1, f_2, f_3 sont construites, le système demande à l'utilisateur de pondérer la somme

$$S = c_0 f_0 + c_1 f_1 + c_2 f_2 + c_3 f_3$$

en indiquant les valeurs des coefficients c_i.
Le signal S est alors synthétisé sur l'étage de sortie d'un terminal de parole (BELLISSANT, C., 1978). L'utilisateur a la faculté de le réécouter tel quel ou de modifier les valeurs des coefficients c_i ou encore de calculer d'autres sinusoïdes en introduisant d'autres paramètres fréquentiels, ou enfin travailler sur un autre signal.
Toutes ces opérations s'effectuent en mode interactif sur un ordinateur IBM 360/67 fonctionnant en temps partagé. Les calculs de synthèse s'effectuent en 4 ou 5 fois le temps réel de prononciation.

RESULTATS

Ce système expérimental a permis d'engendrer des signaux synthétiques en en contrôlant complètement les diverses composantes. Les sons engendrés sont d'assez faible qualité lorsque la durée Δt est de l'ordre de 10 ms. Une amélioration sensible est envisagée en tenant compte d'une certaine largeur de formant au lieu de calculer simplement 3 sinusoïdes (LIENARD, J.S., 1977). Lorsque la durée Δt est de l'ordre de 1 ms, les signaux engendrés sont d'une intelligibilité suffisante pour pouvoir reconnaître l'identité du locuteur.

REFERENCES BIBLIOGRAPHIQUES

BELLISSANT, C., 1978, Contribution à l'analyse et à la reconnaissance automatique de la parole. Thèse I.M.A.G. Grenoble.

FLANAGAN, J.L., 1972, Speech Analysis Synthesis and Perception.Springer-Verlag. Berlin.

LIENARD, J.S., 1977, Les processus de la communication parlée. Introduction à l'analyse et à la synthèse de la parole. Masson. Paris.

10èmes JOURNÉES D'ÉTUDE SUR LA PAROLE

GRENOBLE - 30 Mai - 1er Juin 1979

REALISATION ET EXPERIMENTATION D'UN SYSTEME HYBRIDE DE SIMULATION DYNAMIQUE EN TEMPS REEL DES PHENOMENES DE PRODUCTION DE LA PAROLE.

A. BOURJAULT

Laboratoire d'Automatique de Besançon - ENSMM - Route de Gray - 25030 - Besançon Cédex.

RESUME

Dans le but de développer un outil de recherches destiné à mieux connaître les mécanismes de production de la parole naturelle, nous avons réalisé et testé un système hybride pour simuler en temps réel ces phénomènes. En considérant le conduit vocal comme un système mécanique globalement générateur de signaux, continûment déformable dans l'espace et le temps par le locuteur, nous avons établi un modèle mathématique effectivement dynamique. Les équations rendent compte de l'évolution de la pression et de la vitesse de l'air au sein de l'appareil phonatoire, sous l'action des variations des aires transversales. Ainsi, les sources d'excitation (source glottique, sources de bruit ...) ne sont plus traitées à part ; elles apparaissent dès lors que sont simulées les conditions qui permettent leur existence : ouverture et fermeture de la glotte, constriction, occlusion. Les équations sont résolues sur le Simulateur Analogique Modulaire rapide S.A.M., les variables de commande (les fonctions d'aire) étant fournies au modèle par un calculateur numérique. Une première série d'expériences a permis de réaliser les 12 voyelles orales du français et des groupements voyelle-consonne-voyelle.

REALISATION ET EXPERIMENTATION D'UN SYSTEME HYBRIDE DE SIMULATION DYNAMIQUE EN TEMPS REEL DES PHENOMENES DE PRODUCTION DE LA PAROLE.

A. BOURJAULT

SUMMARY

In order to develop a research device intended to investigate the production mechanism of natural speech, we have performed and tested a hybrid system to simulate these phenomenons in real time.

We have considered the vocal tract as a mechanical system, globally productive signals, continuonsly variable in space and time, and we have made a really dynamic mathematical model. The equations account for air pressure and air particle velocity inside the vocal tract, under the effects of cross areas variations. Thus, we don't deal with excitation sources (Glottic or noise sources) separately. These sources appear as long as the conditions of their existence are simulated : Glottis opening and closing, constriction, occlusion.

The systematic exploitation of such a model requiring a real time simulation, we have chosen a hybrid set as a realisation medium aid. However, the solution of discretized equations imposes an high speed performing analog computer of great capacity ; one was especially made in the laboratory. The control (programming of time and space variations of vocal tract cross area) has been attributed to a numerical computer.

A first series of experiments enabled us to perform the twelwe french oral vowels and some vowel/consonant/vowel Clusters :
/apa/, /ada/, /aka/, Lara/, /ala/, /aza/ etc... . The particularly simple programming shape allows us to study several parameters such as duration, transition, articulation points, etc... .

10èmes JOURNÉES D'ÉTUDE SUR LA PAROLE
GRENOBLE - 30 MAI - 1ER JUIN 1979

REALISATION ET EXPERIMENTATION D'UN SYSTEME HYBRIDE DE SIMULATION DYNAMIQUE EN TEMPS REEL DES PHENOMENES DE PRODUCTION DE LA PAROLE.

A. BOURJAULT

INTRODUCTION

Si la synthèse s'intéresse surtout au signal de parole, la simulation prend davantage en compte les processus mécaniques et acoustiques de production. A ce titre, les simulateurs permettent généralement d'approfondir l'ensemble des phénomènes intervenant dans la production de la parole. C'est dans cet esprit que nous avons élaboré et expérimenté un système hybride de simulation dynamique en temps réel.

MODELE MATHEMATIQUE

Pour prendre réellement en compte l'aspect dynamique du processus, nous avons dû repenser le modèle mathématique généralement utilisé. L'étude de la genèse des sons met en évidence deux fonctions distinctes mais complémentaires de l'appareil phonatoire : la génération d'un ébranlement sonore et son modelage. De notre point de vue, l'appareil phonatoire est un système mécanique global, continuellement déformable dans l'espace et le temps par le locuteur. Nous ne faisons (a priori) aucune distinction entre ces deux fonctions. Les sources d'excitation sont alors décrites comme résultant d'une action mécanique : constriction, occlusion, ouverture et fermeture de l'orifice glottique, etc... . Le modèle mathématique résultant, dont les équations sont établies en appliquant les lois de la mécanique des fluides, prend en compte aussi bien les aspects stationnaires que transitoires de la parole et intègre de ce fait les sources (CHEVILLARD A.,BOURJAULT A.,LHOTE F., 1975).

Ainsi, l'évolution de la pression $p(x,t)$ et de la vitesse $v(x,t)$ de l'air au sein du conduit vocal, sous l'action des variations de l'aire des sections droites $A(x,t)$ s'exprime par :

$$\begin{cases} -\frac{\partial p}{\partial t} = \rho C^2 \frac{\partial v}{\partial x} + \rho C^2 \left(\frac{1}{A}\frac{\partial A}{\partial x}\right) v + \rho C^2 \left(\frac{1}{A}\frac{\partial A}{\partial t}\right) + \rho C^2 \frac{K_1}{\sqrt{A}} p \\ -\frac{\partial v}{\partial t} = \frac{1}{\rho}\frac{\partial p}{\partial x} + \frac{1}{\rho}\frac{K_2}{\sqrt{A}} v \end{cases}$$

équations unidimensionnelles linéarisées, obtenues moyennant certaines hypothèses classiques.

SIMULATION HYBRIDE : PARTIE ANALOGIQUE

Ces équations ont été discrétisées (CHEVILLARD A., 1973) par rapport à la variable d'espace x (Δx = 1 cm) ; à chacune des trois parties du conduit vocal (figure 1) correspond un système d'équations du type :

$$\begin{cases} -\frac{dp_k}{dt} = A\,(v_k - v_{k-1}) + B\left(\frac{1}{A}\frac{\partial A}{\partial x}\right)_k v_k + C\left(\frac{1}{A}\frac{\partial A}{\partial t}\right)_k + D\frac{1}{\sqrt{A_k}} p_k \\ -\frac{dv_{k-1}}{dt} = E\,(p_k - p_{k-1}) + F\frac{1}{\sqrt{A_{k-1}}} v_{k-1} \end{cases}$$

auxquelles s'ajoutent deux conditions de raccordement du conduit nasal (égalité des pressions et égalité des débits au point de raccordement) et trois conditions limites (radiation aux lèvres, au nez, et pression subglottique constante). Il est à noter que la première tranche représente la glotte.

La résolution en temps réel par rapport aux inconnues p(x,t) et v(x,t) nécessite l'utilisation d'un calculateur analogique rapide et de grande capacité; un tel outil n'existant pas sur le marché, nous avons réalisé au Laboratoire le Simulateur Analogique Modulaire SAM répondant à ces exigences (ANDRE P. et al., 1975).

SIMULATION HYBRIDE : PARTIE NUMERIQUE

Les variables de commande représentant des excitations d'origine mécanique apparaissent dans les équations sous la forme $\left(\frac{\partial}{\partial x} \text{Log } A\ (x,t)\right)_k$ et $\left(\frac{\partial}{\partial t} \text{Log } A(x,t)\right)_k$ A partir des données utilisées par d'autres auteurs (MRAYATI, M., 1976) nous avons déterminé et stocké sur calculateur numérique les suites :

$\{A_{k,i}\} = \{A(x = x_k, t = t_i)\}$ décrivant l'évolution des aires au cours d'une séquence de parole. ($t_{i+1} - t_i = 4$ ms). Les variations de l'aire glottique, représentées par la suite $\{A_{o,i}\}$ ont été assimilées dans un premier temps à celles d'un signal triangulaire, de fréquence 125 Hz.

Les commandes proprement dites sont élaborées par une interface constituées de mémoires analogiques et d'interpolateurs linéaires (figure 2).

EXPERIMENTATION

La nécessité de tester la validité de notre modèle nous a conduit à réaliser une série progressive d'expériences (BOURJAULT A., 1978) ; pour des raisons d'ordre matérielles nous avons limité nos premières investigations au seul conduit vocal.

Etude des caractéristiques des 12 voyelles orales du français :

Etude sur un modèle statique à 17 tranches, destinée à vérifier l'aptitude de la partie analogique à rendre compte de configurations fixes dans le temps ; pour chaque voyelle nous avons mesuré les valeurs des trois premiers formants (figure 3) ainsi que la distribution spatiale de la pression à ces fréquences.

Réalisation dynamique et en temps réel de voyelles et de groupements de voyelles :

Etude sur un modèle à 13 tranches ; par rapport aux expériences précédentes, nous avons une tranche supplémentaire, correspondant à la glotte, commandée par la fonction d'aire de l'orifice glottique. La qualité (sonore) des phonèmes et diphonèmes ainsi réalisés n'est peut-être pas aussi bonne que nous l'espérions; le signal de sortie est parfois entaché de bruit dont sont responsables les intégrateurs et les multiplieurs. Des expériences en temps ralenti, actuellement en cours, permettront d'éclaircir ce point.

Réalisation dynamique et en temps réel de consonnes dans des groupements de type voyelle-consonne-voyelle :

Notre modèle rendant compte des variations de pression et de vitesse en tout point de l'appareil phonatoire et à tout instant, ce ne sont pas les phénomènes eux-mêmes (qu'ils soient acoustiques ou mécaniques) qui sont simulés, mais davantage les circonstances qui les engendrent ; les sources de bruit peuvent donc apparaître d'elles-mêmes, dès lors que sont simulées les conditions (occlusion, constriction ...) qui leur donnent naissance, (BOURJAULT A., CHEVILLARD A., 1976) c'est-à-dire dès lors que la fonction d'aire traduit ces faits.

Nous avons ainsi réalisé les groupements /apa/, /aka/, /ada/, /ala/, /ara/, /aza/, /a a/, /asa/. Dans tous les cas (sauf /asa/) ces séquences ont été identifiées facilement par des personnes prises au hasard et n'ayant jamais entendu de parole artificielle ; malgré un certain niveau de bruit accompagnant la voyelle, les consonnes réalisées sont toutes reconnues.

A partir d'une configuration donnée, nous avons exploré presque systématiquement les configurations immédiatement voisines, c'est-à-dire en modifiant un seul paramètre tel que : durée de la consonne, transition consonne-voyelle, longueur et section de la constriction, longueur de l'occlusion.

CONCLUSION

En élaborant et en expérimentant un système hybride de simulation dynamique en temps réel des phénomènes phonatoires, notre but était double : créer un outil destiné à mieux connaître ces phénomènes, et vérifier un certain nombre d'hypothèses propres à notre modélisation. Malgré une réalisation que nous considérons encore comme provisoire, nous estimons avoir atteint ces deux objectifs.

Les expériences que nous avons menées montrent l'aptitude de notre modèle à rendre compte des phénomènes engendrés au sein de l'appareil phonatoire : les variations de section suffisent à elles seules, sans l'aide d'aucune source extérieure d'excitation, à produire un voisement ou un bruit. De plus, la simulation s'effectuant en temps réel, le résultat est entendu immédiatement et la programmation de la commande est suffisamment simple pour répercuter immédiatetoute modification.

REFERENCES

ANDRE P., BOURJAULT A., CHEVILLARD A., HENRIOUD J.M.
1975, Calculateur analogique rapide pour la simulation en temps réel des phénomènes de phonation ; Actes du Symposium International Simulation'75, Zürich, pp.486-490

BOURJAULT A.,
1978, Réalisation d'un système hybride de simulation dynamique en temps réel des phénomènes acoustiques au sein du canal vocal ; thèse de docteur-ingénieur, Université de Franche-Comté

BOURJAULT A., CHEVILLARD A.
1976, Le problème des sources dans la simulation dynamique du tractus vocal ;
7ièmes Journées d'Etudes sur la Parole, Nancy, pp.145-151

CHEVILLARD A.
1973, Contribution à l'étude analogique de l'appareil phonatoire ;
Thèse de Docteur-Ingénieur, Université de Franche-Comté

CHEVILLARD A., BOURJAULT A., LHOTE F.
1975, Nouvelle approche pour la simulation dynamique du canal vocal;
Revue d'Acoustique, n° 38, pp.25-29

MRAYATI M.
1976, Contribution aux études sur la production de la parole ;
Thèse d'état, Université de Grenoble

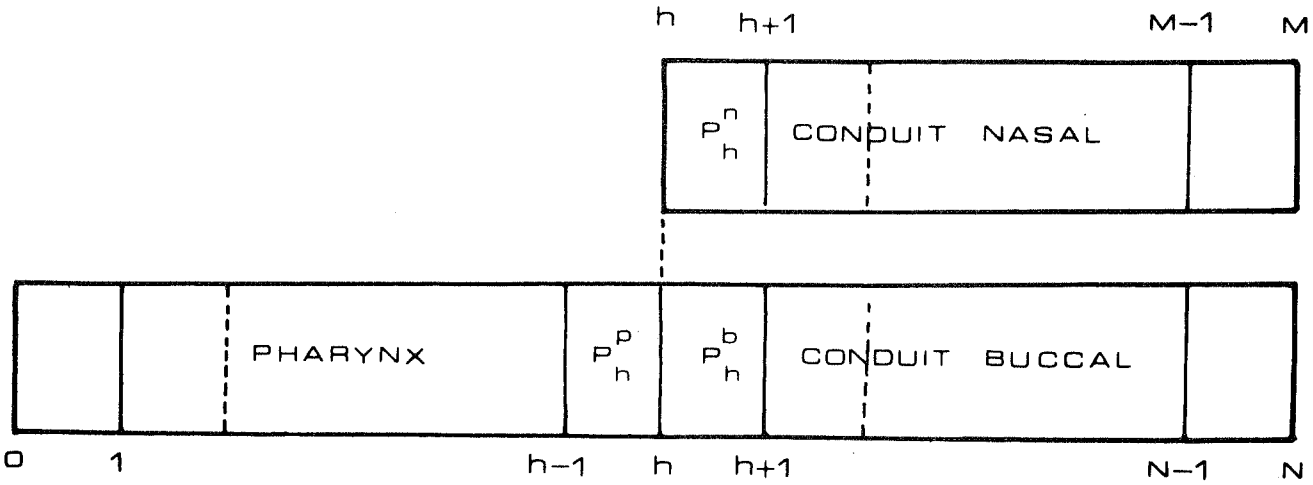


FIGURE 1 : Représentation symbolique du découpage en trois conduits de l'appareil phonatoire.

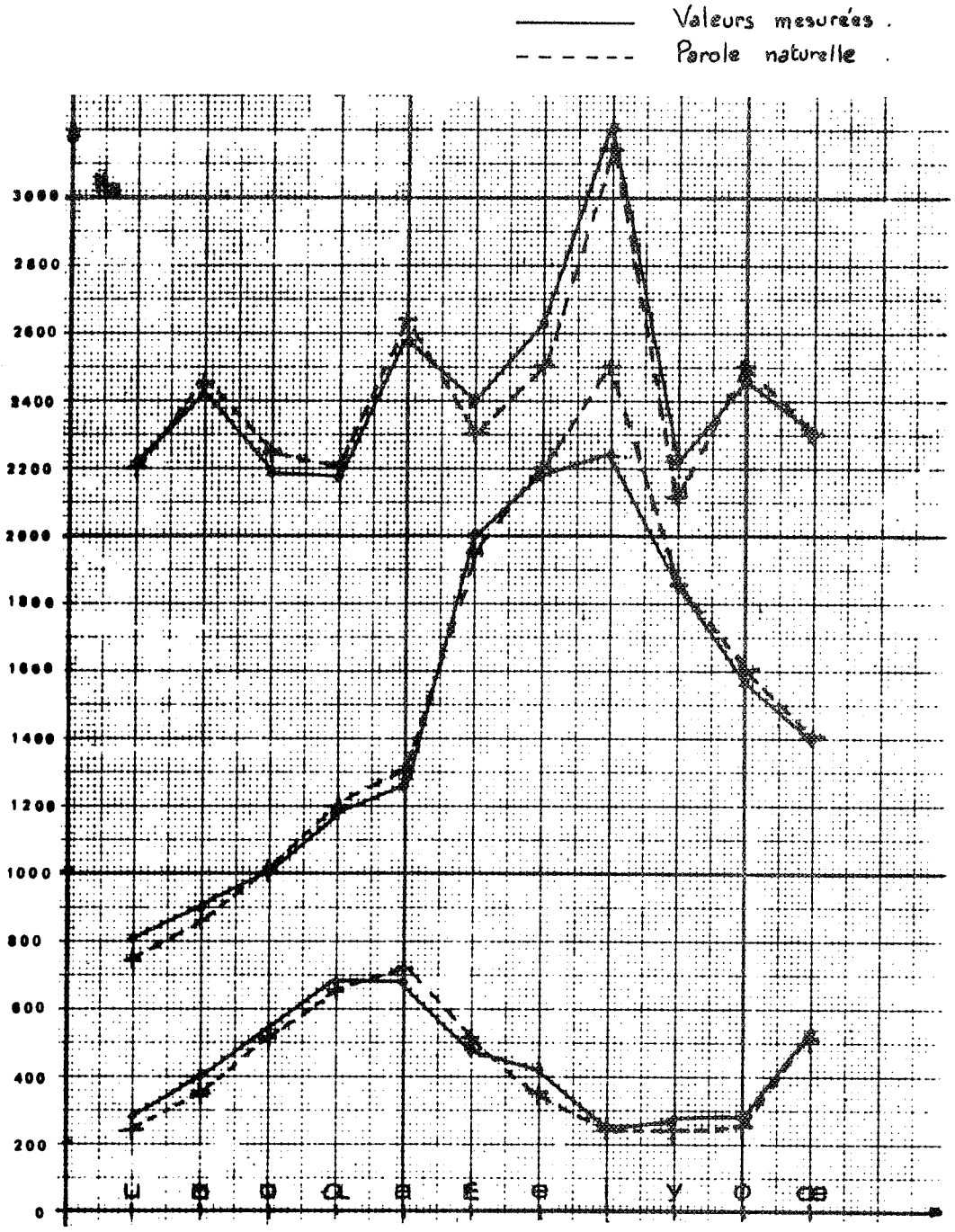


FIGURE 2 : Mesure des 3 premiers formants des voyelles françaises.

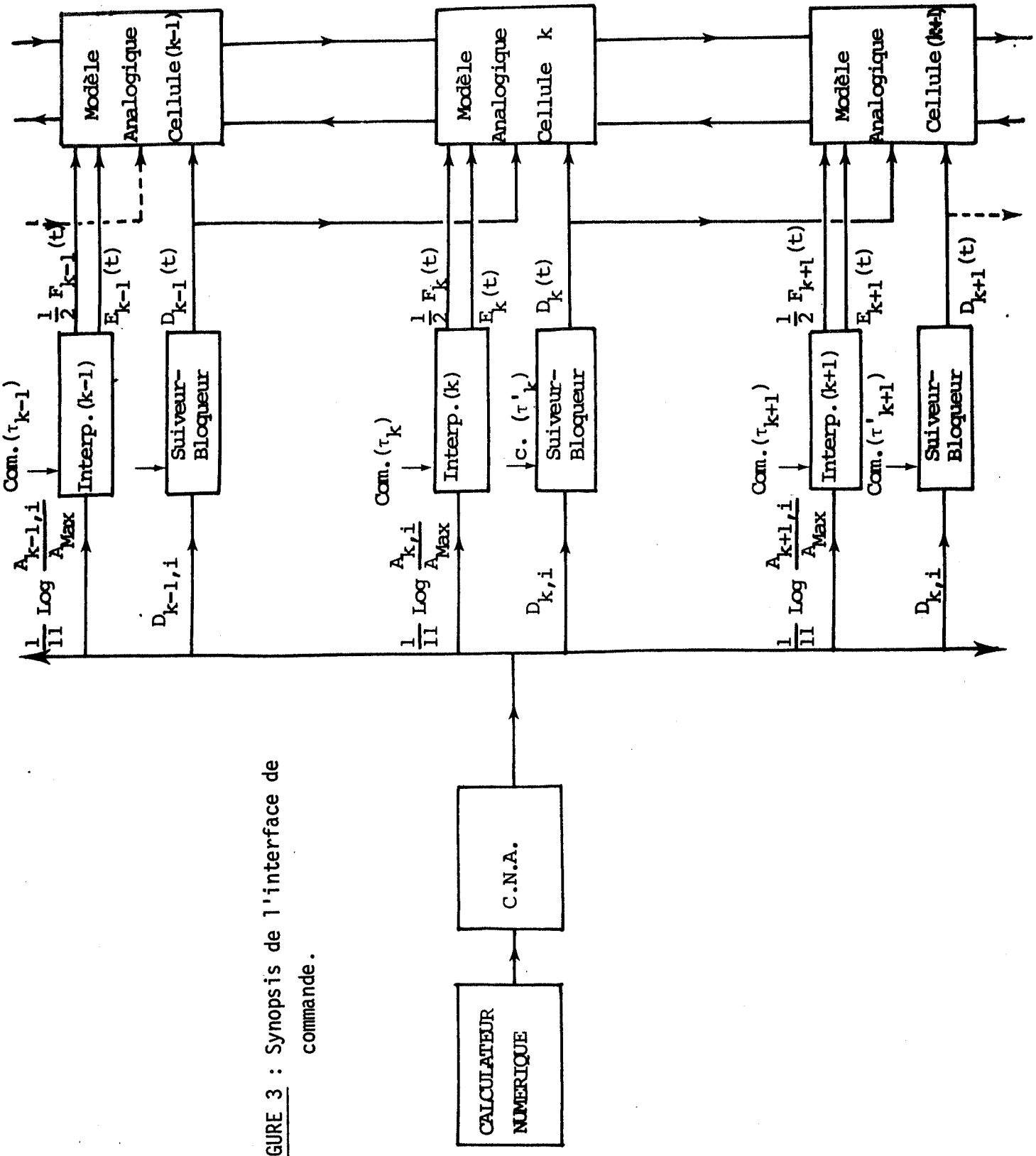


FIGURE 3 : Synopsis de l'interface de commande.

10èmes JOURNÉES D'ÉTUDE SUR LA PAROLE

GRENOBLE - 30 MAI - 1ER JUIN 1979

MISE EN OEUVRE D'UN CALCULATEUR SPECIALISE POUR LA SYNTHESE EN TEMPS REEL DE LA PAROLE.

D. DEGRYSE - J.F. SERIGNAT - P. ABAUZIT

Laboratoire de la Communication Parlée.
E. N. S. E. R. G.
23, Avenue des Martyrs - 38031 GRENOBLE CEDEX
Equipe de Recherche associée au C.N.R.S. N° 366

RESUME

Pour simuler en temps réel des synthétiseurs numériques de parole, on utilise un calculateur spécialisé pour le traitement du signal. Ce calculateur a été conçu et réalisé au Laboratoire de la Communication Parlée.

Ce calculateur spécialisé programmable, associé à un mini ordinateur. présente un certain nombre de particularités qui permettent d'atteindre des performances élevées en traitement du signal.

Dans le cadre de cette communication nous décrirons brièvement ce calculateur et indiquerons les performances obtenues lors de la simulation d'un synthétiseur à formants et d'un synthétiseur à prédiction linéaire.

MISE EN OEUVRE D'UN CALCULATEUR SPECIALISE POUR LA SYNTHESE EN TEMPS REEL DE LA PAROLE.

D. DEGRYSE - J.F. SERIGNAT - P. ABAUZIT.
Laboratoire de la Communication Parlée. E.N.S.E.R.G.
23, avenue des Martyrs - 38031 GRENOBLE CEDEX
Equipe de Recherche associée au C.N.R.S. N° 366

SUMMARY

A specialized speech processor is used to simulate a digital synthesizer in real time. This processor has been constructed at the " Laboratoire de la Communication Parlée".
This speech processor, which can be connected with a LSI 11, is programmable and uses pipeline schemes, parallel processing and high speed memories to increase the operating speed.
In this paper, we describe the general structure of this processor and its uses for the simulation of a formant synthesizer and a linear predictive synthesizer.
In the case of the formant synthesizer we have a real time simulation at 16 khz sampling rate. This simulation includes also parameters filtering and filter coefficient calculation.
For the predictive linear synthesizer, each output sample is computed in less than 20 μs for a 10 sections filter at a sampling rate of 10 KHZ.

10èmes JOURNÉES D'ÉTUDE SUR LA PAROLE
GRENOBLE - 30 MAI - 1er JUIN 1979

MISE EN OEUVRE D'UN CALCULATEUR SPECIALISE POUR LA SYNTHESE EN TEMPS REEL DE LA PAROLE.

D. DEGRYSE - J.F. SERIGNAT - P. ABAUZIT.
Laboratoire de la Communication Parlée - E.N.S.E.R.G.
23 Avenue des Martyrs - 38031 GRENOBLE CEDEX
Equipe de Recherche associée au C.N.R.S. N° 366

1. INTRODUCTION

Dans le cadre des études sur la parole éffectuées au Laboratoire de la Communication Parlée, nous avons conçu et réalisé un calculateur spécialisé pour le traitement de signal.

Ce calculateur, dont nous décrivons la structure générale dans une première partie, fait appel à des techniques de parallélisme afin d'obtenir des performances élevées dans le cadre du traitement du signal de parole.

Dans une deuxième partie, nous décrivons les performances de ce calculateur lors de la simulation d'un synthétiseur à formants dont les paramètres de commande sont fournis par un miniordinateur. Le calculateur spécialisé assure les opérations de lissage des paramètres, de calcul des coéfficients du filtre de synthèse et de synthèse du signal de parole.

Dans un dernière partie, nous décrivons une application de synthèse à l'aide d'un filtre en échelle dont les coéfficients de commande sont obtenus par une analyse à prédiction linéaire.

2. CALCULATEUR SPECIALISE

2.1. Ensemble du matériel.

L'ensemble du matériel utilisé pour la synthèse comprend :

- un calculateur spécialisé pour le traitement du signal de parole. Ce calculateur est équipé d'un système d'entrée sortie analogique.

- un microordinateur LSI 11 de chez DIGITAL EQUIPEMENT équipé principalement de 56 Koctects de mémoire.

- d'une unité de deux disques souples d'une capacité totale de 512 Koctets
- d'un terminal de visualisation alphanumérique et graphique du type 4010 de TEKTRONIX associé à une table traçante.

2.2. Stucture du calculateur spécialisé.

Notre calculateur spécialisé pour le traitement de la parole (DEGRYSE 1976), est organisé sous la forme de plusieurs modules fonctionnant en parallèle avec un cycle de base de 300 ns (Figure 1) :

- un module arithmétique et un opérateur du type somme de produits, dans lesquels sont éffectués les opérations relatives aux échantillons du signal. Ce module éffectue les opérations sur des nombres de 24 bits et possède un ensemble de 16 registres banalisés de 24 bits ; il est associé avec un opérateur somme de produits, ce type d'opérations étant trés courant en traitement numérique du signal.

- une mémoire de données MD composée de 2 plans de 1 K mots de 24 bits et d'un plan de 1 K mots de 16 bits où sont rangés les valeurs relatives au signal et les constantes de clacul.

- un module de calcul d'adresse où sont éffectuées toutes les opérations nécessaires pour accéder à la mémoire MD. Ce module qui effectue ses calculs sur 12 bits, possède 16 registres de 12 bits qui peuvent être utilisés comme registres de base, d'index ou de compte de boucle.

- une mémoire de programme MP, d'une capacité de 1 K mots de 32 bits dans laquelle sont rangées les instuctions.

- un module de contrôle, comprenant le compteur ordinal associé à une pile à 16 niveaux, dans lequel sont gérés les appels de sous programme, les boucles de programme et les divers tests d'indicateurs.

- un module d'interface assurant, en particulier, la connexion des mémoires MD et MP avec le calculateur LSI 11 (voir § 2.3.)

Tous ces modules sont interconnectés par l'intermédiaire de trois bus :

. un bus instruction contenant l'instruction suivante à exécuter. Le contenu de ce bus peut être fourni soit par la mémoire MP, soit par le module arithmétique ou le module de calcul d'adresse. Cette dernière possibilité permet l'échange de données entre ces deux modules ;

. un bus mémoire, BM, assurant la connexion entre les divers plans de la mémoire MD et le module arithmétique ;

. un bus entrée sortie, BES, assurant la connexion entre les différentes mémoires et le module d'interface.

Les 32 bits, I0 à I31, d'une instruction sont utilisés de la manière suivante :

. les bits I16 à I31 contrôlent le module arithmétique et l'opérateur "somme de produits" ;

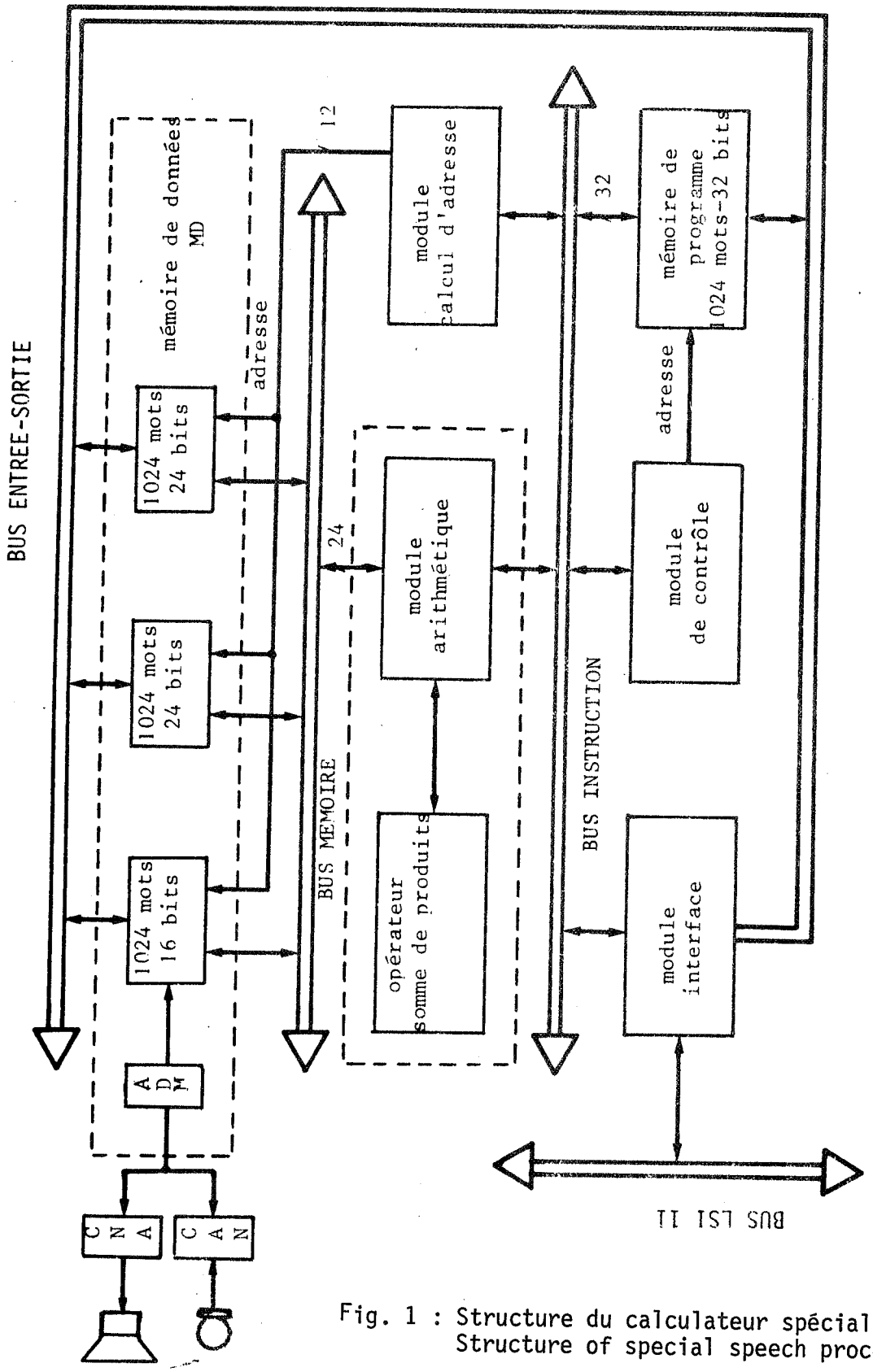


Fig. 1 : Structure du calculateur spécialisé .
Structure of special speech processor.

. les bits I0 à I15 peuvent soit contenir une constante destinée au module arithmétique, soit contrôler le module de calcul d'adresse ou le module de contrôle.

Tous ces modules sont indépendants et fonctionnent en parallèle avec un cycle de base de 300 ns.

Les tâches suivantes sont éffectuées à chaque cycle :

- une opération de calcul relative aux échantillons de signal est effectuée dans le module arithmètique et l'opérateur somme de produits ;

- une opération, permettant au cycle suivant d'accéder à la memoire MD, est réalisée dans le module de calcul d'adresse ;

- une donnée peut être lue ou écrite dans la mémoire MD à l'adresse calculée au cycle précédent par le module de calcul d'adresse ;

- la valeur du compteur ordinal est mise à jour ;

- la prochaine instruction à exécuter est lue dans la mémoire MP.

Cette structure parallèle permet d'obtenir des performances élevées, mais entraîne une grande complexité de programmation.

2.3. Liaison du calculateur spécialisé avec l'extérieur

Le module d'interface permet la liaison de notre calculateur spécialisé avec un calculateur du type LSI 11 du laboratoire. Ce module se décompose en deux parties :

- la liaison des divers modules avec le LSI 11
- la liaison des mémoires de données et de programme avec le LSI 11, chaque plan de mémoire (Fig. I) étant accéssible soit depuis le calculateur spécialisé, soit depuis le module d'interface.

Pour faciliter le traitement en temps réel du signal de parole, un jeu de convertisseurs numérique-analogique et analogique-numérique vient d'être connecté via un accés direct à un des plans de la mémoire de données. La fréquence maximale d'échantillonnage (limitée par le convertisseur analogique-numérique) est de 50 KHz.

2.4. Logiciel de base

Le calculateur spécialisé étant dans l'impossibilité de fonctionner en autonome, le développement et l'exploitation des programmes est assuré par l'ordinateur LSI 11 auquel il est connecté :

- un assembleur élémentaire est disponible sous forme de macroinstructions qui sont exploitées par le logiciel MACRO 11 du LSI 11 ;

- un ensemble de sous-programmes, compatibles avec le système FORTRAN, permet d'assurer les fonctions nécessaires à l'exploitation du calculateur spécialisé. De plus, un module permet d'effectuer, en conversationnel, la mise au point des programmes du calculateur spécialisé.

3.SYNTHESE A FORMANTS

Un synthètiseur à formants du type série est composé d'un canal vocal, d'un canal nasal, d'un canal de bruit et d'une et d'une exécution. Chacun de ces canaux est constitué principalement par une mise en cascade de plusieurs filtres de formants et éventuellement de filtres d'antiformants. Chacun de ces filtres de formant a une fonction de transfert numérique (GOLD B. et al.,1968) qui s'écrit :

$$H(z) = \frac{1 - 2r\cos(2\pi f_i T) + r^2}{1 - 2r\cos(2\pi f_i T)\, z^{-1} + r^2 z^{-2}}$$

où $r = e^{-2\pi\Delta f_i T}$; Δf_i bande passante ;
f_i = fréquence de formant
T = période d'échantillonnage

La simulation numérique d'un tel filtre entraîne le calcul d'une somme de trois produits par échantillon de signal de synthèse (Fig. 2).

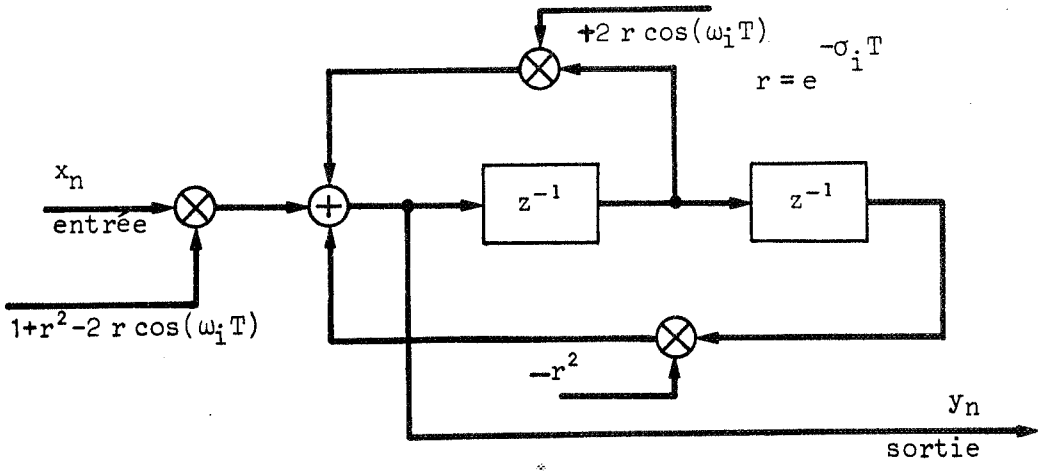


Figure 2 : Stucture d'un filtre de formant numérique.
Structure of a digital formant filter

La simulation du synthétiseur à formants est effectuée avec une fréquence d'échantillonnage de 16 KHz. Les paramètres de commande, fournis par le LSI 11, peuvent être renouvellés à des périodes multiples de 5 ms. Les paramètres de commande sont filtrés avec une constante de temps variable entre 5 et 20 ms, ce filtrage étant effectué toutes les 2,5 ms.

Le programme de synthèse se décompose de la manière suivante :

- un ensemble de sous programmes, chacun étant affecté à la simulation d'une fonction particulière : source vocale, source de bruit, filtre de formant, filtre d'anti-formant.
- un programme principal appelant ces divers sous programmes ; l'ordre dans lequel sont appelés les divers sous programmes définissant la structure du synthétiseur simulé.

Le sous programme de filtre de formant effectue la synthèse d'une séquence de 80 échantillons de signal. Il assure, pour les filtres dont la fréquence peut être commandée, le filtrage de la commande, le calcul des coéfficients du filtre de synthèse. La fonction cosinus est obtenue par un polynome d'approximation du quatrième ordre et la fonction exponentielle par un développement au premier ordre ; cette méthode ne nécessite pas de tableaux volumineux tout en assurant une précision suffisante et une grande rapidité de calcul. Le temps moyen de calcul par filtre et par échantillon de signal est de 4,5μs pour un filtre commandé et de 3μs pour un filtre à fréquence fixe. Le temps permet donc la simulation complète en temps réel d'un synthétiseur à formants, la période d'échantillonnage étant de 62,5μs

En comparaison avec un synthétiseur réalisé en technique analogique, une simulation numérique permet d'obtenir un trés bon rapport signal sur bruit, les calculs internes étant effectués sur 24 bits. Par ailleurs, l'aspect programmable d'une simulation numérique offre un maximum de souplesse pour définir la structure du synthétiseur.

4. SYNTHESE A PREDICTION LINEAIRE

Dans le cas d'une synthèse à partir de paramètres fournis par une analyse du type à prédiction linéaire, nous utilisons un filtre numérique récursif en échelle dit "à deux multiplieurs" (SERIGNAT, 1974).
Les paramètres de commande d'un tel filtre sont
- les coéfficients de réflexion
- l'amplitude du signal d'exitation
- la valeur de la période de mélodie

Le nombre de cellules (compris entre 8 et 14) peut être facilement modifié.

La synthèse en temps réel est effectuée à partir de paramètres rangés dans la mémoire du LSI 11, ces paramètres ayant été, au préalable, lus depuis un fichier sur disque.
L'ensemble des programmes de simulation sont organisés de la manière suivante :

- un programme de transfert s'exécutant sur le LSI 11, assure le passage à la cadence de 20 ms des paramètres de synthèse depuis la mémoire du LSI 11 vers celle du calculateur spécialisé

- un programme de synthèse s'exécutant sur le calculateur spécialisé :

. calcule les échantillons du signal de parole
. effectue toutes les 2,5 ms l'interpolation des paramètres de synthèse à l'intérieur d'une séquence de 20 ms
. gère l'ensemble d'accès direct mémoire pour effectuer la sortie en temps réel des échantillons de parole vers le convertisseur numérique analogique à une fréquence d'échantillonnage de 10 KHz

- la désaccentuation est effectuée au niveau de la sortie du convertisseur numérique analogique.

Le programme de synthèse comprend de l'ordre de 60 instructions et le temps moyen d'exécution par échantillon de signal pour une synthèse à 10 coéfficients se décompose en :

- 1,5 µs pour le calcul du signal d'excitation
- 15 µs pour la simulation du filtre en échelle soit 1,5 pour chaque cellule du filtre
- 3,5 µs pour l'interpolation des coéfficients et la gestion du système d'accés direct mémoire.

Soit un total de 20 µs par échantillon pour une période d'échantillonnage de 100 µs

CONCLUSION

Les techniques numériques de traitement du signal de parole etant de plus en plus couramment utilisées, nous avons développé un calculateur spécialisé utilisant des concepts de parallélisme associés avec un cycle de base rapide , afin de pouvoir effectuer des simulations en temps réel.
Nous avons simuler en temps réel un synthètiseur à formants avec une fréquence d'échantillonnage de 16 KHz. Ce synthètiseur est caractérisé principalement, vis à vis des synthètiseurs analogiques, par un excellent rapport signal sur bruit-dû à la précision des calculs- et par une trés grande souplesse de modification de sa structure.
La simulation d'un synthètiseur à prédiction linéaire permet de montrer la grande rapidité de ce calculateur.
Au cours d'autres études sur la production de la parole ménées au Laboratoire de la Communication Parlée, ce calculateur sera utilisé pour simuler un modèle du conduit vocal à commande articulatoire.

REFERENCES

DEGRYSE D., 1976. Etude et réalisation d'un calculateur spécialisé pour le traitement du signal de parole. Thèse de Docteur Ingénieur, GRENOBLE.

GOLD B., RABINER L.R., 1968. Analysis of digital and analog formant synthesizers. IEEE Trans., C. 20, pp 33-38.

PAILLE J., BEAUVIALA J.P., CARRE R., 1970. Synthèse de la parole : description et utilisation d'un synthétiseur du type à formants. Revue de Physique Appliquée, vol. 5, pp 785-792.

SERIGNAT J.F., 1974. Etude et simulation d'un vocodeur à prédiction linéaire. Thèse de Docteur Ingénieur, GRENOBLE.

10èmes JOURNÉES D'ÉTUDE SUR LA PAROLE

GRENOBLE - 30 MAI - 1ER JUIN 1979

SIMULATION NUMERIQUE DU CONDUIT VOCAL

ESPESSER R.

Institut de Phonétique, Université de Provence
29, Avenue R. Schuman - 13621 AIX-en-PROVENCE

RESUME

Nous présentons une simulation basée sur le modèle de KELLY J. , LOCHBAUM J. dont nous avons amélioré le comportement en dynamique (prise en compte des changements de longueur, élimination du bruit lors des mouvements du conduit).

Le système est actuellement limité à la synthèse des voyelles orales et est implanté sur un ordinateur T 1600.

A DIGITAL SIMULATION OF THE VOCAL TRACT

ESPESSER R.

SUMMARY

We present a digital simulation of the vocal tract (VT) based on the model outlined in KELLY J., LOCHBAUM J. (1962) which consists in dividing the VT into a number of elementary tubes and calculating the propagation of pressure wave through the successive junctions between these tubes.

The dynamic performance of this model has been improved on the followings points :

- changes in the length of the VT - important in French - are taken into account by the adjonction/suppression of elementary tube (cf. Fig. 1.1 1.2, 1.3)
- noise (clicks) produced during transitione have been suppressed by means of a sufficiently five linear interpolation between two target configurations of the VT. The suppression of noise produced by the changes in the length of the VT required special treatment.

The vocal source is an "oscillator" as defined in MUSIC V (cf appendix). The system is at present limited to the synthesis of oral vowels. Despite the simplifycing hypothesis concerning losses and lip impedance, the formant values obtained are satisfactory (cf comparative table 1) and the quality of the synthesis is judged as good by listeners.

10èmes JOURNÉES D'ÉTUDE SUR LA PAROLE
GRENOBLE - 30 MAI - 1ER JUIN 1979

SIMULATION NUMERIQUE DU CONDUIT VOCAL *

ESPESSER R.

INTRODUCTION

Le modèle de KELLY J. LOCHBAUM J. (1962), permettant une simulation éventuellement détaillée du fonctionnement du conduit vocal (source d'excitation diverse, impédance aux lèvres, couplage larynx-conduit vocal), et fournissant d'autre part directement des échantillons de parole, paraît être un outil intéressant pour des phonéticiens. Les simulations sur ordinateur basées sur ce modèle (MERMELSTEIN (1969), RICE (1971), RUIZ (1969), TITZE (1973)) laissent de côté les variations de longueur du tractus, paramètre non négligeable, du moins en français, et l'aspect "outil" s'en trouve diminué. L'étude présentée tend donc à intégrer ce paramètre dans cette méthode de synthèse.

I - Equations

Le conduit vocal est assimilé à un tube acoustique déformable transversalement et longitudinalement, et découpé en une succession de n tubes élémentaires de 1 cm.

I-1 Propagation

Soit Δt le temps de transit par l'onde de vitesse volumique dans un tube élémentaire.

$V(+,n,k)$ la vitesse volumique se propageant dans le sens positif (axe orienté positivement de la glotte vers les lèvres), incidente sur la jonction des tubes n, n+1 à l'instant $k\Delta t$.

$V(-,n+1,k)$ la vitesse volumique se propageant dans le sens négatif, incidente sur la même jonction.

On tient compte des pertes en supposant que la vitesse volumique s'atténue d'un facteur α durant sa propagation dans un tube élémentaire ; il vient :

$$v(+,n+1,k+1) = \alpha\,(1+R_n)\;V(+,n,k) - R_n V(-,n+1,k)$$

$$v(-,n,k+1) = \alpha\,(1-R_n)\;V(-,n+1,k) + R_n V(+,n,k)$$

avec $R_n = \dfrac{A_{n+1} - A_n}{A_{n+1} + A_n}$ coefficient de réflexion de la jonction n, n+1

Ai section d'aire du tube i

A des tubes de 1 cm de long et pour une vitesse de propagation de 350 m/s correspond une fréquence d'échantillonnage de 17,5 KHz.

I-2 Terminaison aux lèvres

On considère que le conduit vocal se termine sur un tube d'aire Ar très supérieure à l'aire aux lèvres et de longueur infinie (RICE 1971)
le dernier coefficient de réflexion vaut donc :

$$Rt = \frac{Ar - Al}{Ar + Al}$$

$Ar = 400\ cm^2$
Al = aire aux lèvres

La pression à une distance r de la tête est donnée par la dérivée temporelle de la vitesse aux lèvres (FLANAGAN J.L., 1972)

I-3 Le larynx

On considère que la glotte a une impédance infinie ; on ajoute à l'onde réfléchie un échantillon d'onde glottale. Celle-ci est obtenue à la fréquence et l'amplitude désirées par un oscillateur, au sens de MUSIC V (MATHEWS (1969), cf. annexe)); la forme d'onde utilisée est une de celles données par ROSENBERG (1968) (forme d'onde dite polynominale).

II - DYNAMIQUE DU CONDUIT VOCAL

Nous appelons configuration un ensemble de n sections d'aire (Ci, i croissant de la glotte aux lèvres) où n est la longueur du conduit en cm.

II-1 Mouvements transversaux

On suppose qu'entre deux configurations cibles, chaque aire évolue linéairement par rapport aux temps. Les "clicks" durant les transitions - dus aux erreurs inhérentes à ce modèle, exact seulement pour des configurations statiques - ont été éliminés par une interpolation suffisamment fine entre les cibles ; cette technique paraît plus simple que celle utilisée par RICE (1971) (attente de la fin de l'impulsion glottale) ou RUIZ (1971) (reformulation des équations aux jonctions).

II-2 Mouvements longitudinaux

Ces mouvements, essentiellement localisés aux lèvres et au larynx, peuvent atteindre 2 à 3 cm d'amplitude.

En première approximation, les variations sont reportées en totalité aux lèvres. La discrétisation spatio temporelle choisie implique une variation brusque de la longueur, égale à 1 cm ; on admet que cette variation a lieu au milieu de la transition (Fig. 1.1). Les figures 1.2, 1.3 montrent les valeurs cibles de l'aire aux lèvres du tube rajouté ou enlevé.

Les figures 2.1 2.2 schématisent la situation des échantillons vitesse, pression.

Considérons le cas d'un allongement (fig. 2.1) ; une difficulté est apparue pour la valeur de v (-,n+1) à t = t/2 : la prendre nulle (valeur justifiée a priori) née en fait un "clic" prononcé, nous avons en fait considéré la situation à t/2 comme stabilisée, et donné à V (-,n+1) la valeur qu'elle aurait eu si l'allongement s'était produit à t/2 - 2 Δt, soit :

$$v(-,n+1) = \alpha\ R_{n+1}\ \ \alpha(1+R_n)\ \ V(+,n)$$

avec $Rn+1 = (400 - A_{n+1})/(400 + A_{n+1})$ et $R_n = 0 \quad (A_n = A_{n+1})$

Pour un raccourcissement, le même raisonnement conduit à v(-,n+1) = 0 (fig.2.2) Ces valeurs "fictives" donnent des transitions exemptes de bruit.

III - PROGRAMMES

Ils sont au nombre de 4, en FORTRAN IV.

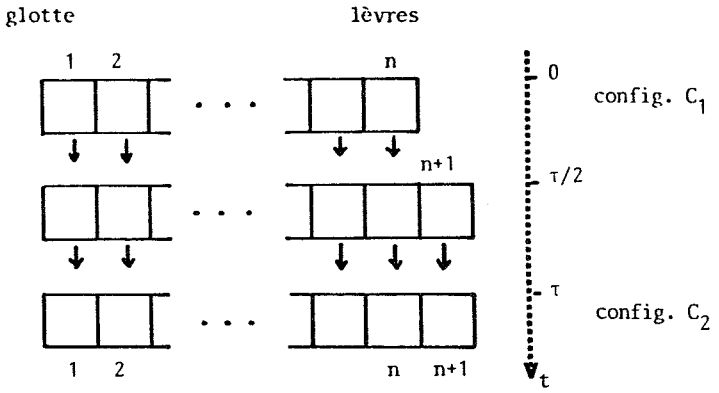


Fig. 1.1 Allongement
Lengthening

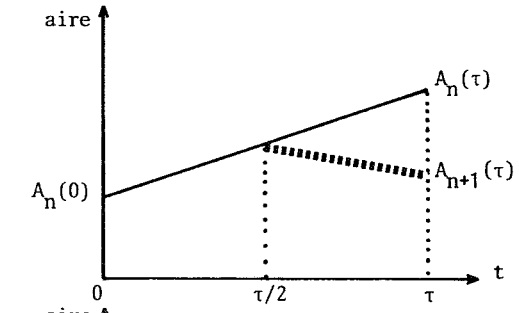


Fig. 1.2

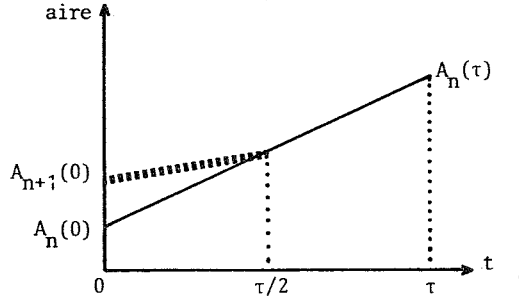


Fig. 1.3

évolution de A_{n+1}

évolution de A_n

Evolution des aires aux lèvres
Lip-area evolution

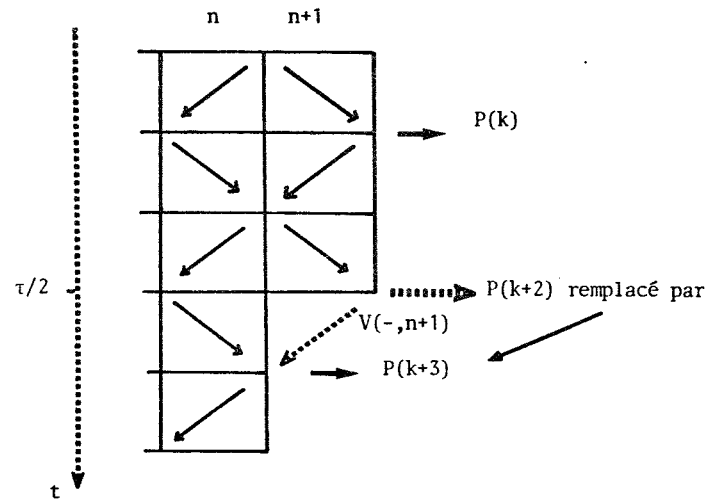


Fig. 2.1

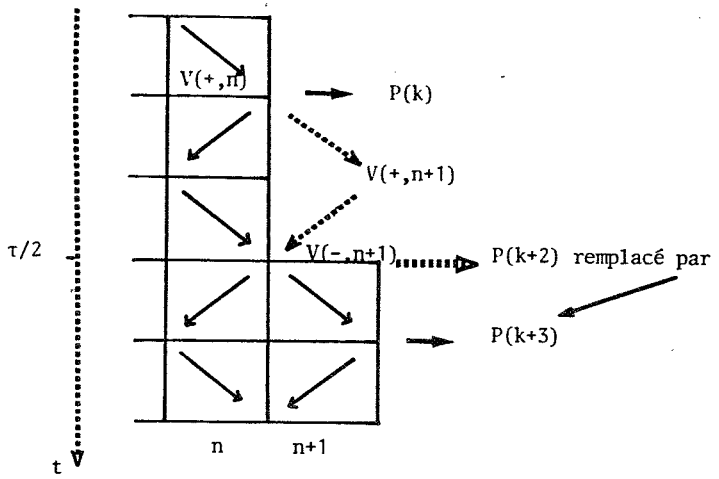


Fig. 2.2

Echantillons vitesse/pression lors d'un changement de longueur
Volume velocity, pressure samples during a change in length

- calcul d'une période d'onde glottale
- génération de valeurs de Fo (par calcul ou utilisation) d'un système de détection de la mélodie (Bernard TESTON et Mario ROSSI, 1977) et d'amplitude.
- interpolation des fonctions d'aires.
- calcul des échantillons de parole (avec lecture préalable du chronométrage).

IV - RESULTATS ET CONCLUSION

Les fonctions d'aires utilisées sont celles données par MRAYATI, M., GUERIN, B. (1976). On peut comparer sur le Tableau N° 1 les valeurs de formants obtenues(B) avec celles (A) données par MRAYATI et GUERIN, et avec celles (C) que nous avons calculées par la méthode de COKER (1968), reprises par ZERLING (1974).

Les séquences synthétisées (exclusivement composées de voyelles orales ; ex. [aijaja jə], [awi e : wi] , etc...) avec utilisation d'une mélodie "naturelle", ont été jugées de bonne qualité. Il s'est avéré que le modèle choisi a permis une intégration simple du changement de longueur ; il devrait être également relativement aisé de simuler séparément les mouvements longitudinaux des lèvres et du larynx.

Parallèlement, un découpage "naïf" de l'évolution des fonctions d'aires, en cibles d'une part, transitions linéaires d'autre part, paraît donner de bons résultats sur le plan perceptif, approche semblant justifiée par OLIVE et SPICKENAGEL (1976).

* Ce travail a fait l'objet d'une thèse 3ème cycle d'acoustique (janvier 1977).

ANNEXE

. OSCILLATEUR MUSIC V

Soit TAB un tableau de N échantillons (512 dans notre cas) d'une période d'amplitude normalisée à 1, du signal périodique à générer

f la fréquence fondamentale voulue
F la fréquence d'échantillonnage

le principe consiste à parcourir ce tableau circulairement avec un pas P_j fonction de f

$P_j = f.512/F$ pas associé à l'échantillon S_j, généré par :

$I_j = I_{j-1} + P_j$

$S_j = A_j \ TAB((I_j) \bmod 512)$

A_j : amplitude voulue

REFERENCES

COKER, C.H., 1968, Speech synthesis with a parametric articulatory model Proc. Kyoto Speech Symp, Kyoto A41-A46.

FLANAGAN, J.L., 1972, Speech analysis synthesis and perception, Springer Verlag, Berlin.

FLANAGAN, J.L., ISHIZAKA, K., & SHIPLEY, K.L., 1975, Synthesis of speech from a dynamical model of vocal words and vocal tract. The Bell Syst. Tech. J, 54, n° 3, pp. 485-506.

KELLY, J.L., LOCHBAUM, J.R. & cc., 1962, Speech synthesis, Proc. 4th Sca Congress Copenhague.

MATHEWS, M.V., 1969, The technology of computer music, MIT Press Cambridge.

MERMELSTEIN, P., 1969, Computer simulation of articulatory activity in speech production, Joint conf. on Artil-Intelligence, Washington.

MRAYATI, M., GUERIN, B., 1976, Etude des caractéristiques acoustiques des voyelles orales françaises par simulation du conduit vocal avec pertes. Revue d'Acoustique, 36, pp. 18-32.

OLIVE, J.P. & SPICKENAGEL, N., 1976, Speech resynthesis from phoneme related parameters, JASA 59, p. 993-996.

RICE, L., 1971,A new line analog speech synthesizer for the PDP 12, working papers in phonetics, 17, pp. 58-75.

ROSENBERG, A.E., 1971, Effect of glottal pulse shape on the quality of natural vowels, JASA, 49, pp. 583-588.

RUIZ, P., MERMELSTEIN, p., 1969, Speech generator with a computer simulated vocal tract, JASA, 110 (A).

RUIZ, P., 1971, A digital simulation of the time variying vocal tract, JASA 49, p. 123 (A).

TESTON, B., & ROSSI, M., 1977, Un système de détection automatique des éléments prosodiques, 8èmes Journées d'Etude sur la parole, Aix-en-Provence.

TITZE, J.R., 1973, The human vocal cords : a mathematical model part I, Phonetica 28, n° 3-4, p. 129-170.

ZERLING, J.P., 1974, Etude d'un programme d'ordonateur permettant de calculer les trois premiers formants à partir de la fonction d'aire. Travaux de l'institut de Phonétique de Nancy, vol. 1, pp. 257-291.

TABLEAU 1

Phonèmes		[ɑ]	[a]	[e]	[ɛ]	[o]	[ɔ]	[ø]	[œ]	[i]	[y]	[u]
Calculs de MRAYATI & GUERIN (A)												
	F_1	676	678	418	458	402	540	384	535	244	280	296
	F_2	1 192	1 300	2 111	1 881	797	1 007	1 592	1 411	2 250	1 846	783
	F_3	2 208	2 583	2 665	2 584	2 417	2 480	2 461	2 269	3 082	2 230	2 286
Nos résultats (B)												
	F_1	685	685	380	450	410	545	380	550	220	270	275
	F_2	1 290	1 430	2 220	1 970	820	1 160	1 800	1 500	2 255	1 980	770
	F_3	2 270	2 666	3 110	2 830	2 560	2 600	2 530	2 390	3 260	2 290	2 290
Calculs d'après ZERLING (C)												
	F_1	699	705	399	448	378	525	346	529	195	231	248
	F_2	1 304	1 475	2 226	1 995	875	1 175	1 786	1 542	2 257	1 986	770
	F_3	2 313	2 759	3 201	2 856	2 530	2 672	2 572	2 432	3 271	2 348	2 293

Moyenne des écarts

- B comparé à A
  - sur F_1 = 3,9 %
  - sur F_2 = 6,2 %
  - sur F_3 = 5 %
- B comparé à C
  - sur F_1 = 7 %
  - sur F_2 = 1,5 %
  - sur F_3 = 1,7 %

10èmes JOURNÉES D'ÉTUDE SUR LA PAROLE

GRENOBLE - 30 MAI - 1ER JUIN 1979

DESCRIPTION D'UN SYSTEME DE SYNTHESE DE LA PAROLE.

APPLICATION A LA SYNTHESE DE SONS ISOLES.

J.L. GARNELL

Laboratoire CERFIA - Université Paul Sabatier
118, route de Narbonne - 31077 TOULOUSE CEDEX

RESUME

Un système de synthèse de la parole a été réalisé sur mini-ordinateur Télémécanique T1600.

Il comprend un synthétiseur numérique à formants de type parallèle commandé par programme ainsi que divers utilitaires permettant d'étudier le signal vocal naturel ou synthétique.

Le choix de cette forme de synthèse a été fait afin de pouvoir étudier l'influence relative des paramètres acoustiques en vue d'une aide à l'analyse et la reconnaissance de la parole.

DESCRIPTION OF A SPEECH SYNTHESIS SYSTEM SYNTHESIS OF ISOLATED UTTERANCES

J.L. GARNELL

SUMMARY

A speech synthesis system has been developped on a minicomputer (Téléméca-nique T1600).

This system includes :

- a parallel type digital formant synthesizer driven by programs
- a study and analysis software.

The aim of the related study is to weight the relative influence of the acoustic parameters, as a help to speech analysis and recognition.

10èmes JOURNÉES D'ÉTUDE SUR LA PAROLE
GRENOBLE - 30 MAI - 1er JUIN 1979

DESCRIPTION D'UN SYSTEME DE SYNTHESE DE LA PAROLE APPLICATION A LA SYNTHESE DE SONS ISOLES.

J.L. GARNELL

I. INTRODUCTION

Cette forme de synthèse vise à reconstituer un fragment de parole à partir de ses principaux paramètres acoustiques. Le principe en est le suivant : pour produire des sons voisés, un générateur simule les vibrations des cordes vocales et envoie une série d'impulsions dans des circuits de résonance dont la fonction de transfert est équivalente à celle du conduit vocal. Les sons de frictions sont produits par un générateur de bruit blanc d'amplitude variable, filtré.

Deux types de synthétiseurs à formants ont ainsi été utilisés :

- du type série ou en cascade (circuits de résonance successifs). Il représente le modèle le plus exact de la fonction vocale (Flanagan 1957), mais il ne permet pas de contrôler séparément l'amplitude des formants.

- du type parallèle où chaque formant est contrôlé en fréquence et en intensité. Il permet une bonne réalisation des sons complexes (plosives, constrictives sonores).

Holmes a montré qu'avec un synthétiseur de ce type il pouvait générer de la parole synthétique indiscernable de la parole naturelle (Holmes 1973).

Le synthétiseur réalisé dans cette étude est de ce second type.

But de l'étude

Le système de synthèse a été développé afin d'étudier l'influence relative des divers paramètres acoustiques (fréquences de formants, amplitudes, bandes passantes, forme de l'onde glottale etc...) en vue d'une aide à l'analyse et la reconnaissance de la parole.

Dans un premier temps, nous n'avons pas cherché à réaliser un système produisant de la parole continue en temps réel.

Moyens utilisés

L'ordinateur utilisé est un télémécanique T1600 de 64 k mots mémoire, doté de périphériques visuels : disque, bande, visu graphique et d'une sortie vocale : coupleur, convertisseur ADIDA relié à un magnétophone.

La configuration est la suivante :

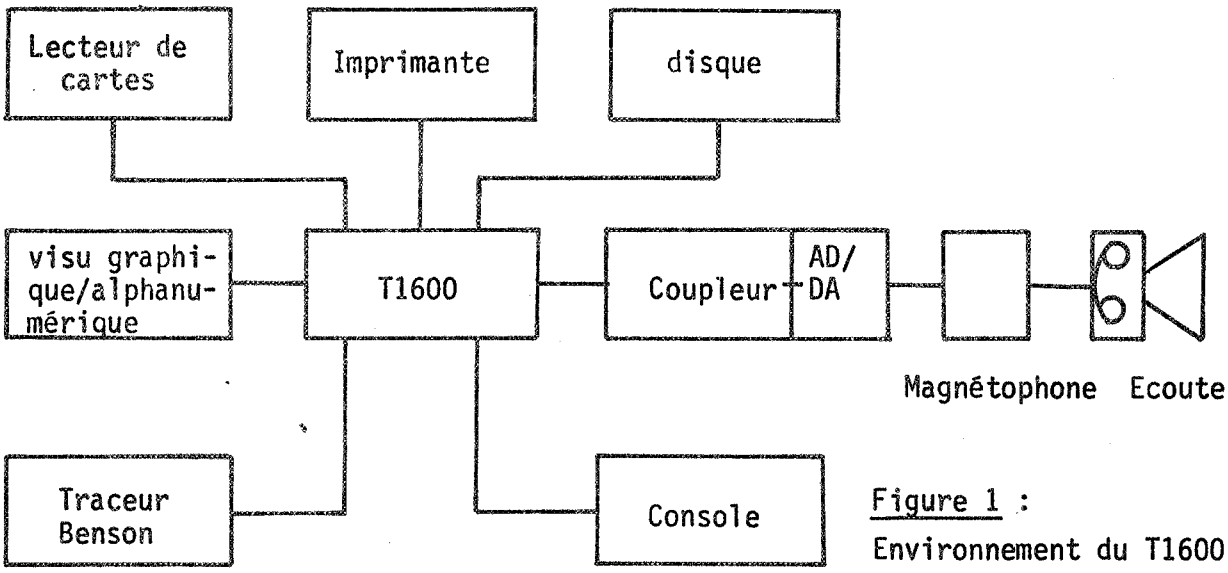


Figure 1 :
Environnement du T1600

II. DESCRIPTION DU SYSTEME DE SYNTHESE

1) Le synthétiseur :

Il est de type parallèle (fig. 2). Un générateur d'onde glottale variable attaque quatre filtres numériques du second ordre simulant le conduit nasal (1 filtre) et le conduit vocal (3 filtres).

Un générateur de bruit attaque deux filtres symétriques du second ordre, donnant les formants de bruit. Les sons simultanément voisés et bruités sont produits en modifiant le système d'excitation par superposition de l'onde glottale à un bruit d'amplitude variable.

Les paramètres de commande sont au nombre de 15 : fréquences et amplitudes des 4 formants de voisement et des deux formants de bruit, fréquence du fondamental, amplitude du bruit dans le cas de sons voisés et bruités, un indice donnant le caractère du son émis : voisé, bruité, ou les deux simultanément.

Les bandes passantes des formants sont fixées par programme et peuvent être modifiées par dialogue.

Les paramètres sont fournis au synthétiseur en synchronisme avec le fondamental dans le cas de sons voisés, et 117 fois/seconde pour les sons bruités ou les silences. Le signal est recalculé point par point à raison de 15000 points par seconde(fréquence d'échantillonnage choisie).

2) La génération des paramètres fournis au synthétiseur :

L'évolution du fondamental et des formants en fréquence et en amplitude, est calculée à partir de courbes de base : sinusoïdes, portions d'exponentielles, droites. On peut visualiser ou imprimer ces schémas représentant des "stratégies de synthèse" (fig. 3).

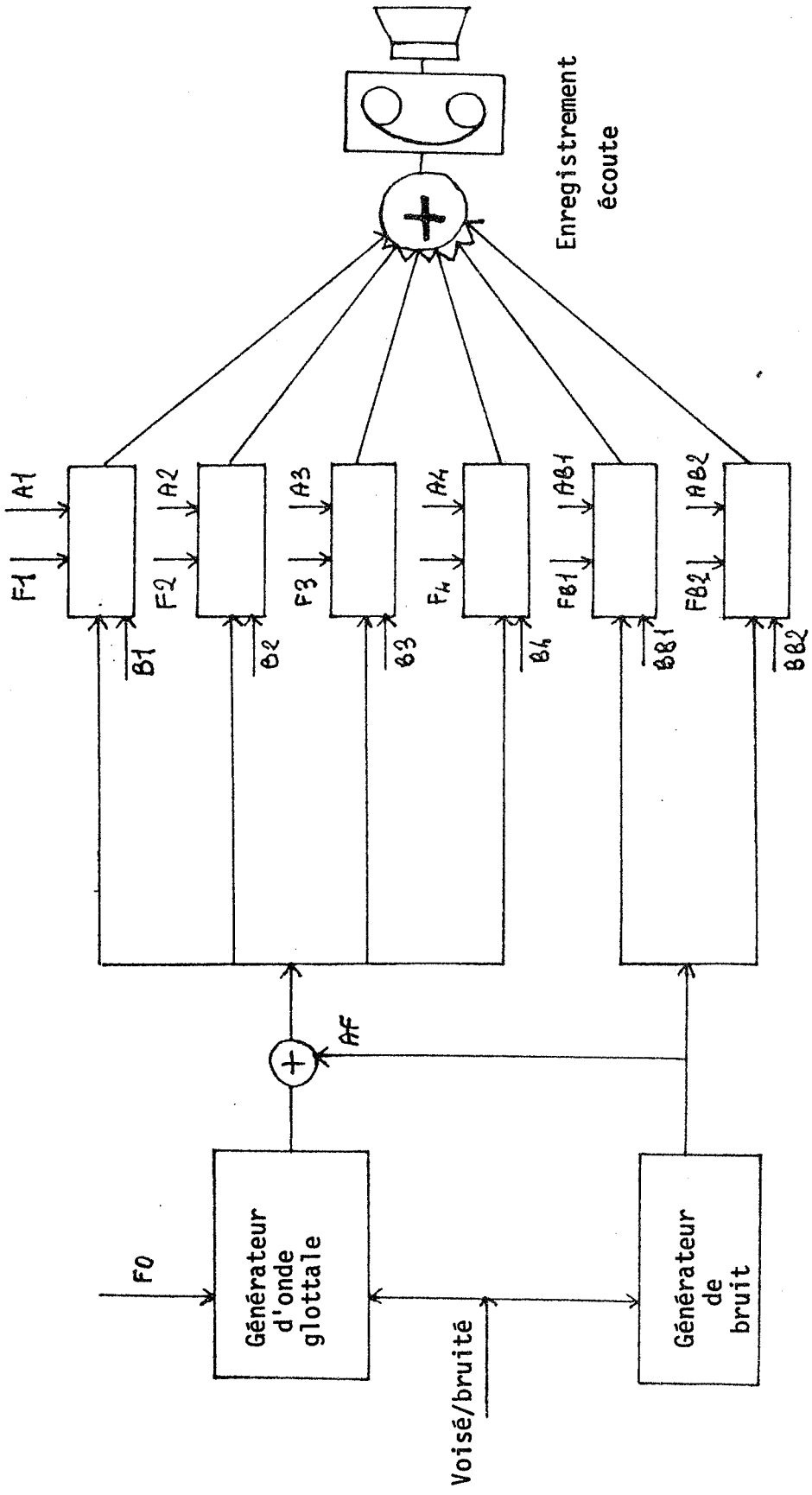


Figure 2 : Structure du synthétiseur

Les paramètres ainsi générés sont stockés dans une table fournie ensuite au synthétiseur. Les programmes réalisés sont conversationnels et permettent l'accès à toutes les tables de paramètres afin d'en modifier le contenu.

3) Autres traitements réalisables :

Divers programmes ont été mis au point principalement par P.Y. CAZENAVE et permettent de :

- faire l'acquisition/restitution d'une portion quelconque de parole,
- analyser une portion de signal échantillonné stocké sur disque par le modèle de cochlée développé par J. CAELEN (Caëlen 1974) et tracer les sonagrammes correspondants,
- visualiser, écouter, inverser, tronquer, tracer un signal réel ou synthétique stocké sur disque.

III. RESULTATS OBTENUS

Nous avons synthétisé différents sons isolés de bonne qualité : séries de voyelles orales, voyelles nasales (sans zéros), semi voyelles, liquides à l'initiale.

Différentes études sont en cours concernant les fricatives sourdes, sonores (qualité de la source de bruit), les plosives, les consonnes nasales, l'influence de la forme de l'onde glottale.

Cette étude fait l'objet d'une thèse de docteur ingénieur (à paraître courant 79) et où les résultats détaillés seront donnés.

IV. CONCLUSION

Nous avons cherché à développer un outil à la fois d'une grande précision et souplesse d'utilisation nous permettant d'acquérir une pratique en synthèse de la parole et de mener diverses études concernant les paramètres acoustiques.

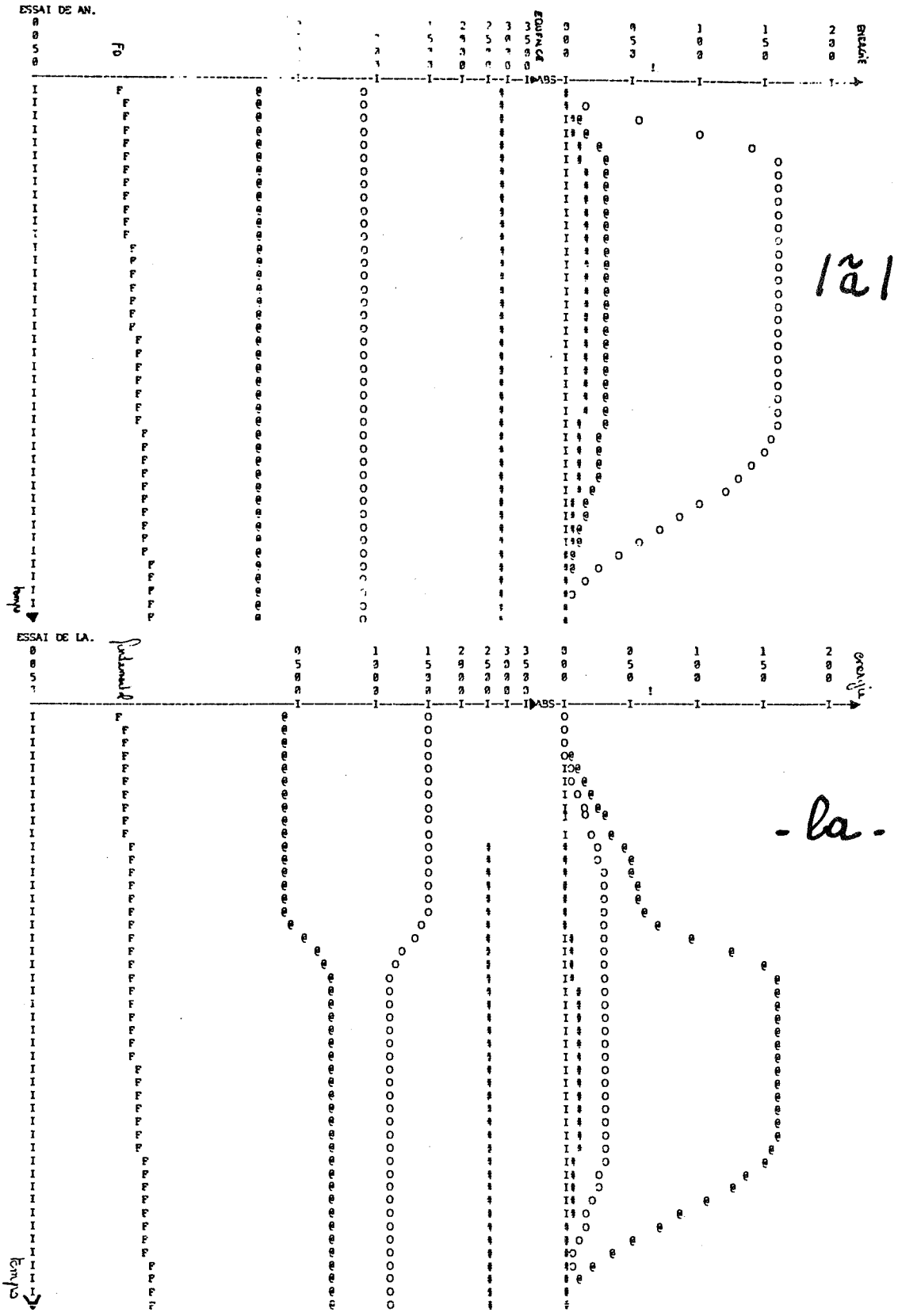


Figure 3 : Exemples de stratégies de synthèse

BIBLIOGRAPHIE

(1) J. CAELEN (1974)
Un modèle mathématique de cochlée et son application à l'analyse de la parole.
Thèse de docteur ingénieur. Toulouse 1974

(2) CARRE J. (1971)
Contribution aux études sur l'analyse et la synthèse de la parole.
Rôle et importance des formants.
Thèse d'état. Grenoble

(3) CHAFCOULOFF N. (1976)
25 années de recherche en synthèse de la parole.
Editions du CNRS. 1976

(4) FLANAGAN J.L. (1972)
Speech analysis, synthesis and perception.
Berlin, Heidelbery, New-York, Springer Verlay

(5) HOLMES J.N. (1973)
Influence of glottal wareforms on the naturalness of speech from a parallel formant synthetiser.
IEEE, TAE, AV21, 1973, Vol.3, pages 295-305

(6) PERENNOU G., CAELEN J. (1975)
Localisation des voyelles dans le plan (F_1,F_2). Application à la reconnaissance de la parole.
9ème J.E.P. Lannion. Pages 223-232

(7) RABINER L.R. (1969)
A model for synthesizing speech by rule.
IEEE, TAE, AU17, Vol.1, pages 7-13

10èmes JOURNÉES D'ÉTUDE SUR LA PAROLE

GRENOBLE - 30 Mai - 1er Juin 1979

DESCRIPTION D'UNE UNITE DE REPONSE VOCALE DE DONNEES NUMERIQUES DECIMALES

OUAKNINE Maurice et TESTON Bernard

Laboratoire de PsychoPhysiologie
Université de Provence

Institut de Phonétique
Université de Provence

RESUME

Le système que nous décrivons, est une unité de réponse vocale exclusivement réalisée pour synthétiser des chiffres et des nombres de 1 à 999. Son utilisation est réservée à des applications multimétriques. Elle est connectée à tout appareil disposant des données numérisées en décimal codé binaire. Elle est constitué d'un décodeur lexical, et d'une mémoire de 23 segments de longueur variable qui permettent, présentés dans un ordre adéquat, une synthèse par concaténation. La définition est de 4 bits, le nombre d'échantillons de 3.300 par seconde. On peut envisager très simplement une extension jusqu'aux milliers ainsi qu'une mémoire additionnelle pour stocker quelques unités de mesure.

DESCRIPTION OF A VOCAL RESPONSE UNIT FOR DECIMAL DATA

(1) OUAKNINE Maurice et (2) TESTON Bernard

SUMMARY

The system we describe is a vocal response unit designed sedely for the synthesis of figures and number from 1 to 999, which has been conceived with a view to multemetric application.

It can be connected to any apparatus supplying numerised data in the form of binary coded decimal. It comprises a lexical decoder and a memory of 23 segments of variable length which after appropriate ordering allows synthesis by concatenation.

The definition is of 4 bits and the sampling rate 3.300 per second. The system can easily be extended to bigher numbers and on additionnal memory could used to stock the units of measurement.

(1) Laboratoire de Psychophysiologie
Université de Provence.
(2) Institut de Phonétique
Université de Provence.

10èmes JOURNÉES D'ÉTUDE SUR LA PAROLE
GRENOBLE - 30 MAI - 1ER JUIN 1979

DESCRIPTION D'UNE UNITE DE REPONSE VOCALE DE DONNEES NUMERIQUES DECIMALES.

(1) OUAKNINE Maurice et (2) TESTON Bernard

INTRODUCTION

L'unité de réponse vocale que nous nous proposons de décrire a été développée dans un but bien précis ; transmettre vocalement la mesure d'une grandeur physique par l'intermédiaire d'un multimètre ou tout autre instrument disposant d'une sortie en système décimal codé binaire.

L'application originale d'un tel synthétiseur était, de donner à un sujet, l'information de ses scores au cours d'expériences de psychologie sur la vision, cette information ne devant pas être lue sous peine de perturber les tests. Au début, un opérateur donnait vocalement le résultat des scores aux sujets par l'intermédiaire d'un interphone. Mais, les erreurs de lecture, le retard entraîné par cet intermédiaire humain (tout le reste de la manipulation est automatisé) et les perturbations du sujet au plan de sa concentration, nous ont fait envisager très vite, l'utilisation d'une unité de réponse vocale capable de transmettre toutes les valeurs numériques comprises entre 1 et 999, nécessaires à notre manipulation expérimentale.

La limitation du lexique, ainsi que quelques réalisations industrielles (Master Specialities Co Model 1.700), nous ont fait choisir immédiatement, comme technique de synthèse, la concaténation directe de mots stockés au préalable dans des mémoires numériques.

II - PRINCIPE DE SYNTHESE

La concaténation de segments ou préalablement enregistrés, et découpés en autant d'unité, qu'il apparaissait nécessaire n'a jamais donné de bons résultats pour synthétiser de la parole (CHAFCOULOFF 1976). Pour la langue anglaise, HARRIS (1953) puis WANG et PETERSON (1958) ont avancé le plus dans cette direction, bloquée dès le départ par des problèmes de transitions entre les segments. Cette technique a été complètement abandonnée sauf pour certaines expériences bien particulières (AUTESSERRE et DI CRISTO 1971). Elle a été remplacée avantageusement par les techniques de synthèse par règle. A propos du Français, nous ne connaissons pas de travaux de ce genre dans le passé. Cependant, une tentative actuelle semble se développer sans que nous en ayons connaissance sous forme de publications, mais cela ne tardera-t-il pas malgré le peu d'avenir du système et la médiocrité des résultats.

Si l'on restreint le vocabulaire à concatener à des chiffres et des nombres, on s'aperçoit que l'on n'a pas de problèmes de transition particuliers ; ni de

(1) Laboratoire de PsychoPhysiologie
Université de Provence.
(2) Institut de Phonétique
Université de Provence.

liaison, ni prosodique. Pour compter de 1 à 999 on a besoin que de 23 segments de concaténation. Cette méthode est donc bien adaptée pour résoudre notre problème.

III - DECODEUR LEXICAL

Les 23 segments que nous devons utiliser pour compter de 1 à 999 sont : 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 20, 30, 40, 50, 60, 100 et la liaison particulière ET (Tableau N° 1).
Un décodeur lexical particulier est nécessaire pour une bonne numérotation car le système lexical du français des dizaines est particulièrement exceptionnel (dans le sens où il existe de nombreuses exceptions à la règle générale). Sur ce plan, l'énumération en langue anglaise ou allemande est plus simple. Il nous faut réaliser de nombreux masquage selon les ordres de sortie des groupes de chiffres. (Tableau 2)
Le décodage du nombre (digit) en rapport au nombre effectivement prononcé est parfois complexe (Exemple : 97 → 4, 20, 10, 7) et le décodeur lexical que nous avons voulu le plus simple possible nous a occasionné quelques difficultés. Plutôt qu'une explication confuse, nous reportons le lecteur au tableau 2 pour en saisir le principe de fonctionnement, ainsi qu'à la figure 3.

Les informations numériques, images des grandeurs auparavant mesurées, représentée en décimal codé binaire, sont envoyées en unités, dizaines et centaines, dans sept registres d'adresse sur 5 bits et un registre à décalage de 1 bit pour l'inhibition des chiffres qui ne doivent pas être émis vocalement (Figure 4). Le code binaire sur 5 bits des différents segments ou groupe de chiffres est donné au Tableau 5.

IV - SYNTHETISEUR

Il est essentiellement constitué par une mémoire de 23 mots d'un nombre de bits variables, plus ou moins important selon la durée du segment.
Cette mémoire morte est constituée par 23 boitiers de 4 x 1.024 bits.

Les adresses des mots ainsi que le balayage des bits successifs constituant ces mots sont réalisés au moyen de compteurs - décompteurs synchrones, comparateurs et décodeurs.

La sortie des informations se fait sur 4 bits qui attaquent un convertisseur numérique - analogique suivi d'un filtre passe bande dissymétrique (150 Hertz - 36 dB/octave - 2,5 K Hertz - 96 dB/octave). L'horloge de lecture bat à la fréquence de 3.300 Hertz (Figure 6).

V - PREPARATION DU CORPUS DE SYNTHESE

La durée des différents chiffres et nombres étant très variable (Tableau 7) nous avons enregistré en chambre sourde, 10 locuteurs choisis dans notre entourage. Les 23 séquences ont été prononcées naturellement sans accentuations particulières, recto tono et bien séparées par un important silence (2 à 3 secondes). Ceci fait, nous avons réalisé une moyenne sur les différents segments, et nous avons choisi le locuteur qui se rapprochait le plus de cette moyenne mais aussi dont la voix restait esthétique et intelligible malgré les traitements que nous lui faisions subir.

Le cadrage à la longueur exacte des segments a été effectué au moyen d'un adressage numérique sur une RAM après conversion sur 8 bits du signal analogique. La diminution n'a été que de quelques bits afin d'assurer un bon cadrage en fonction du nombre de bits disponible dans la mémoire (Tableau N° 8). La translation cadrée en durée sur les PROM a été réalisée manuellement sur 4 bits. L'opération

s'est avérée longue et fastidieuse, mais nous ne pouvions pas utiliser notre calculateur pour ce premier essai. La fréquence d'échantillonnage a été choisie à la valeur de 3.300 Hertz après avoir tenu compte de la durée totale des segments et de notre capacité de mémoire maximale.

Avant de mémoriser définitivement les segments sur les PROM, nous avons testé différents procédés dans le but d'améliorer la qualité de la voix restituée dans les conditions de compression d'information précisées précédemment.

Tout d'abord, nous avons essayé de doubler la fréquence d'échantillonnage pour augmenter la largeur de bande du spectre des consonnes constructives surtout dans 6 (six) et 10 (dix). Les essais furent concluants et l'amélioration de la distinction de ces deux chiffres très sensibles. Cependant, le fait de devoir, outre la fréquence d'échantillonnage, changer la valeur de la fréquence de coupure du filtre passe bas de sortie, ainsi que le doublement de la capacité mémoire de ces deux chiffres nous ont dissuadé d'employer cette solution. Toujours dans le même sens, nous avons également testé un système d'échantillonnage continuement variable en fonction de la fréquence des signaux à échantillonner. Ce système est très efficace pour optimiser l'encombrement des mémoires, mais il est compliqué à mettre en oeuvre.

Pour augmenter la dynamique de l'amplitude des signaux, nous avons essayé de comprimer puis d'expanser le signal au moyen de circuits appropriés analogiques, mais nous ne les avons pas retenus car ils compliquaient par trop le système, malgré une nette amélioration de la dynamique de synthèse.

VI - RESULTATS

Les résultats que nous obtenons, avec les durées du Tableau 8, 4 bits de définition du signal et 3.300 Hertz de fréquence d'échantillonnage sont satisfaisants. Le taux d'erreur de compréhension des valeurs numériques ainsi transmises aux sujets est inférieur à 3 %. Dans la version actuelle, nous avons supprimé le ET de liaison (Exemple : 20 et 1, 30 et 1 etc...). Cette simplification importante ne semble pas perturber la compréhension. Ceci malgré la présence dans les segments de nombreuses erreurs de programmation de la PROM, dues à la manipulation manuelle des données. Le décodeur lexical fonctionne parfaitement et semble ne pas pouvoir être plus simple.

La qualité de la voix ainsi restituée peut être avantageusement améliorée en utilisant pour les conversions AN et NA, des convertisseurs à compression-expension de dynamique (P.M.I. type DAC 76 par exemple) qui viennent d'apparaître sur le marché. On peut envisager avec ces systèmes de 4 bits effectifs, d'obtenir une dynamique de 6 bits. On peut également augmenter la largeur du spectre jusqu'à la bande passante téléphonique, en doublant la capacité des mémoires, car ces dernières deviennent de plus en plus compactes et coûtent de moins en moins cher.

VII - CONCLUSION

Le système de réponse vocale que nous venons de décrire peut être très facilement étendu pour devenir un système complet pour multimètre numérique.

Le décodeur lexical comprend déjà les milliers, ce qui permet de compter jusqu'à 999.999 avec un mot de mémoire supplémentaire (1.000). Pour des applications multimétriques, une mémoire spéciale pour le stockage des unités peut être réalisée très simplement (Hertz - Volts - Décibels - Millibars etc...). De tels systèmes de réponse vocale peuvent être utilisés dans de nombreuses occasions. Dans certaines conditions de travail difficile, où toute l'attention d'un sujet

est prise par une tache principale nécessitant la connaissance de nombreux paramètres (pilotage d'hélicoptères - interventions chirurgicales, maintenance technique dans de mauvaises conditions d'accessibilité etc... etc...). On peut également envisager ainsi, la scrutation de données centralisées au moyen d'une simple ligne téléphonique à grande distance.

Dès maintenant, un système de 4 bits de définition avec compression expension de dynamique, comptant de 1 à 999.999 et d'une bande passante de 100 à 3.300 Hertz peut être réalisé avec moins d'une vingtaine de boîtiers de C.I. logiques.

On peut envisager même un circuit complexe contenant sur la même puce tout le synthétiseur de grandeur numérique décimale. Le constructeur pourrait à la demande proposer trois types de voix ; homme, femme, et type "Aéroport" pour horloge parlante publique par exemple. Le marché d'un tel composant existe dès maintenant. Son coût de production devra pourtant être comparé à des systèmes de synthèse à prédiction linéaire qui viennent de sortir sur le marché nord américain (WIGGINS, R. 1978). Si l'on se place strictement dans notre application, il nous semble que la synthèse par concaténation a des chances de se développer.

10^1 \ 10^0	0	1	2	3	4	5	6	7	8	9
0	0	1	2	3	4	5	6	7	8	9
1	10	11	12	13	14	15	16	10 → 7	10 → 8	10 → 9
2	20	20 et 1	20 → 2	20 → 3	20 → 4	20 → 5	20 → 6	20 → 7	20 → 8	20 → 9
3	30	30 et 1	30 → 2	30 → 3	30 → 4	30 → 5	30 → 6	30 → 7	30 → 8	30 → 9
4	40	40 et 1	40 → 2	40 → 3	40 → 4	40 → 5	40 → 6	40 → 7	40 → 8	40 → 9
5	50	50 et 1	50 → 2	50 → 3	50 → 4	50 → 5	50 → 6	50 → 7	50 → 8	50 → 9
6	60	60 et 1	60 → 2	60 → 3	60 → 4	60 → 5	60 → 6	60 → 7	60 → 8	60 → 9
7	60 → 10	60 → 11	60 → 12	60 → 13	60 → 14	60 → 15	60 → 16	60 → 17	60 → 18	60 → 19
8	4 20	4 → 20 → 1	4 → 20 → 2	4 → 20 → 3	4 → 20 → 4	4 → 20 → 5	4 → 20 → 6	4 → 20 → 7	4 → 20 → 8	4 → 20 → 9
9	4 20 10	4 → 20 → 11	4 → 20 → 12	4 → 20 → 13	4 → 20 → 14	4 → 20 → 15	4 → 20 → 16	4 → 20 → 10 → 7	4 → 20 → 10 → 8	4 → 20 → 10 → 9

- Composition des nombres de 1 à 99.

Tableau N° 1

- Number composition (1 to 99).

Ordre d'apparition des groupes de chiffres		Commentaire particulier de masquage
A) 1,2,3,4,5,6,7,8,9	10^2	1 masqué
B) 100		0 masqué
C) 4	10^1	Masqué sauf pour 8 + 9
D) 20,30,40,50,60		20 sort pour : 8 + 9 (2 + 8 + 9)
		60 pour 7 (6 + 7)
E) ET → $(n)_{10}{}^1$ $(1)_{10^\circ}$		Masqué pour $(0)_{10}{}^1$ et $(1)_{10}{}^1$ $(1)_{10^\circ}$
F) 10,11,12,13,14,15,16		Sort pour (1 + 7 + 9) x $(0 + 1 + 2 + 3 + \overline{4 + 5 + 6})_{10^\circ}$ $(1 + 7 + 9)_{10}{}^1$ x $\overline{(7 + 8 + 9)}_{10^\circ}$
G) 1,2,3,4,5,6,7,8,9		Sort lorsque F est masqué

- Ordre de sortie des différents groupes de chiffres avec masquage éventuel.

Tableau N° 2

- Output sequence of different number groups with eventual mask.

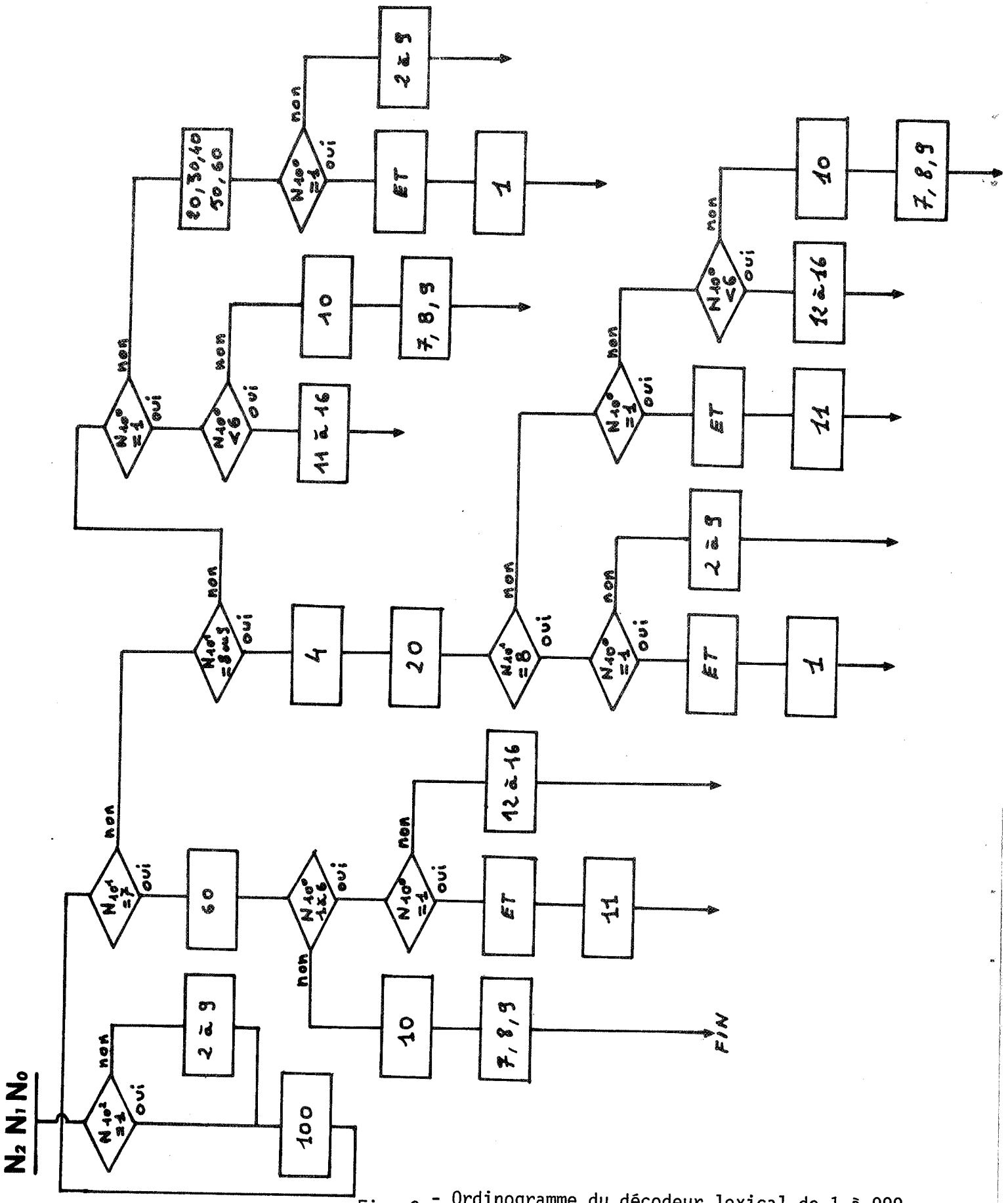


Fig. 3 - Ordinogramme du décodeur lexical de 1 à 999.
- Flowchart of the lexical decoder (1 to 999).

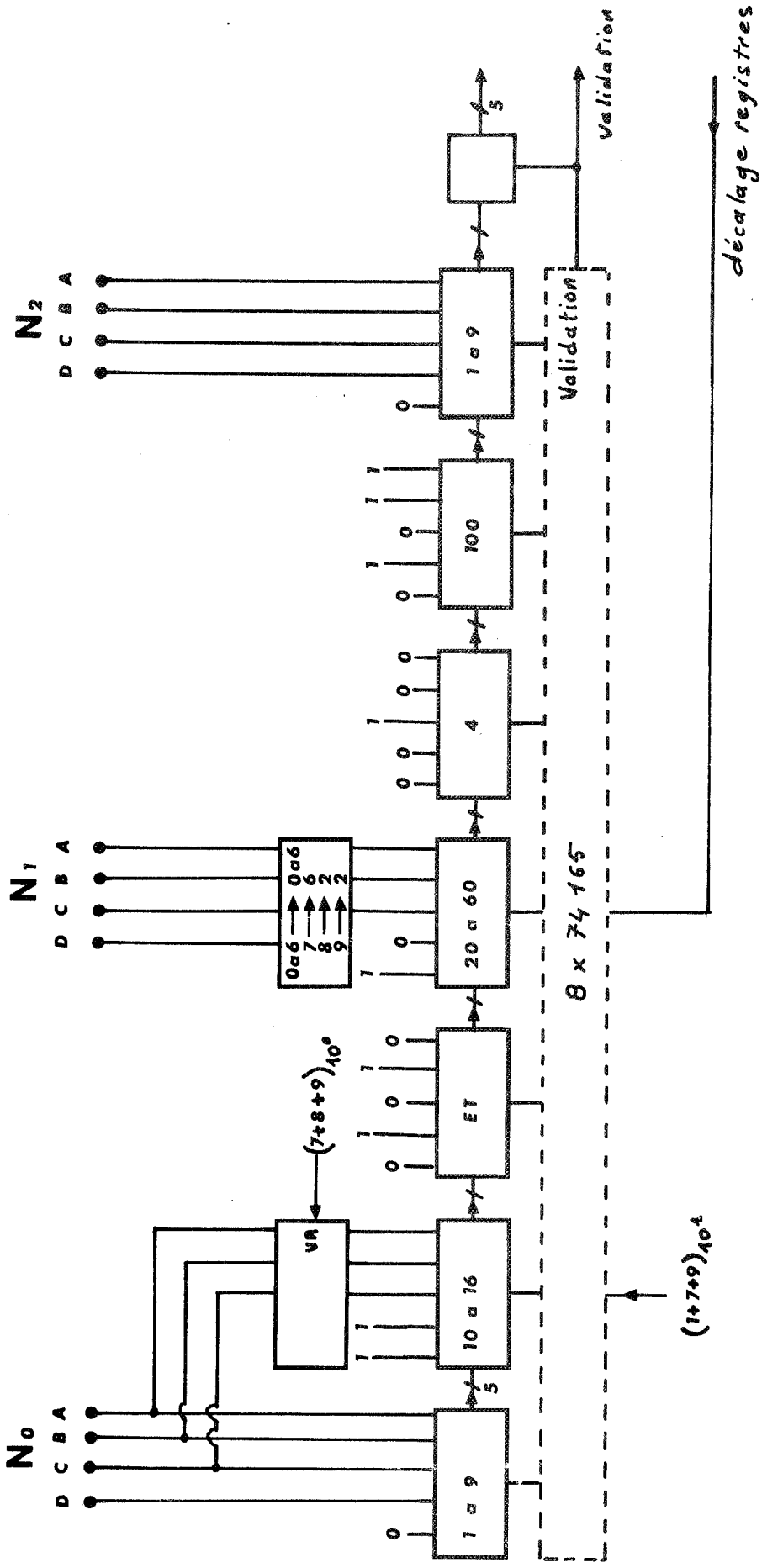


Fig. 4 - Schéma de principe du décodeur lexical.
- General outline of the lexical decoder.

Groupe de chiffre	Chiffre ou Nombre (23)	Code Binaire	
A	1	0 0 0 0 1	
	2	0 0 0 1 0	
	3	0 0 0 1 1	
	4	0 0 1 0 0	
	5	0 0 1 0 1	
	6	0 0 1 1 0	
	7	0 0 1 1 1	
	8	0 1 0 0 0	
	9	0 1 0 0 1	
B	100	0 1 0 1 1	
C	4	0 0 1 0 0	
D	20	1 0 0 1 0	$(2 + 8 + 9)_{10^1}$
	30	1 0 0 1 1	
	40	1 0 1 0 0	
	50	1 0 1 0 1	
	60	1 0 1 1 0	$(6 + 7)_{10^1}$
E	ET	0 1 0 1 0	

- Code binaire des différents segments.

Tableau N° 5

- Binary code of the different segments.

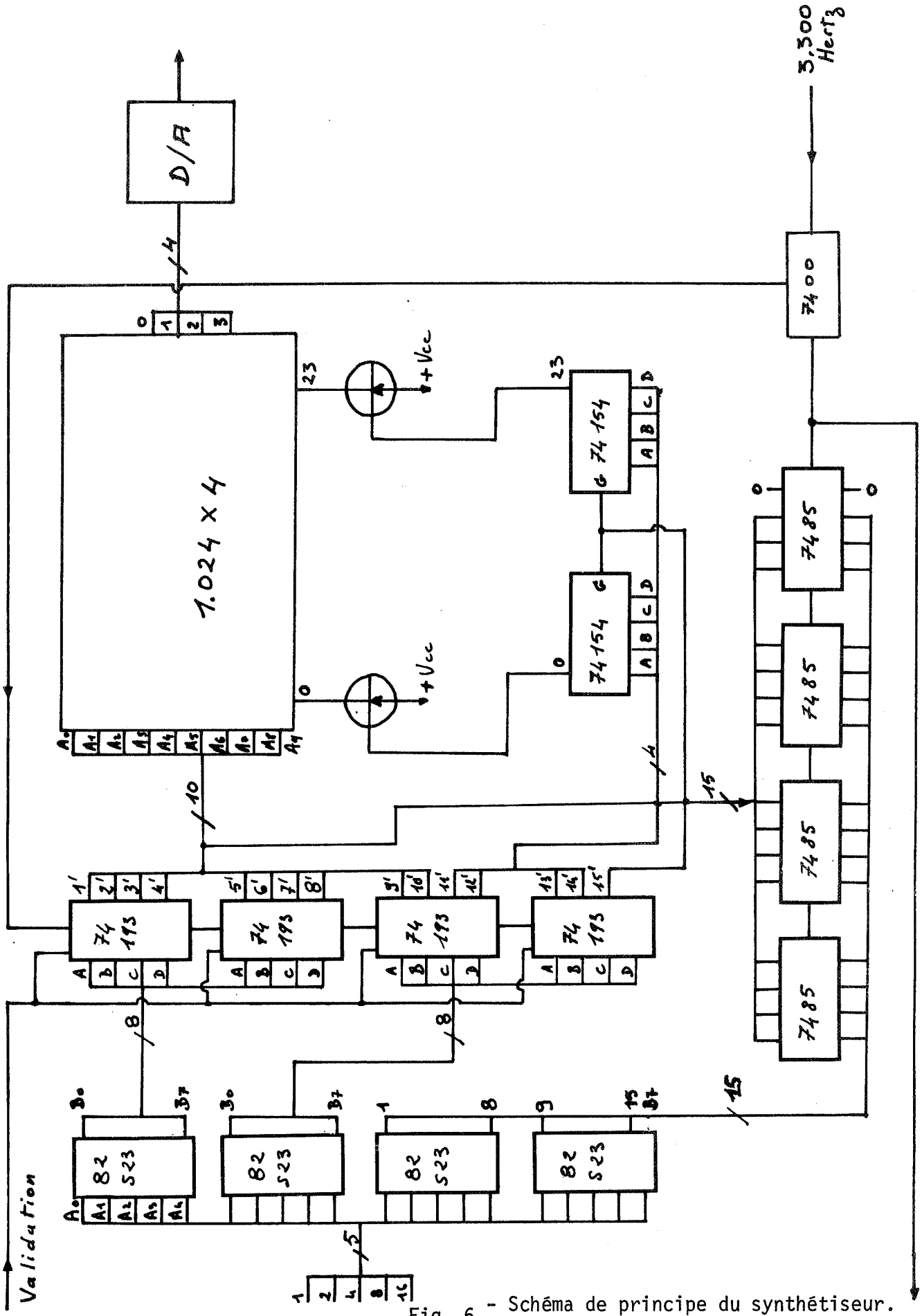


Fig. 6 - Schéma de principe du synthétiseur.
- General outline of the synthesizer.

N° du Locuteur	Esthétique	Intelligibilité	Durée Moyenne	Durée du segment 5 (cinq) en milli-secondes
1				450
2				500
3		x	x	350
4			x	375
5				450
6				400
7	x	x	x	375
8				385
9			x	375
10				410

- Durée du chiffre cinq prononcé par 10 locuteurs différents.

Tableau N° 7

- Duration of the number five, pronounced by 10 different speakers.

SEGMENT (chiffre ou nombre)	DUREE CHOISIE en millisecondes	NOMBRE DE BITS dans la Mémoire
1	231	765
2	189	624
3	330	1.023
4	310	1.023
5	370	1.217
6	280	924
7	250	831
8	314	1.023
9	310	1.023
10	239	791
11	300	975
12	314	1.023
13	350	1.151
14	566	1.871
15	355	1.167
16	310	1.023
20	250	835
30	310	1.023
40	430	1.415
50	540	1.783
60	465	1.535
100	250	831
1000	465	1.535

- Durée moyenne choisie pour la durée des segments et leur taille mémoire respective.

Tableau N° 8

- Mean duration elected for the segment duration and the memory wide.

REFERENCES

AUTESSERRE, D. et DI CRISTO, A., 1972, "Recherche sur l'intonation du Français : Traits significatifs et non significatifs", Proceedings of the VIIth International Congress of Phonetic Sciences, Mouton, The Hague, 1972, pp. 842-859.

CHAFCOULOFF, M., 1976, Vingt cinq années de recherches en synthèse de la parole, Editions du C.N.R.S., Paris, p. 283, 1976.

HARRIS, C.M., 1953, "A Study of the building blacks in speech", J.A.S.A., 25, 5, pp. 962-969.

HNATEK, E.R., 1976, A user's hondboak of semiconductor memories. J. Wiley, Londres, p. 688, 1976.

HNATEK, E.R., 1977, A user's hondboak of D/A and A/D converters. J. Wiley, Londres, p. 472, 1977.

LEE, S.C., 1976, Digital circuits and logic design Prenctice Hall, New York, p. 594, 1976

ROSS, E.J., 1977, Modern digital communications Mc Grow Hill, New-York, p. 308, 1977

WANG, W., PETERSON, G.E., 1958 "Segment Inventory for Speech Synthesis". J.A.S.A., 30, 8, pp. 743-746, 1958.

WIGGINS, R., BRANTINGHAM, L., 1978, "Three chip system synthesizes human speech" Electronics 51,18, August, pp. 109-116.

10èmes JOURNÉES D'ÉTUDE SUR LA PAROLE

GRENOBLE - 30 Mai - 1er Juin 1979

Un système de traitement rapide de signaux digitaux utilisé en synthèse de la parole.

Xavier RODET - Jean-Luc DELATRE (IRCAM)

Institut de Recherche et de Coordination Acoustique/Musique
Département Diagonal, 31, rue Saint Merri 75004 PARIS
tél: 277 12 33 poste 4820 ou 4827

RESUME

Nous décrivons le système SARA destiné à la production de signaux acoustiques complexes par création, manipulation et mixage de flots d'échantillons. C'est un système à usage général sur lequel nous implantons des programmes de synthèse de parole et de musique utilisant de nouveaux développements des techniques mises au point dans l'ancien système SARA (Synthèse de la parole par règles, utilisation de fonctions d'ondes dans le domaine temporel). Cependant il faut noter que ce système peut servir à de nombreuses autres applications.

Dans SARA, pour chaque pseudo-période, le signal acoustique est construit, échantillon par échantillon comme une somme de fonctions d'ondes. La valeur de chaque échantillon résulte donc d'une combinaison de calculs élémentaires dont le nombre et le type peuvent changer très souvent en raison de la complexité des trajectoires des paramètres acoustiques dans la parole. Pour traiter ce problème, le système SARA-JUNIOR considère chaque calcul élémentaire comme une tâche indépendante, permettant ainsi des modifications dynamiques aisées du traitement des données. Les changements dynamiques aisés dans le traitement des données permettent au système d'optimiser le temps de calcul propre à la construction des signaux, en détectant et supprimant les exécutions non-indispensables de certaines opérations comme X + 0, X * 1 etc...

Le langage de commande permet à l'utilisateur de déclencher à des instants précis, le démarrage, la modification ou la suppression de séquences de calculs. La syntaxe de ce langage autorise l'écriture de commandes complexes sous forme de règles de réécriture de type LL1. Ainsi peut être décrite, par exemple, la traduction d'une chaîne de phonèmes en une suite d'événements nécessaires à la synthèse du signal vocal correspondant.

A flexible and fast signal management system for speech synthesis.

Xavier RODET - Jean-Luc DELATRE (IRCAM)

SUMMARY

We describe the SARA system for generating, editing and merging streams of samples to produce complex audio signals. It is a general-purpose software management system in which we have embedded processes for synthesis of speech and music using new developments in the techniques pioneered in the previous SARA software (Speech Synthesis-by-Rules using time-domain functions). Although it should be noted that the process management system is quite general in its applicability.

In SARA, acoustic parameters (i.e. pitch period duration, formant frequencies and formant amplitudes) are calculated for each pitch period by positioning in time segments of pitch contour and formant trajectories to form the complete trajectory. At certain points along these trajectories, changes often are necessary in the treatment of data. We shall call these changes "events".

According to the acoustic parameters we construct each pitch period waveform sample by sample as a summation of calculated time-domain functions. Thus the value of each sample results from a combination of elementary calculations, the kind and the number of which can change quite often due to the complex trajectories of acoustic parameters in speech. To deal with these problems the SARA-Junior system treats each elementary calculation as an independent task, thus allowing easy dynamic changes of the treatment of the data. The fact that each such task is executed only once for each sample (if needed at all) allows the tasks scheduler to be extremely simple (round-robin), consequently the scheduling overhead is minimized. Moreover the easy dynamic changes in the treatment of the data enable the system to optimize the runtime by dynamically detecting and removing superfluous operations like repeated executions of X + 0, X * 1 etc... The tests performed by the events controller are also kept to a minimum through the use of a distributed control structure not requiring any search of an events list in most cases.

The user interface to the system is through a command language with which one can trigger at specific instants the start, alteration or cancellation (events) of calculation sequences. The language syntax is such that complex commands can be stated as LL(1) rewriting rules. Allowing, for example, the translation of a string of phonemes into a sequence of events driving the synthesis of the corresponding speech.

10èmes JOURNÉES D'ÉTUDE SUR LA PAROLE
GRENOBLE - 30 MAI - 1ER JUIN 1979

Un système de traitement rapide de signaux digitaux utilisé en synthèse de la parole.

Xavier RODET, Jean-Luc DELATRE (IRCAM)
Institut de Recherche et de Coordination Acoustique/Musique
Département Diagonal, 31, rue Saint Merri, 75004
tél: 277 12 33 poste 4820 ou 4827

I. FINALITES DU SYSTEME

Nous avons défini et construit un système et un langage à usage général orientés vers le calcul de signaux digitaux complexes. Ce système doit faciliter l'installation et l'usage de programmes de synthèse de parole et de musique. Cependant, il y a de nombreuses autres applications lorsque le problème central est le calcul de fonctions échantillonnées. Ces applications se ramènent à la transcription d'un flot d'information en entrée tel que texte, phonèmes, partition musicale, paramètres, chaines de codes de façon générale, en une ou plusieurs séquences d'échantillons qui sont la réalisation du "sens" de la chaine d'entrée.

En premier lieu, nous avons choisi des moyens de commande qui facilitent l'analyse des informations d'entrée, considérées comme appartenant à un langage défini par une grammaire LL(1). Par ailleurs, ce système doit faciliter la création, la manipulation et le mixage de flots d'échantillons quelconques: ceux-ci pouvant représenter n'importe quel niveau de structure du signal. Ainsi nous écrivons dans ce langage une nouvelle version du programme SARA de synthèse de la parole. Dans SARA les paramètres acoustiques (durée de pseudo-période, fréquence et amplitudes des formants) sont recalculés pour chaque pseudo-période. Ce calcul est effectué par positionnement en temps de portions de trajectoires mélodiques et formantiques, de façon à construire la trajectoire définitive. De même pour chaque pseudo-période, le signal acoustique est construit, échantillon par échantillon comme une somme de fonctions d'onde.

Nous définissons de la sorte un langage grâce auquel il est possible de déclencher et de contrôler la production de flots d'échantillons en se référant au contenu d'un ou plusieurs flots d'information en entrée aussi bien qu'à l'écoulement d'un temps courant ou à des critères propres aux étapes intermédiaires des calculs. De plus nous tenons à ce que ce langage présente sensiblement les mêmes possibilités qu'un langage classique de programmation. Enfin ce langage peut servir au musicien comme outil de composition, car il représente les processus les plus généraux (évolution dynamique d'une phrase, ou manipulation symbolique des structures) de la même manière que les processus au niveau de l'échantillon. Cette "homogénéité" du langage est tout à fait nouvelle dans un langage de commande pour la synthèse du signal sonore.

II PREMIERE APPROCHE DU PROBLEME

Comme il s'agit en fin de compte de calculer des signaux digitaux, la fréquence d'échantillonnage détermine une unité temporelle, un "pas" élémentaire dans notre système. D'autre part, les flots d'échantillons en sortie sont construits par une suite de transformations dont les opérandes et les

résultats sont des flots d'échantillons dont les valeurs se trouveront à chaque pas dans des variables (correspondant chacune à un flot). Il suffira donc, de spécifier à chaque pas les expressions qui définissent les valeurs de ces variables. Une telle expression pourra faire référence aux valeurs d'autres variables, soit telles qu'elles étaient au pas précédent soit telles qu'elles doivent être au présent pas. Ce dernier cas implique un ordre d'élaboration des valeurs qui n'introduise pas de cercle vicieux: certaines variables devant contenir en quelque sorte des résultats intermédiaires qui concourent à l'élaboration d'autres valeurs, elles ne peuvent elles-mêmes être définies par rapport au résultat auquel elles participent.

En général, l'expression qui définit la valeur instantanée d'un flot d'échantillons (contenue dans une variable) ne change pas pendant des durées notables. Alors il n'est nécessaire que de spécifier à quels instants une telle expression doit changer et non plus de spécifier une expression à chaque pas. Nous nommons "événements" ces définitions ou redéfinitions de valeurs.

Sous-programmes itératifs et variables échantillonnées.

Le genre d'expression mentionné ci-dessus apparaît tout naturellement comme un sous-programme itératif qui calcule à chaque pas une nouvelle valeur de la variable. Nous parlons alors de variable échantillonnée et elle représente un flot d'échantillons. D'autre part, la définition du contenu d'une telle variable est susceptible de changer et cette redéfinition (ou définition initiale) consiste à lier un (autre) sous-programme à la variable échantillonnée. Ce qui s'exprimera dans notre langage par une instruction d'affectation particulière dite attribution:

<variable échantillonnée>:=<expression>

qui signifie non pas seulement que la variable reçoit la valeur que dénote l'expression à cet instant, mais également qu'elle recevra désormais à chaque "pas" la valeur correspondant à l'expression. Il faut noter que l'attribution d'une instance de sous-programme à une variable n'exclut pas l'attribution de cette même instance à une ou plusieurs autres variables.

Structure de base d'un programme.

Pour décrire un programme de manipulation de flots d'échantillons il suffit de noter sur l'axe du temps, l'occurence des attributions <variable échantillonnée>:=<expression>. L'exécution d'un programme consiste à calculer, pour chaque pas, les valeurs prises par les variables échantillonnées: la structure d'un programme est donc une simple boucle (implicite) formée d'appels aux sous-programmes de calcul (Fig. 1). Chaque sous-programme peut être conçu comme un processus indépendant.

Exécution des attributions.

L'instant où s'effectue l'attribution d'un sous-programme à une variable échantillonnée ("événement") est fonction des valeurs de certaines variables. La procédure qui décide de la prise d'effet de chaque (re)définition de variable échantillonnée peut donc être vue elle aussi comme un sous-programme répétitif qui, au lieu de délivrer une valeur à chaque pas, teste si l'instant d'une (re)définition particulière est venu et, si oui, attribue à la variable

échantillonnée le sous-programme qui va désormais calculer sa valeur.

III OPTIMISATION DE L'USAGE DES RESSOURCES ET AUTRES NECESSITES

A cause de sa souplesse le système décrit ci-dessus risque de consommer une grande part des ressources de calcul d'un processeur. Nous avons donc introduit d'autres améliorations pour minimiser l'usage des ressources.

En premier lieu, il n'est pas nécessaire de recommencer un calcul dont le résultat est déjà connu: par exemple, lorsque les opérandes qui permettent d'obtenir sa valeur ne changent pas. Toute instance de sous-programme dont le résultat est connu jusqu'à un certain instant est donc extrait de la boucle d'appel des sous-programmes et mis dans une file d'attente. Une routine de service se charge au moment opportun de réintroduire dans la boucle les sous-programmes ainsi suspendus (Fig. 1).

Cette gestion des stabilisations des sous-programmes est effectuée grâce à des liens qui identifient dans chaque instance de sous-programme élémentaire quels sont les opérandes de celui-ci et quels sont les usages du résultat.

Pour que cette technique soit profitable, il est exclu d'avoir à tester les conditions de stabilité de chaque calcul. C'est pourquoi, on procédera à l'envers, c'est-à-dire que seuls certains sous-programmes auront l'initiative de ces tests de stabilité et propageront cette information vers ceux qui font usage de leur résultat, aussi bien que vers leurs opérandes (lesquels ne sont plus nécessaires dès lors qu'ils n'ont plus d'usage). Par exemple, certains segments d'enveloppes temporelles, ou des bornes de paramètres tabulés peuvent constituer des cas de stabilité. De plus, lors de la propagation d'une information de stabilité, de nouvelles stabilités peuvent être détectées (par exemple si tous les opérandes d'une fonction deviennent stables, ou si certains atteignent des valeurs particulières comme 0 dans une multiplication). Et ces nouvelles stabilités, à leur tour sont propagées vers les opérandes et les usages concernés.

Purge des ressources en fin d'utilisation.

Les liens entre opérandes et usages servent également à déterminer l'obsolescence des instances de sous-programmes et à libérer les ressources associées à celles qui, du fait des redéfinitions de variables échantillonnées n'ont plus aucun usage. Cette libération est faite par une technique de "garbage collection" incrémental, c'est-à-dire exécuté partiellement à chaque création d'une instance de sous-programme.

IV DESCRIPTION GENERALE DU SYSTEME

Programme

Un programme se compose de:
- une suite de déclarations de variables
- une suite d'instructions. Elles seront exécutées l'une après l'autre dès le début de l'exécution du programme jusqu'à la dernière ou jusqu'à ce que l'une de ces instructions suspende sa propre exécution et celle des instructions qui la suivent. Cette suspension intervient si une condition qui figure dans une instruction de test ('IF') n'est pas remplie au départ

et peut l'être par la suite.

Variables

Il existe deux catégories bien distinctes de variables:
- les variables échantillonnées (identifiées par le mot clé 'SAMPL')
- les variables non échantillonnées qui, devant servir à l'élaboration de résultats intermédiaires à l'intérieur d'un sous-programme, et n'étant pas censées transmettre une valeur d'un pas à l'autre, ne sont pas soumises à la gestion usages/opérandes. Pour ces dernières, l'instruction d'affectation retrouve son sens habituel d'altération immédiate et unique de la valeur.

Fonctions et procédures.

Les fonctions et procédures que l'on peut déclarer permettent de définir et de structurer les calculs nécessaires à l'obtention des résultats désirés.

Celles qui sont déclarées sans le préfixe 'SAMPLE' sont analogues aux fonctions et procédures classiques de la programmation: elle indiquent une action immédiate et unique à leur point et instant d'invocation.

Au contraire, les procédures et fonctions échantillonnées quand elles sont invoquées sont placées dans la boucle de calcul, soit pour fournir des valeurs à des variables échantillonnées (fonctions) soit pour modifier des relations entre variables et fonctions à des instants précis soit pour d'autres effets tel l'écriture sur un support externe (procédures). Lors de l'appel une instance de la fonction ou procédure échantillonée est créee (allocation de mémoire pour les variables et la pile, initialisations etc...), introduite dans la boucle de calcul, et le contrôle retourne immédiatement à l'appellant.

Instructions de contrôle du déroulement.

D'une part, nous avons besoin de spécifier à quels instants doivent se produire les "évènements", d'autre part nous voulons "transcire" un flot d'information en entrée qui se présente sous forme d'une séquence de signes.

Nous pouvons décrire une telle séquence de signes et les transformations que l'on désire y apporter par une grammaire LL(1) déterminant les constructions possibles au moyen de ces signes et définissant leurs transformés grâce à des annotations sémantiques mêlées aux règles. Il faut donc pouvoir reconnaître chaque construction sans ambiguité et en déduire des créations et transformations de flots d'échantillons. Pour permettre la reconnaissance aisée des constructions, nous vérifierons que les grammaires que nous utilisons sont LL(1) et au besoin nous les modifierons en subdivisant certaines règles et/ou en introduisant des règles auxiliaires afin de satisfaire à cette restriction. Cependant, d'autres méthodes d'analyse syntaxique sont utilisables en cas de nécessité. Le contrôle du déroulement se fait au moyen de la construction:

```
DO
IF <condition 1><action 1/1><action 1/2> ...
IF <condition 2><action 2/1><action 2/2> ...
...
IF <condition n><action n/1><action n/2> ...
END
```

ou les différentes conditions sont supposées mutuellement exclusives.

Cette construction est, du point de vue sémantique strictement équivalente à une règle de grammaire LL(1) car si aucune condition n'est satisfaite la construction elle-même a la signification d'une condition fausse. Chaque construction non échantillonnée du langage a une valeur logique en sus de sa propre action ou valeur numérique et peut dont être utilisée en tant que condition. Il suffit pour cela qu'elle apparaisse à droite de 'IF' ou '&'.

Ainsi n'importe quelle grammaire LL(1) peut être réécrite comme un ensemble de procédures.

Syntaxe du langage.

Nous utilisons une variante de la notation BNF:
- Les mot-clés apparaissent entre apostrophes
- L'occurence optionnelle d'une séquence de symboles est indiquée par des crochets [...]
- L'occurence répétitive (zéro, une, ou plusieurs occurrences) est indiquée par une paire d'accolades {... }

Selon ces conventions un programme est donc décrit par:

```
program = 'PROGRAM' {declaration ';'}
          .    {instruction separator} .
separator = ',' | ';' .
declaration = vardecl | procedure | function.
vardecl = [ 'SAMPL'] [ 'LONG'] vartype ident
          {',' ident} .
vartype = 'INT' | 'FRAC'.
procedure = ['SAMPL'] 'PROC' ident [parms] ';'
            procbody .
parms =  '(' vardecl {';' vardecl } ')'
procbody = {declaration;} blockbody 'END'
           | 'EXTERNAL' .
function =  ['SAMPL'] ['LONG'] vartype
            'PROC' ident[parms]';' funcbody
funcbody = {declaration ';'} valuespec 'END'
           | 'EXTERNAL' .
blockbody = altrules | {instruction } .
altrules = clause {clause} [endclause] .
clause = 'IF' conjunction {instruction } .
conjunction = cond {'&' cond} .
endclause = 'DONE' | 'ELSE' 'DONE'
            | 'ELSE' {instruction} .
cond = expr comp expr | parg | procall.
```

```
parg = 'DO' altrules 'END' .
procall = procname [args] .
args = '(' expr {',' expr} ')' .
expr = term {('+'|'-') term} .
term = fact {('*'|'/') fact} .
fact = label ':' fact | '-' fact | funccall
       const | varname | '(' valuespec ')' .
funccall = funcname [args] .
valuespec = valtrules | valtext .
valtext = {instruction} expr {instruction} .
valtrules = 'IF' vclause {'IF' vclause}
            ['ELSE' valtext] .
vclause = conjunction valtext
        | valcond {instruction} .
valcond = cond '&' valcond | expr {'&' cond} .
instruction = label ':' instruction | procall
              parg | varname ':=' expr
             | 'RESUME' | 'EXIT'.
```

Les occurences de ident, varname, procname, funcname, label représentent des noms symboliques du langage. Le langage a une structure de blocs et suit les conventions d'ALGOL en ce qui concerne la portée des déclarations.

V DEROULEMENT D'UN PROGRAMME

Si une construction qui ne suit pas un 'IF' ou '&' a la valeur logique "faux", cette construction contient quelque condition non satisfaite et aucune autre alternative (à la satisfaction de cette condition) n'est possible. Alors la procédure contenant cette construction doit être interrompue. Cependant, si une variable échantillonnée figure parmi les variables qui concourent à l'évaluation "faux", la condition peut se trouver satisfaite au "pas" suivant. La procédure ayant obtenu cette condition d'arrêt est donc suspendue jusqu'au "pas" suivant ou son calcul est réexécuté à partir du point qui précède immédiatement la construction qui provoque l'arrêt.

On voit donc qu'une procédure une fois lancée, au cours d'un cycle ou "pas" de calcul, se déroule à l'intérieur de ce cycle jusqu'à ce qu'elle rencontre une condition qui détermine pour elle la fin de ce cycle. Cette fin de cycle peut être spécifiée explicitement au moyen de l'instruction 'RESUME' qui passe le contrôle au prochain sous-programme dans la boucle d'appel. Chaque tour de la boucle de calcul correspond à un "pas", c'est-à-dire au calcul d'un échantillon pour chaque flot dont la valeur n'est pas prévisible ou est différente de la valeur au pas précédent.

Le système de gestion interne examine les stabilités des valeurs logiques et extrait de la boucle de calcul les sous-programmes qui sont suspendues (qui sont immobilisés par une condition non-satisfaite) pour une durée prévisible.

La suspension peut être de durée infinie, par exemple lorsque la condition non-satisfaite ne se réfère à aucune variable échantillonnée. L'instruction 'EXIT' permet également de provoquer la suspension définitive d'une procédure. Si une fonction est suspendue définitivement sa dernière valeur reste disponible tant qu'elle conserve quelque usage.

VI CONCLUSION

Le langage de commande que nous avons décrit permet à l'utilisateur de contrôler des "événements" précis et leurs interactions, de déclencher le démarrage, la modification ou la suppression de séquences de calculs. Chaque calcul élementaire est considéré comme une tâche indépendante qui délivre un échantillon à chaque pas (si nécessaire). Cette structure autorise un enchaînement des tâches très simple (Round Robin) ce qui réduit d'autant le temps de calcul nécessaire à cette gestion. Le temps de calcul propre à la construction des flots d'échantillons est réduit grâce à la suspension des calculs redondants tandis que les tests effectués par le contrôleur d'événements sont minimisés par l'emploi d'une structure de contrôle répartie. Enfin la syntaxe de ce langage autorise l'écriture de commandes complexes sous forme de règles de réécriture de type LL(1).

Ce système est en cours d'installation sur un ordinateur PDP-11 pour la mise en oeuvre du programme SARA de synthèse de la parole et d'autres projets de synthèse musicale.

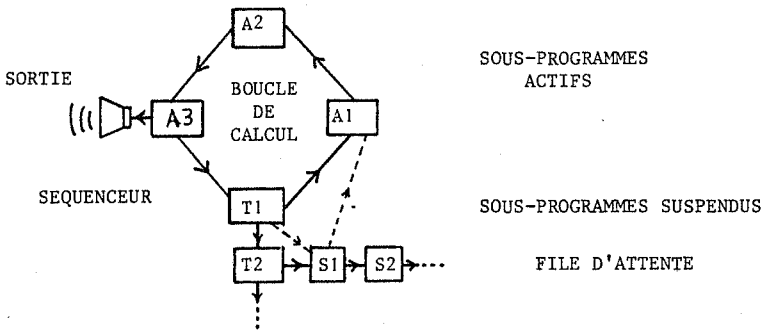


FIG. 1 . BOUCLE DE CALCUL DES FLOTS D'ECHANTILLONS

10èmes JOURNÉES D'ÉTUDE SUR LA PAROLE

GRENOBLE - 30 MAI - 1ER JUIN 1979

Construction du signal vocal dans le domaine temporel

Xavier Rodet - Jean-Luc Delatre (IRCAM), Mohammed Razzam (C.E.N.S.)

- Institut de Recherche et de Coordination Acoustique/Musique
  Département Diagonal, 31, rue St. Merri 75004 PARIS tél 2771233
- Centre d'Etudes Nucléaires de Saclay S.E.S./S.I.R.
  BP2 91190 Gif-sur-Yvette tél: 9418000 poste 5626

RESUME

Nous étudions et comparons deux méthodes pour construire les parties voisées de signaux de parole, ou de signaux musicaux, directement dans le domaine temporel au moyen de fonctions d'onde formantiques. Dans la première méthode de synthèse, les fonctions d'onde formantiques sont calculées en appliquant à des ondes sinusoïdales une enveloppe spécifique (du genre: fenêtres temporelles d'analyse). Ceci permet d'éliminer tout bruit parasite et de produire, suivant l'enveloppe, des spectres de formes variées. Cette technique a permis de synthétiser une parole de bonne qualité. Dans la seconde méthode les fonctions d'onde formantiques sont extraites d'une période de voix naturelle. Pour ce faire, un filtre prédictif est calculé sur une période fondamentale du signal de la voix naturelle. La forme parallèle de ce filtre est divisée en q groupes de sections du 2ème ordre suivant la répartition des fréquences des pôles. Le filtrage de l'erreur de prédiction à travers chacun des q groupes fournit q fonctions d'onde formantiques de base.

Pour reconstruire une période du signal, chacune de ses fonctions de base subit indépendemment une affinité suivant l'axe du temps et l'axe des amplitudes. Ceci afin d'engendrer les fréquences et amplitudes de formants voulues. Le signal final est la somme point par point des q fonctions modifiées.

Dans ces deux méthodes, le calcul d'une forme d'onde D_i est commencé au début de chaque pseudo-période i. Il est prolongé jusqu'à atténuation suffisante pendant la pseudo-période suivante (i+1) durant laquelle la forme D_i est ajouté en chaque point à la nouvelle forme D_{i+1}.

Construction of the sound signal in time-domain.

Xavier RODET - Jean-Luc DELATRE (IRCAM), Mohammed Razzam (C.E.N.S.)

SUMMARY

In this paper we study and compare two methods for constructing voiced parts of speech signals directly in time-domain by use of formant time-domain functions. Beforehand, speech signals are analysed to extract the acoustic parameters for each pitch period. These parameters are the duration of the period, the formant frequencies F_i and the formant amplitudes ($1 \leq q \leq 6$ formants). The signal is then computed period by period using a summation of q time-domain functions W_i, each corresponding to one or more of the q formants detected in the original signal period.

In the first method, each formant time-domain function W_i is constructed by applying a specific time envelope to a damped sinusoïd of frequency F_i. The shape of this envelope is altered to best fit the natural speech spectrum and minimize any resulting noise (Fig 1 to 5).

In the second method, basic formant wave-functions are obtained from an L.P.C. model of a natural speech pitch period. For this purpose, the paralel form of the corresponding filter is divided into q groups of second order sections according to the placements of the frequencies of the poles. Filtering of the L.P.C. error signal by each group gives p basic formant time-domain functions. To reconstruct a pitch period of the signal, each of these functions is independently modified by time and amplitude compression (or extension to yield the desired values of the formant frequencies and amplitudes. At the end, the signal is a summation, sample by sample, of those q modified wave-functions.

In both methods, the calculation of a wave-form is initialized at the beginning of each pitch period, and is continued until sufficiently attenuated during the following pitch period where it is added sample by sample to the new initialized wave-form.

Good quality speech can be produced by these both methods in a relatively efficient manner.

10èmes JOURNÉES D'ÉTUDE SUR LA PAROLE
GRENOBLE - 30 MAI - 1er JUIN 1979

Construction du Signal Sonore dans le Domaine-Temporel.

Xavier RODET, Jean-Luc DELATRE (IRCAM), Mohammed Razzam (C.E.N.S.)
- Institut de Recherche et de Coordination Acoustique/musique, Département Diagonal, 31, rue St. Merri, 75004 PARIS, tél 277.12.33 poste 4820 ou 4827
- Centre d'Etudes Nucléaires de Saclay S.E.S./S.I.R. BP2 91190 Gif-sur-Yvette, tél: 941.80.00 poste 5626

I INTRODUCTION

Nous présentons deux méthodes de construction des parties voisées du signal sonore de la parole qui étendent et améliorent la méthode utilisée dans le programme SARA [1] (fonctions d'onde formantiques, dans le domaine temporel). Nous appellons "caractéristiques acoustiques" du signal, les fréquences F_i et amplitudes A_i d'un certain nombre q de formants ($0<q\leqslant 6$) ainsi que la fréquence fondamentale F_0. Ces paramètres sont des fonctions discrètes du temps qui prennent une nouvelle valeur pour chaque nouvelle pseudo-période fondamentale.

II EXTRACTION DES CARACTERISTIQUES ACOUSTIQUES

Pour tester nos méthodes de synthèse, nous avons choisi en premier lieu d'extraire ces paramètres de la parole naturelle, avec une procédure très simple:

- Enregistrement et digitalisation d'une phrase prononcée par un locuteur.
- Détection du pitch au moyen d'un programme de corrélation [2]. Ce programme fournit une table des durées T_j des pseudo-périodes successives du signal.
- Calcul du spectre de chaque pseudo-période j (F.F.T.).
- Extraction de l'amplitude A_{ij} et de la fréquence F_{ij} des maximums i du spectre par un programme semi-automatique. Par extension, nous appellerons "formants" les maximums retenus.

Dans cette première étude nous n'avons pas cherché à extraire d'information concernant la largeur de bande.

III MODELISATION DE SPECTRES AU MOYEN DE FONCTIONS D'ONDE FORMANTIQUES

1. Construction du signal correspondant à un formant.

Nous pouvons construire directement dans le domaine temporel et de façon simple une fonction d'onde dont le spectre est analogue à celui d'un formant. En première approximation, considérons une fonction d'onde telle que (Fig. 1):

$$Y_{(t)} = A_i\, e^{-\alpha_i t} \sin(2\pi F_i t)\ ,\quad t \geqslant 0$$

Cette fonction d'onde est le produit d'une enveloppe $E(t) = e^{-\alpha_i t}$ et d'une onde sinusoïdale, avec un coefficient d'amplitude A_i. Le spectre de cette onde (Fig. 2) est alors le produit de convolution $E_{(f)} \circledast C_{(f)}$ des spectres respectifs de l'enveloppe ($E_{(f)}$) et de l'onde sinusoïdale ($C_{(f)}$).

Cette formule permet de contrôler facilement la fréquence centrale du formant F_i , sa largeur de bande $\Delta F_i = \frac{\alpha_i}{\pi}$ et son amplitude A_i. Mais le spectre de cette envelope n'est pas limité. Nous avons donc utilisé l'enveloppe $e^{-\alpha_i} \frac{(t-c)^2}{t}$ dont le spectre est bien limité. (Fig. 3 et 4)

(Fig. 5). Une autre solution consiste à filtrer la discontinuité intitial de l'enveloppe $e^{-\alpha_i t}$, par exemple en la multipliant par une fonction sinusoïdale croissant de 0 à 1. En fonction de la forme des spectres à reproduire (par exemple pour des sons instrumentaux), on peut envisager d'autres types d'enveloppes d'amplitude dont les transformées présentent un pic principal plus ou moins étroit, et des lobes latéraux d'amplitudes plus ou moins élevés. On peut trouver de bons exemples d'envelopppes parmi les fenêtres temporelles utilisées en analyse spectrale (3). Les critères de choix peuvent être, entre autres:

- concentration de l'énergie dans une bande étroite en fréquence et sur une fenêtre temporelle de durée aussi réduite que possible (pour diminuer les calculs et respecter les transitions).
- approximation d'une portion de spectre de forme quelconque (Exemples Fig 6).

Finalement une fonction d'onde est définie comme la somme de q fonctions d'onde élémentaires, appellées par extension fonctions d'onde formantiques, chacune par exemple approchant un formant ou une partie du spectre de façon à respecter au mieux un spectre donné.

2. Construction d'un signal harmonique.

Nous construisons un signal de fréquence fondamentale Fo et ayant les caractéristiques spectrales voulues en répétant une des fonctions d'onde composites définies ci-dessus à des intervals de temps correspondant à $\frac{1}{F_o}$ (Fig 7). Le signal résultant est la somme de toutes les fonctions d'onde présentes à chaque instant.

Pour diminuer les calculs, il est pratique, si l'enveloppe temporelle n'atteint pas la valeur zéro assez rapidement, de la limiter arbitrairement en la multipliant dans sa partie finale déjà amortie, par une fonction d'aténuation (par exemple de la forme $\frac{1}{2}(1 + \cos(\frac{t}{T_o}))$.)

FONCTIONS D'ONDE FORMANTIQUES EXTRAITES D'UNE PERIODE DE VOIX NATURELLE

Cette deuxième méthode permet de décomposer le signal d'une période fondamentale d'un son quelconque en des fonctions d'ondes partielles (dites "formantiques") correspondant à des régions distinctes de l'axe des fréquences. Et la somme de ces fonctions est identique du signal de la période fondamentale que l'on veut reproduire.

1. Calcul d'un filtre prédictif.

Dans la méthode de prédiction linéaire, le spectre d'un signal s(k) (de transformée en z: S(z) est modélisé sous la forme d'un filtre:

$$H(z) = G.\ \frac{1}{A(z)} = G.\ \frac{1}{1 + \sum_{k=1}^{M} a_k z^{-k}}$$

dont la fonction de transfert est la plus proche possible du spectre du signal au sens des moindres carrés.

Pour le calcul d'un tel filtre, nous avons utilisé le programme de J.A. MOORER [2] (méthode de BURG [4]).

Nous calculons alors le signal d'erreur de prédiction

$$G.\ E(z) = S(z)\ .\ A(z)$$

où G est un facteur de gain.

Remarquons que : $S(z) = G\ .\ E(z)\ .\ \frac{1}{A(z)}$

2. Calcul du filtre parallèle équivalent.

Le filtre prédictif $\frac{1}{A(z)} = \frac{1}{\sum_{k=0}^{M} a_k z^{-k}}$, $a_0 = 1$

peut s'écrire (décomposition en fractions rationnelles):

$$\frac{1}{A(z)} = \frac{1}{\prod_{k=1}^{M} (1 - R_k z^{-1})}$$

où les R_k sont les racines du polynome A(z). De la même façon, si nous choisissons M pair (M=2m)

$$\frac{1}{A(z)} = \frac{1}{\prod_{k=1}^{m} (1 + u_k z^{-1} + v_k z^{-2})}$$

ou encore

$$\frac{1}{A(z)} = \sum_{k=1}^{m} \frac{c_k + d_k\ z^{-1}}{1 + u_k z^{-1} + v_k z^{-2}}$$

(on suppose que les racines sont distinctes).

(c_k, d_k, u_k, v_k réels)

Ce filtre peut alors être considéré comme un ensemble de m sections du 2ème ordre disposées en parallèle. Nous appelons F_r la fréquence des pôles d'une section k, le cas des pôles réels ne posant pas de difficulté.

3. Découpage des filtres en q groupes distincts.

Nous choisissons sur l'axe des fréquences q(q 1) régions distinctes (pas nécessairement d'un seul tenant) f_i , $f_i + 1$ qui couvrent respectivement les q zones du spectre du signal origine que nous désirons séparer. Nous découpons alors la forme parallèle du filtre prédictif en q groupes $\gamma_i(z)$, chaque groupe étant constitué des r sections du 2ème ordre telles que $f_i < F(k) \leqslant f_{i+1}$

$$\gamma_i(z) = \sum_{k = k_i}^{k_i + r-1} \frac{c_k + d_k z - 1}{1 + u_k z - 1 + v_k z - 2}$$

$$\text{et } \frac{1}{A(z)} = \sum_{i=1}^{q} \gamma_i(z)$$

4. Calcul des fonctions d'onde formantiques.

Chacune des q fonctions d'onde formantiques $O_i(z)$ est le signal obtenu en sortie d'un des q groupes $\gamma_k(z)$ lorsque le signal d'entrée est le signal d'erreur G.E(z) précédemment calculé

$$O_i(z) = G.E(z) \cdot \gamma_i(z)$$

6. Discussion

Pour une meilleur précision du découpage en fréquence, il n'est pas coûteux d'utiliser un filtre d'ordre très élevé car les calculs sont effectués sur une seule période fondamentale et non pas en continu.

Ces fonctions d'onde formantiques peuvent être utilisées au même titre que celles définies au § III. En particulier, elles permettent de faire varier l'amplitude et, dans une certaine mesure, la fréquence de chaque composante du signal. Ceci grâce au fait que les flancs du spectre de chaque composante sont inclinés (Fig. 8) et non droits (comme ce serait le cas si les fonctions provenaient d'un découpage du spectre suivi d'une transformée de Fourier inverse).

Elles présentent également l'intérêt d'être amorties sur leur fin ce qui permet de construire un signal harmonique (de fréquence fondamentale F_0 suivant la méthode du § III 2) avec de très faibles variations de la forme du spectre résultant.

Enfin, on peut se contenter d'une seule fonction s'il n'est pas nécessaire de modifier séparemment les composantes. Dans ce cas, un spectre très étendu peut être reproduit moyennant un coût de calculs réduit au minimum.

V CONCLUSION

Nous avons présenté deux méthodes de calcul de fonctions d'onde formantiques qui permettent une modélisation aisée de divers types de spectres, en particulier de ceux de la parole. De plus ces fonctions, grâce à leur forme amortie se prêtent bien à la synthèse de signaux harmoniques. Et il est possible de faire varier séparemment et de façon très précise l'amplitude et la position sur l'axe des fréquences, de chaque composante; par exemple pour reproduire les variations des caractéristiques acoustiques dans les transitions phonémiques ou celles des timbres musicaux.

Enfin, il faut remarquer que ces méthodes de synthèse sont bien adaptées aux ordinateurs et microprocesseurs classiques, peu coûteuses en puissance de traitement et ne nécessitent q'une précision de calcul de l'ordre de celle souhaitée pour le signal à construire.

BIBLIOGRAPHIE

(1) RODET, Xavier, "Anayse du signal vocal dans sa représentation amplitude-temps.
Synthèse de la parole par règles", THESE Université Paris 6, 1977.

(2) MOORER, James Anderson, "The use of linear prediction of speech in computer music applications",
Pre-print n. 1320, 59th Convention of the Audio Engineering Society, Hamburg, Feb-Mar 1978.

(3) HARRIS, Frederich, "On the Use of Windows for Harmonic Analysis with Discrete Fourier Transform".

(4) MAKHOUL J., "Stable and Efficient Lattice Methods for Linear Prediction" IEEE Trans. or Acoust. Speech and Sig. Frec. Vol. ASSP-25, N° 5, October 1977.

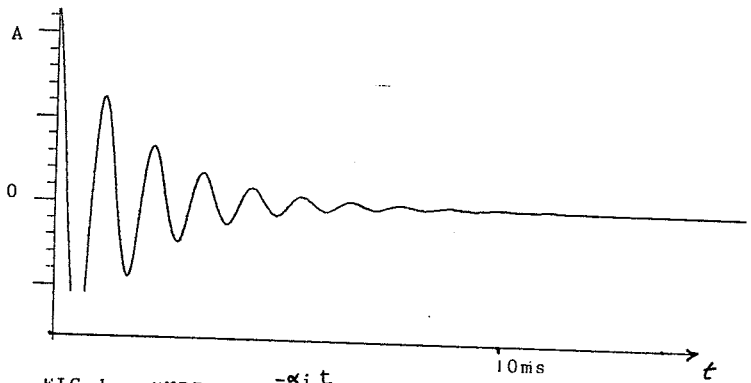


FIG. 1 _ ONDE $A_i e^{-\alpha_i t} . \sin(2\pi F_i t)$

$(c=3.5ms, \frac{\alpha_i}{\pi}=50Hz, F_i=.5KHz)$

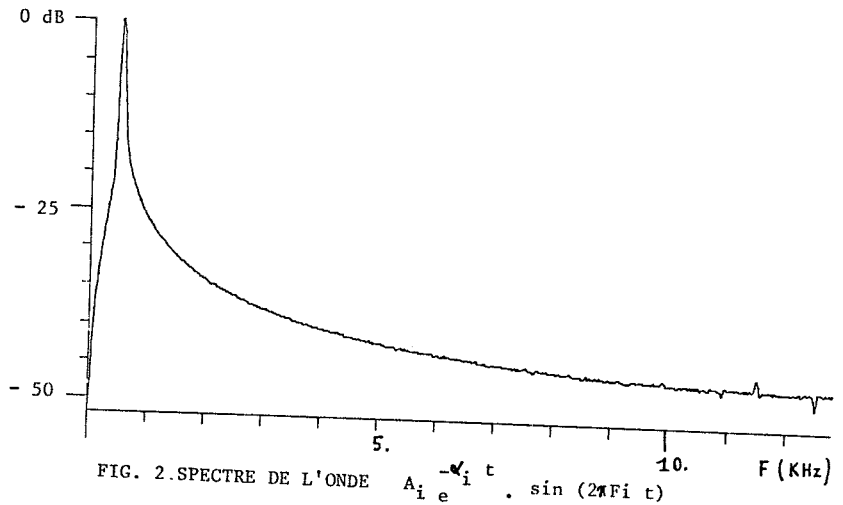


FIG. 2. SPECTRE DE L'ONDE $A_i e^{-\alpha_i t} . \sin(2\pi F_i t)$

$(\frac{\alpha i}{\pi} = 50$ Hz, Fi = .5KHz)

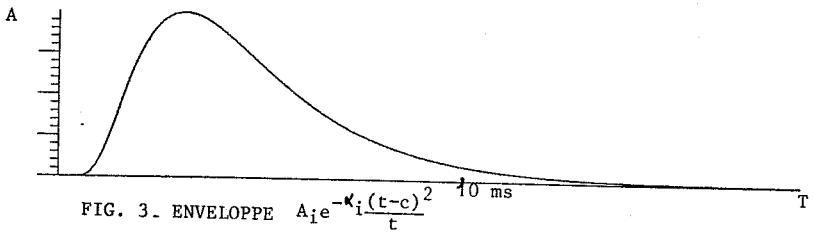


FIG. 3_ ENVELOPPE $A_i e^{-\alpha_i \frac{(t-c)^2}{t}}$

$(\frac{\alpha i}{\pi} = 50$ Hz, C = 3.5 ms)

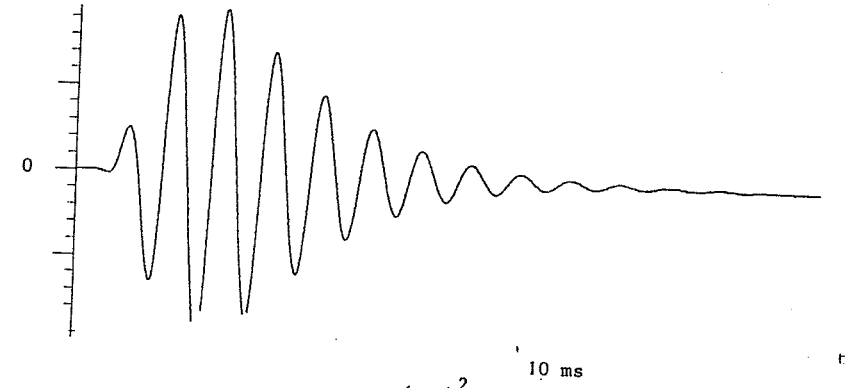


FIG. 4 _ ONDE $A_i e^{-\alpha_i \frac{(t-c)^2}{t}} . \sin(2\pi F_i t)$

$(c = 3.5$ ms, $\frac{\alpha i}{\pi} = 50$ Hz, $F_i = 500$ Hz)

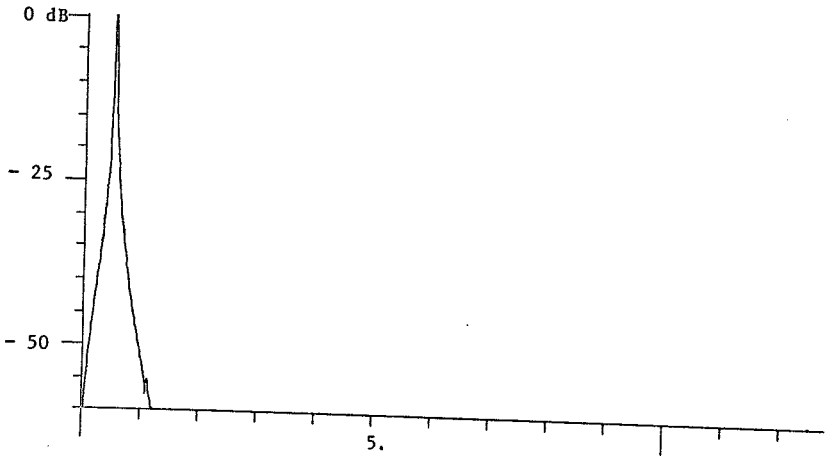


Fig. 5. SPECTRE DE L'ONDE $A_i\, e^{-\alpha_i \frac{(t-c)^2}{t}} \cdot \sin(2\pi F i t)$

(c = 3.5 ms, $\frac{\alpha i}{\pi}$ = 50 Hz, Fi = .5 KHz)

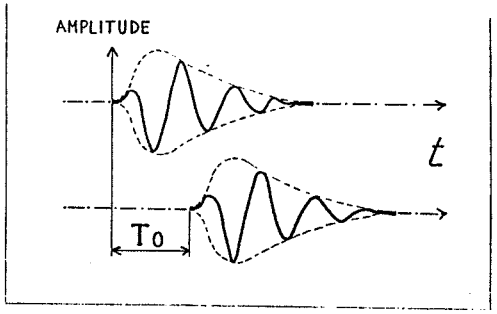
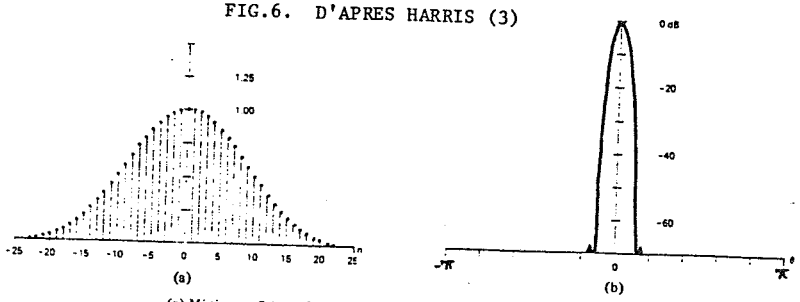


FIG. 7. CONSTRUCTION D'UN SIGNAL HARMONIQUE
PAR REPETITION D'UNE FONCTION D'ONDE

FIG. 6. D'APRES HARRIS (3)



(a) Minimum 3-term Blackman-Harris window. (b) Log-magnitude of transform.

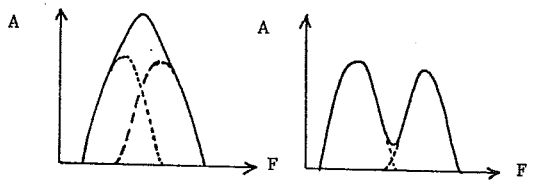


FIG. 8 a. RAPPROCHEMENT ET ELOIGNEMENT
DE DEUX FORMANTS A FLANCS INCLINES

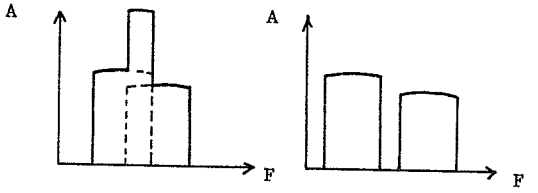


FIG. 8 b. RAPPROCHEMENT ET ELOIGNEMENT
DE DEUX FORMANTS A FLANCS DROITS

10èmes JOURNÉES D'ÉTUDE SUR LA PAROLE

GRENOBLE - 30 MAI - 1ER JUIN 1979

STRUCTURE D'UN SYNTHETISEUR DE PAROLE A PREDICTION LINEAIRE UTILISABLE EN PERIPHERIQUE D'ORDINATEUR

J.F. SERIGNAT - M. TIBI - D. DEGRYSE

E.N.S.E.R.G.
23 Avenue des Martyrs
38031 GRENOBLE CEDEX

RESUME

Nous décrivons un synthétiseur de parole à prédiction linéaire fonctionnant en temps réel qui peut être relié à un ordinateur par une ligne asynchrone comme un simple périphérique.

Le filtre de synthèse mis en oeuvre est du type en échelle à deux multiplieurs choisi en raison de sa grande stabilité et de son indépendance vis à vis de l'amplitude du signal d'exitation.

La structure se compose d'un filtre de synthèse (élaborant un échantillon du signal de sortie) associé à un microprocesseur qui contrôle les coëfficients du filtre en assurant leur interpolation et élabore le signal d'excitation du filtre.

Le microprocesseur est chargé également de gérer le protocole de communication sur la ligne asynchrone dont le débit peut être variable entre 2400 et 9600 bits par seconde.

L'intérêt du synthétiseur tient d'une part, à sa structure non figée (programmable) en ce qui concerne le nombre de coëfficients de réflexion utilisés, l'intervalle de temps entre les renouvellements de ces coefficients, la forme du signal d'excitation du filtre et la fréquence d'échantillonnage du signal de sortie, et d'autre part, à son mode de couplage banalisé (ligne asynchrone) avec un ordinateur.

STRUCTURE OF AN LPC SPEECH SYNTHESIZER USABLE AS A COMPUTER PERIPHERAL.

J.F. SERIGNAT - M. TIBI - D. DEGRYSE
E.N.S.E.R.G.
23 Avenue des Martyrs - 38031 GRENOBLE CEDEX

SUMMARY

We describe a real time linear prediction speech synthesizer which can be connected to a computer by an asynchronous line as a standard peripheral.

To set great stability and independence with the magnitude of the excitation signal a twomultiplier-lattice filter was chosen.

The synthesizer is composed of the lattice filter, computing one sample of the speech signal connected to a 8085 A microprocessor which checks the filter coefficients, performs their interpolation and generates the excitation signal.

The microprocessor has to manage the communication protocol over the asynchronous line with a transmission rate which may range from 2,400 to 9,600 bits per second.

The interest of this synthesizer lies in its programmable structure allowing easily changes of the reflection coefficient number, the time interval between two transferts, the excitation signal shape and the output signal sampling frequency. On the over hand, a standard coupling mode with a computer has been adopted.

10èmes JOURNÉES D'ÉTUDE SUR LA PAROLE
GRENOBLE - 30 MAI - 1ER JUIN 1979

STRUCTURE D'UN SYNTHETISEUR DE PAROLE A PREDICTION LINEAIRE UTILISABLE EN PERIPHERIQUE D'ORDINATEUR.

J.F. SERIGNAT - M. TIBI - D. DEGRYSE .

E.N.S.E.R.G.
23 Avenue des Martyrs
38031 GRENOBLE CEDEX

INTRODUCTION

On connaît depuis une dizaine d'années maintenant (ATAL B.S. - SCHROEDER M.R., 1968) l'intérêt que présente les méthodes de prédiction linéaire dans leur application à l'analyse et à la synthèse de la parole. L'éfficacité de ces méthodes, notamment dans le domaine de la compression de bande en télécommunications, n'est plus à démontrer et la tendance actuelle se situe dans la conception et la réalisation de matériels permettant d'appliquer ces méthodes en temps réel. En ce qui concerne la synthèse, de tels appareils apparaissent déjà aux Etats-Unis sous forme de composants (WIGGINS R. - BRANTINGHAM L., 1978 ; SMITH S., 1978), mais ils sont de structure figée car développés pour des applications particulières.

Le but poursuivi ici est la réalisation d'un synthétiseur de parole pouvant non seulement être utilisé facilement en sortie parlée d'ordinateur mais présentant également une structure programmable permettant de faire varier ses diverses caractéristiques telles que le nombre de paramètres utilisés, la fréquence de renouvellement de ces paramètres et leur quantification, la forme du signal d'excitation et la fréquence d'échantillonnage du signal de sortie. Ainsi, un tel synthétiseur sera considéré également comme un "appareil de laboratoire" rendant encore possibles des recherches sur l'amélioration de la synthèse (qualité, intelligibilité).

STRUCTURE DU FILTRE DE SYNTHESE

Les recherches effectuées sur les méthodes de prédiction linéaire appliquées à l'analyse - synthèse de la parole (ATAL B.S., HANAUER S.L., 1971 ; MAKHOUL J.I., 1973 ; SERIGNAT J.F., 1974) ont montré que celles-ci permettent de définir de façon directe, à partir du signal de parole, un modèle de production possédant la structure d'un filtre numérique récursif dont la fonction de transfert s'apparente à la fonction de transfert globale du système vocal (cordes vocales, conduit vocal, conduit nasal, etc...).

Or, parmi les différentes présentations possibles de ce filtre numérique récursif, les études récentes (MAKHOUL J.I., 1975 ; EL-MALLAWANY I., 1975 ; MARKEL J.D., GRAY A.H., 1976) ont démontré la supériorité des modèles utilisant les coefficients dits " de réflexion"$\langle k_n \rangle$ sur le modèle utilisant les coefficients dits "de prédiction" $\langle a_n \rangle$ bien que leurs structures soient en elles-mêmes plus complexes. Cette supériorité tient avant tout aux caractéristiques des coefficients de réflexion $\langle k_n \rangle$ qui ont un module toujours inférieur à l'unité et dont la sensibilité aux erreurs de quantification

est bien moins critique que celle des coefficients $\langle a_n \rangle$.

Pour cette raison, la structure du filtre utilisé en synthèse est celle du filtre numérique récursif en échelle dite " à deux multiplieurs" car chacune des cellules composant ce filtre nécéssite l'exécution de deux multiplications. C'est aussi la structure qu'a retenue TEXAS INSTRUMENTS pour définir son circuit de synthèse TMC 0280 (WIGGINS R., BRANTINGHAM L., 1978) dont une cellule "n" est représentée sur la Figure 1.

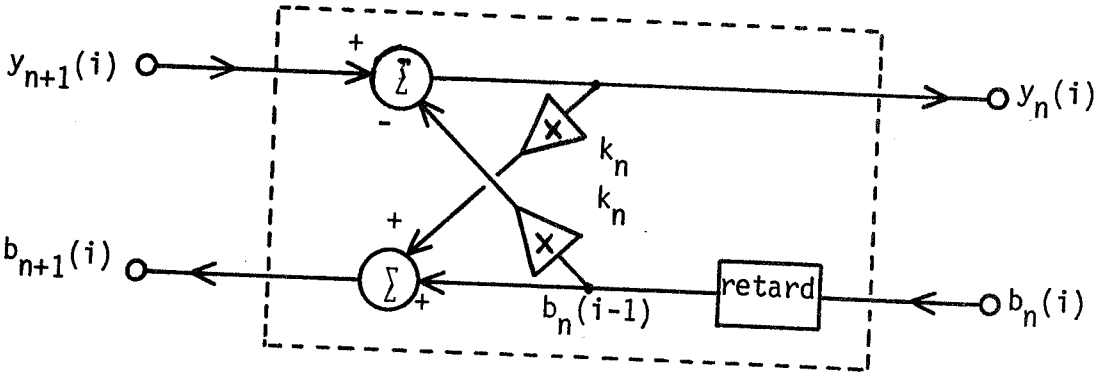


Figure 1. Cellule n / Stage n

$y_n(i)$ est l'échantillon de l'onde incidente à chaque cellule
$b_n(i)$ est l'échantillon de l'onde réfléchie à chaque cellule

$\langle k_n, n=1,2,..,N \rangle$ sont les coefficients de réflexion (N est de l'ordre du filtre)
i désigne l'instant iT considéré où T est la période d'échantillonnage du signal.

A chaque cellule, il y a donc deux multiplications et deux additions à effectuer et les équations exprimant la relation entre les variables y et b sont les suivantes :

$$\left.\begin{aligned} y_n(i) &= y_{n+1}(i) - k_n.b_n(i-1) \\ b_{n+1}(i) &= b_n(i-1) + k_n.y_n(i) \end{aligned}\right\} \quad n = N, N-1, \ldots, 1.$$

Le filtre complet est composé d'au maximum 14 cellules comme celle-là et le fonctionnement avec un nombre n de cellules inférieur à 14 est obtenu simplement en annulant les coefficients $<k_j, j = N+1, N+2,.. 14>$

L'ensemble des multiplications à effectuer pour élaborer un échantillon de signal à la sortie du filtre représenté Figure 2 sont réalisés à l'aide d'un seul multiplieurmultiplexé dans le temps entre toutes les cellules. De plus, l'algorithme appliqué correspond, dans une première phase, au calcul des différents échantillons $<y_n(i), n=14, 13,..., 1>$ de l'onde incidente pour l'instant d'échantillonnage i.T en fonction des échantillons $<b_n(i-1), n = 14, 13,...,1>$ de l'onde réfléchie élaborés à l'instant prédédent (i-1).T et mémorisés dans le filtre ; puis, dans une deuxième phase, on effectue le calcul des échantillons $<b_n(i), n = 14,13,...,1>$ de l'onde réfléchie pour l'instant iT considéré.

Une telle décomposition, qui est aussi celle retenue par TEXAS, autorise le recouvrement des cycles de calcul au niveau de l'exploitation du multiplieur ce qui n'est pas le cas si l'on élabore successivement à chaque cellule du filtre l'échantillon $y_n(i)$ de l'onde incidente et l'échantillon $b_{n+1}(i)$ de l'onde réfléchie puisque $b_{n+1}(i)$ est en fonction de $y_n(i)$.

Le filtre de synthèse, dont la structure synoptique est représentée Figure 3, est constitué principalement d'un multiplieur parallèle TRW du type MPY - 16 AJ effectuant les produits en complément à 2 des coefficients de réflexion $<k_n>$ (8 bits) par les échantillons $<y_n>$ ou $<b_n>$ (16 bits) ; les résultats étant fournis sur 16 bits. Il est suivi d'un additionneur/soustracteur (74 LS 181) sur 16 bits et d'un registre accumulateur sur 16 bits. La sortie de l'accumulateur peut être reliée, d'une part, à la fois à la deuxième entrée de l'additionneur/soustracteur (interrupteur I1) et à l'entrée d'une mémoire de type FIFO (Am 2812 A) qui doit contenir les échantillons $<y_n(i)>$ de l'onde incidente élaborés pendant la première phase et, d'autre part, à l'entrée d'une seconde mémoire de type FIFO (interrupteur I2) qui doit contenir les échantillons $<b_n(i)>$ de l'onde réfléchie élaborés pendant la deuxième phase. De plus, c'est aussi sur la sortie de l'accumulateur qu'est prélevé l'échantillon de signal de synthèse lors de l'élaboration de $y_1(i)$.

Par ailleurs, la sortie de type trois états de la première mémoire FIFO $<y_n(i)>$ communique par un bus avec une entrée du multiplieur ; la sortie de la seconde mémoire FIFO $<b_n(i-1)>$ peut être reliée au même bus (interrupteur I6) et à l'entrée de l'additionneur (interrupteur I3). De plus, l'interrupteur I5 permet la recirculation des échantillons $<b_n(i-1)>$ dans la mémoire FIFO et l'interrupteur I4 sert à enregistrer dans cette mémoire le dernier échantillon $b_1(i) = y_1(i)$ de l'onde réfléchie à la fin du calcul d'un échantillon de signal. Tous les interrupteurs qui assurent la communication entre les différents chemins de données sont des buffeurs à sortie trois états.

Les coefficients de réflexion $<k_n, n = 14, 13,..,1>$ sont enregistrés également dans une mémoire de type FIFO avec recirculation possible et l'échantillon d'excitation du filtre est placé dans un registre 8 bits.

L'ensemble de ce système logique est commandé par un séquenceur réalisé à l'aide d'un microprogramme enregistré dans une mémoire morte de 64 mots de 16 bits.

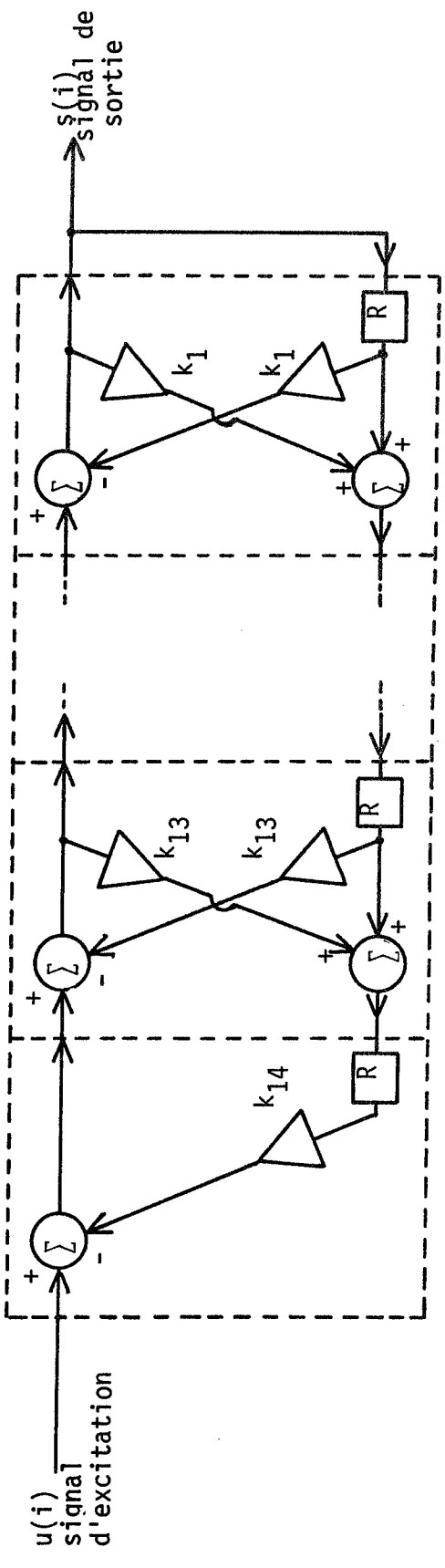


FIGURE 2 : Filtre de synthèse en échelle
Synthesis lattice filter

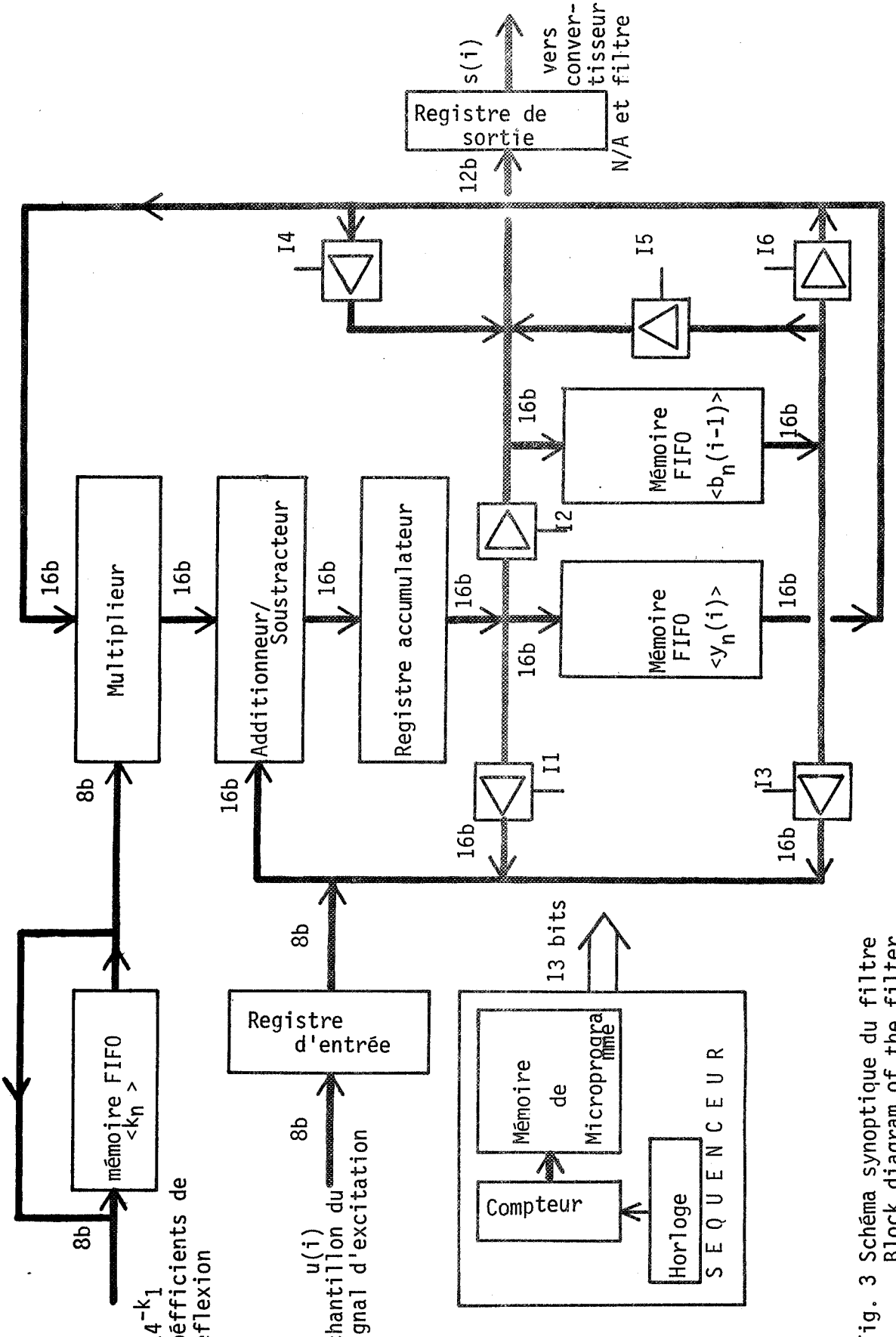


Fig. 3 Schéma synoptique du filtre
Block diagram of the filter

STRUCTURE COMPLETE DU SYNTHETISEUR

Le filtre de synthèse représenté à la figure 3 est contrôlé par un microprocesseur du type 8085 A qui assurent le renouvellement des coefficients de réflexion et des échantillons d'excitation du filtre.

Le microprocesseur est chargé également de l'interpolation des paramètres entre deux réceptions et de l'élaboration du signal d'excitation du filtre dans les cas voisé et non voisé.

Le système à microprocesseur, comme le montre la Figure 4, est composé de cinq circuits principaux qui sont :

- le circuit microprocesseur 8085A et son horloge ;
- un circuit 8755 A comportant principalement 2 K octets de mémoire morte reprogrammable pour le logiciel ;
- un circuit 8156 comportant 256 octets de mémoire vive pour les données, un circuit de temporisation et trois registres d'entrée sortie ;
- un circuit 8251 A (USART) interface de communication pour la réception des paramètres par la ligne asynchrone ;
- un circuit MC 14411 générateur de bits (B.R.G.) associé au 8251 A.

Le logiciel de commande, mémorisé dans le circuit 8755 A, peut être aisément modifié (puisqu'il s'agit de mémoire morte reprogrammable) pour changer, par exemple, les algorithmes d'interpolation des paramètres ou la forme du signal d'excitation du filtre.

L'ensemble des circuits du synthétiseur (filtre de synthèse + microprocesseur) occupe deux plaques de dimensions 14 x 18 cms environ.

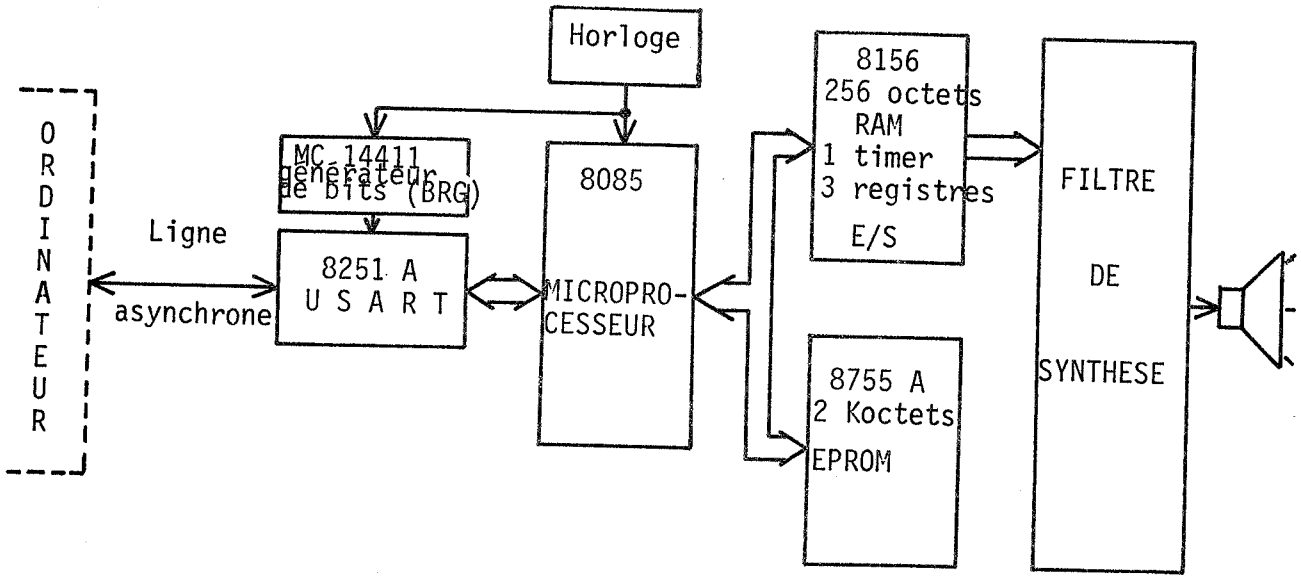


Figure 4 : Schéma synoptique du synthétiseur complet
Block diagram of the synthesizer

MODE DE COMMUNICATION DU SYNTHETISEUR AVEC UN ORDINATEUR

La communication entre le synthétiseur et un ordinateur se fait par une ligne asynchrone grâce au circuit d'interface 8251 qui réalise lui-même le décodage du code asynchrone et la conversion série-parallèle.

Le synthétiseur se comporte, vis à vis de l'ordinateur, toujours comme un esclave qui reçoit les paramètres lorsqu'ils lui sont envoyés ; cependant, à l'initialisation du système, le synthétiseur renvoie à l'ordinateur, sur une demande de sa part, un code particulier indiquant qu'il est en état correct de fonctionnement.

Les paramètres envoyés sur la ligne par l'ordinateur seront codés à raison d'un paramètre, quantifié sur 8 bits maximum, par caractère transmis pour permettre une mise au point rapide de l'ensemble du logiciel. Ultérieurement, on pourra envisager, dans le but de réduire le débit de transmission sur la ligne, un codage plus complexe en regroupant plusieurs paramètres par caractère transmis lorsque ceux-ci sont quantifiés avec un nombre de bits inférieur à 8.

De plus, avant l'utilisation du synthétiseur en tant que périphérique, l'ordinateur doit lui fournir, aprés un code particulier de synchronisation, les caractéristiques de la synthèse à effectuer, c'est à dire, principalement le nombre de coefficients de réflexion et la fréquence de leur renouvellement, de même que le nombre d'interpolations dans une séquence et la fréquence d'échantillonnage du signal de sortie.

CONCLUSION

Parmi les méthodes de synthèse de la parole, la prédiction linéaire a prouvé qu'elle était l'une des plus efficaces pour obtenir de la parole intelligible et même d'assez bonne qualité. L'intérêt est donc grand de disposer d'un synthétiseur à prédiction linéaire fonctionnant en temps réel.

A une époque où commencent à apparaître sur le marché des circuits intégrés de synthèse à prédiction linéaire de structure figée car ils sont développés pour des applications spécifiques, nous avons jugé tout de même intéressant de réaliser un tel synthétiseur en circuits discrets et sous une forme assez réduite, à condition de lui conserver une structure facilement programmable et un mode de couplage standardisé (ligne asynchrone) avec tout type de calculateur.

BIBLIOGRAPHIE

ATAL B.S., SCHROEDER M.R., 1968. Predictive coding of speech signals. 6th Int. Cong. on Acoustics, C 5.4, C13-C16, Tokyo.

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

MAKHOUL J.I., 1973. Spectral analysis of speech by linear prediction. IEEE. Trans., AU 21, pp 140-148

MAKHOUL J.I., 1975. Linear prediction : a tutorial review. Proceedings of I.E.E.E. 63, pp 561-580

MARKEL J.D., GRAY A.H., 1976. Linear prediction of speech. Springer-Verlag, Berlin, Heidelberg, New-York

EL-MALLAWANY I., 1975. Etude de vocodeursà prédiction linéaire-Détermination de l'intervalle de fermeture de la glotte- Détection de mélodie-Extraction de la fonction d'aire du conduit vocal. Thèse de Docteur Ingénieur, I.N.P.Grenoble, Sep. 1975.

SERIGNAT J.F., 1974. Contribution aux recherches sur la communication parlée : travaux sur le vocodeurà autocorrélation, étude et simulation d'un vocodeur à prédiction linéaire. Thèse de Docteur Ingénieur, I.N.P.Grenoble, Juillet 1974.

SMITH S., 1978. Single-chip speech synthesizers. Computer Design Int. Review pp 188-194

WIGGINS R., BRANTIGHAM L., 1978. Three-chip system synthesizes human speech. Electronics Int. Review, 18, pp 109-116.

THEME I (b)

SYNTHESE DE LA PAROLE

Applications - Modèles articulatoires

10èmes JOURNÉES D'ÉTUDE SUR LA PAROLE

GRENOBLE - 30 MAI - 1ER JUIN 1979

LA GEOMETRIE DES LEVRES EN FRANCAIS - Protrusion vocalique et protrusion consonantique.

C. ABRY, L.J. BOË, M. GENTIL, & R. DESCOUT, P. GRAILLOT — Institut de Phonétique de Grenoble & CNET Lannion.

RESUME Nous avons analysé, en vue de la modélisation de la géométrie des lèvres, un corpus de voyelles fermées arrondies [y ø] et non arrondies [i e] en coarticulation avec les sibilantes [s z] et [ʃ ʒ]. Sur 96 réalisations x 5 locuteurs, 12 paramètres ont été mesurés. L'analyse en composantes principales menée avec ces paramètres nous permet de n'en retenir que deux : l'aire S et le facteur de forme K_2 (rapport des deux axes de l'orifice labial) ou le coin des lèvres C, ces deux derniers étant corrélés.

Sur ces trois, S est le seul qui corresponde à la frontière phonémique entre les voyelles [+ rond] et [- rond] que celles-ci soient assimilées ou non. Pour les consonnes, par contre, une telle séparation à 100% n'est généralement pas possible, que ce soit avec S, K_2 ou C.

L'examen des dispersions montre que les voyelles et les consonnes n'ont pas toutes les mêmes comportements du point de vue du maintien des constantes, que ce soit au niveau proprement articulatoire (K2) ou déjà acoustico-articulatoire (S) : les voyelles [- rond] ont une meilleure constante d'aire que les voyelles [+ rond]. Celles-ci et toutes les consonnes ont, par contre, une meilleure constante de forme (celle des consonnes étant généralement la meilleure).

Le maintien de l'aire aux lèvres semblerait ainsi être plus important pour les voyelles [- rond] puisqu'elles peuvent subir une assimilation de protrusion (type [ʃi_{ω}]). Ceci n'est pas le cas des voyelles [+ rond] qui assimilent toujours les consonnes. Quant à celles-ci elles se distinguent suffisamment au niveau de l'articulation linguale.

Mais les trois paramètres retenus ne permettent pas seulement de synthétiser, grâce à S et C, des fonctions d'aire qui préservent, aux lèvres, une distinction phonologique essentielle en français. Grâce à S, K_2 et C, il est possible d'obtenir une bonne géométrisation du pavillon labial en trois dimensions: de face A, grand axe et B petit axe, déterminent la courbure des lèvres; de profil C, généralement bien corrélé aux protrusions supérieure et inférieure, localise à l'extrémité du conduit vocal le biseau des lèvres.

LIP SHAPING FOR FRENCH PROTRUDED VOWELS AND CONSONANTS

C. ABRY, L.J. BOË, M.GENTIL, & R. DESCOUT, P. GRAILLOT — Institut de Phonétique de Grenoble & CNET Lannion

SUMMARY

For purpose of geometrical modeling of the lips, we have analyzed data on French rounded [y ø] and spread vowels [i e], in protrudedand non-protruded CV contexts, with [s z] and [ʃ ʒ] (i.e. 96 combinations in short utterances). A labiographic film has been shot for 5 speakers. 12 parameters have been measured of which factor analysis allows us to retain only 2 : lip area S, and K_2, a shape coefficient (ratio of lip axes) or C, lip outmost corner (these last two being correlated).

Among these, only S can achieve a 100% discrimination corresponding to the phonetic contrast [+ rond / - rond] and that only for vowels.

Analysis of vowels and consonants distributions shows differences of behavior as to the maintenance of articulatory (K_2) and acoustico-articulatory (S) constants : only [- round] vowels present better constance for area than for shape, a fact presumably ascribable to a more important acoustic sensitivity at the vocal tract termination when [- round] vowels are protruded (e.g.[ʃi_{ω}]). [+ round] vowels on the contrary are never assimilated : they assimilate consonants and these ones in fact do not need any constance of lip area, since they are essentially contrasted by lingual articulation.

The three parameters retained allow not only, like S and C, to synthetize good area functions that preserve, at the lips, contrasts essentiel to French; more they provide a satisfactory geometrization of the vocal tract output : available from S and K2, A and B, the frontal axes determine lip curvature and, laterally, C, pretty well correlated with upper and lower protrusions, gives us the lip bevel profile and location.

10èmes JOURNÉES D'ÉTUDE SUR LA PAROLE
GRENOBLE - 30 MAI - 1ER JUIN 1979

LA GEOMETRIE DES LEVRES EN FRANCAIS - Protrusion vocalique et protrusion consonantique.

C. ABRY, L.J. BOË, M. GENTIL, & R. DESCOUT, P. GRAILLOT — Institut de Phonétique de Grenoble & CNET Lannion.

INTRODUCTION

Alors que la phonétique peut depuis quelques temps déjà fournir des informations quantifiées sur les positions linguales, la modélisation de la sortie du conduit vocal manque encore des données les plus élémentaires sur la géométrie des lèvres*.

Les recherches que nous avons entreprises depuis plus d'un an doivent conduire à la mise au point d'un tel modèle pour le jeu des lèvres en français, aux deux niveaux de la géométrie et de la motricité labiales; la première étape étant le pilotage d'un analogue dynamique du conduit vocal.

Toute modélisation d'un stade quelconque de la production de la parole implique un certain nombre d'options, en particulier celles qui concernent l'évaluation du modèle à sa sortie. Cette évaluation peut être en quelque sorte "court-circuitée" si elle se fait à une étape en aval de celle qui est effectivement modélisée : ainsi nombre d'analogues du conduit vocal ont été testés sur leur sortie acoustique, voire sur leur efficacité perceptive (synthèse acceptable), et non directement sur leur bonne prédiction géométrique.

Ces exigences qui situent le "niveau de réalité" du modèle ont des conséquences immédiates en ce qui concerne les données phonétiques qui doivent lui être fournies. Ainsi de ce qu'un [s] coloré de [i] est toujours correctement identifié (CHAFCOULOFF & DI CRISTO, 1978) quel que soit le contexte - alors qu'il n'en va pas de même d'un [s] coloré de [u] - on peut être amené, pour des raisons d'économies, à ne donner au modèle qu'une seule possibilité de réalisa-

*En ce qui concerne le français, il n'existait jusqu'à présent, sur la géométrie des lèvres, rien de comparable à l'étude de V. FROMKIN (1964) pour l'anglo-américain. Pour le suédois, cf. LINDBLOM (1965).

tion articulatoire de cette sibilante (ici [s] à lèvres étirées). Par contre si la raison d'économie n'est pas externe au système modélisé - s'il s'agit en particulier de prédire la coarticulation - les sorties phonétiques reproduites devront être aussi proches que possible des données d'observation articulatoires et les discussions "économiques" auront gagné un niveau en abstraction, celui des cibles.

Le fait que nous nous soyons situés à ce niveau de réalité permet de comprendre les raisons du choix de notre corpus. Nos conditions ont été les suivantes :

1 - Etudier le jeu des lèvres lorsqu'il est maximal, d'où le choix de la composante de protrusion du mouvement labial (plutôt que la seule "compression" cf. LADEFOGED, 1975); d'où aussi le choix des voyelles fermées pour lesquelles la protrusion-rétraction est la plus importante. Cette condition permet ainsi d'explorer en grande partie les frontières de l'espace articulatoire de la labialité. Le choix du français a facilité notre tâche : le jeu des lèvres y est en effet, de notoriété, très marqué.

2 - L'étude des maxima du jeu de la protrusion dans les espaces articulatoires des consonnes et des voyelles aurait pu se faire en minimisant les effets de coarticulation labiale (en les supprimant même pour les voyelles isolées). Mais, de même qu'il est peu pertinent de fournir à la modélisation les valeurs d'un [k] "moyen" qui recouvrirait des réalisations aussi hétérogènes que [ki] et [ku], obtenir des valeurs de [i,e,...] sans explorer leurs variantes contextuelles maximales, du point de vue des lèvres, aurait été d'une faible utilité. Nous pourrons donc définir dans l'espace de la labialité, non seulement les zones vocaliques et consonantiques, mais aussi les trajectoires de coarticulation (par exemple [i_{ω}] derrière [ʃ] , etc.).

3 - Le fait que le jeu des lèvres entre, en français, dans les oppositions phonologiques (au moins au niveau de la phonologie de surface, SCHANE (1969, p. 21)) peut permettre d'appliquer un principe tel que l'assimilation-préservant-les-oppositions (LINDBLOM, 1978, p. 140). Ce principe assigne des limites aux trajectoires d'assimilation pour que soient maintenues certaines constantes relatives qu'il importe de rechercher au niveau modélisé (ex. les constantes de forme de la configuration labiale), ou plus en aval (comme les constantes d'aire).

Parmi les 12 paramètres que nous avons relevés sur nos observations latérales et frontales du pavillon labial, nous proposerons que l'on retienne pour la modélisation de sa géométrie ceux qui :

- maintiennent le mieux les oppositions phonologiques,
- donnent la meilleure prédiction sur leurs partenaires.

LES PARAMETRES

Le corpus, les locuteurs, les conditions d'enregistrement, la technique d'acquisition des données sont les mêmes que pour DESCOUT & al. (1978). Plusieurs paramètres ayant été ajoutés, la liste de ceux-ci est la suivante :

. De face
- A → l'écartement intérolabial — K_1 → S/AB
- B → l'aperture intérolabiale — K_2 → A/B
- M → la position du maxillaire — K_3 → A+B
- S → l'aire intérolabiale

. De profil
- F_1 F_2 → la protrusion-rétraction des lèvres supérieure et inférieure
- C → le point de contact des lèvres
- D → l'aperture intérolabiale, à l'extrémité du conduit vocal
- L → la distance entre C et la tangente F_1F_2

ANALYSE EN COMPOSANTES PRINCIPALES

Les deux premières composantes principales dégagées par l'analyse factorielle menée avec ces 12 paramètres expliquent, à elles seules,95% de la variance; l'axe 1, pour sa part, rend compte d'au moins 80% de celle-ci (de 81 à 87%).

La projection des paramètres révèlent que :

- la plus forte contribution à l'axe 1 est donnée par S;
- K_2 et C apportent chacun une contribution sensiblement équivalente à l'axe 2; ces deux paramètres sont en outre corrélés négativement.

Si l'on projette les données et les paramètres dans le plan des composantes principales dégagées avec seulement S, K_2 et C, on constate une répartition très voisine de celle que l'on a obtenue avec les paramètres au complet. Rappelons que S et K_2 permettent de retrouver les dimensions A et B (connaissant la valeur de K_1) et que C est assez bien corrélé avec F_1 et F_2, ce qui nous donne une bonne approximation de la forme du pavillon labial et du degré de

protrusion/rétraction; d'autre part, alors que K_2 peut s'interpréter en termes articulatoires S et C permettent le passage au niveau acoustique.

En examinant les répartitions des projections par rapport au plan principal, les tendances générales sont les suivantes :
- l'axe 1 oppose les voyelles arrondies aux non-arrondies;
- l'axe 2 les consonnes, c'est-à-dire les sibilantes arrondies [ʃ ʒ] et les non-arrondies [s z];

Si l'on envisage maintenant les séparations à 100 % des termes de ces oppositions, on constate que :
- Pour tous les locuteurs, les voyelles [+ rond] n'ont besoin que du premier axe pour se distinguer des voyelles [- rond] et ceci tous contextes confondus, c'est à dire [y y_ω ø $ø_\omega$]/ [i i_ω e e_ω] (excepté deux réalisations [ø] d'un seul locuteur, sur 12 occurrences de cette voyelle).
- Pour les consonnes, les séparations sont beaucoup moins nettes. Avec un seul axe, le deuxième, on ne peut effectuer cette séparation que pour un seul locuteur (à une occurrence près). Pour les autres, les deux axes sont nécessaires, et pour un locuteur, il est impossible de séparer les réalisations consonantiques.

L'aire aux lèvres S assure donc l'opposition et la séparation des voyelles quel que soit le contexte consonantique; le coefficient de forme K_2 et la protrusion-rétraction C opposent les consonnes. Le fait que la séparation soit moins bonne pour les consonnes ne doit pas nous étonner : en effet, en français, la différenciation des sibilantes ne se fait pas essentiellement au niveau labial; nos données correspondent donc simplement au fait que l'arrondissement consonantique peut être considéré, dans notre langue, comme redondant.

LA FORME ET L'AIRE

Les résultats de l'analyse en composantes principales nous permettent de ne retenir que deux des trois paramètres : K_2 et C étant corrélés, nous avons le choix entre S, K_2 ou S,C. Connaissant K_1, S et K_2 nous permettent de retrouver A et B; par contre, S et C ne nous en donneraient qu'une estimation (K_2 par corrélation). Nous avons donc retenu S et K_2 qui nous donnent la meilleure précision sur le plus grand nombre de paramètres. Nous conserverons ainsi, le plus

fidèlement, les dimensions spatiales du pavillon labial.

Dans le plan S/K_2 (cf. fig. 1-5), l'analyse nous apporte pratiquement les mêmes informations sur les séparations : tous contextes confondus, S suffit pour discriminer à 100% les voyelles, une droite à pente négative n'étant nécessaire que pour un seul locuteur. Pour les consonnes, les séparations ne sont possibles que pour deux locuteurs : l'une avec K_2, l'autre avec S et K_2 (droite à pente positive).

Ces paramètres peuvent, en outre, se prêter à des interprétations physiologiques et acoustiques intéressantes. Si l'on admet qu'une meilleure constante de forme se situe au niveau articulatoire et qu'une meilleure constante d'aire est déjà davantage"orientée vers la sortie", il vaut la peine de se demander quelle est, des deux constantes, la mieux maintenue et ceci respectivement pour les voyelles et pour les consonnes (cf. tableaux I et II).

. Les voyelles

Pour tous les locuteurs, la variance relative* de l'aire est plus petite pour [i e] que pour [y ø] ; celle de forme est pratiquement toujours plus élevée pour [i e]. Sur la dispersion d'un même groupe de voyelles [+ rond] ou [- rond] il est possible d'évaluer si "l'accent" est mis plutôt sur la constante de forme que sur celle d'aire ou inversement. Le rapport des variances relatives de S et K_2, est toujours supérieur à 1 pour les voyelles [+ rond], inférieur ou égal* à 1 pour les voyelles [- rond]. On peut donc dire que les voyelles phonologiquement [- rond] privilégient, relativement à leur aire aux lèvres plus grande, plutôt <u>une constante d'aire</u>, par rapport aux voyelles [+ rond] mais aussi par rapport à leur propre dispersion où la variance de S est en général plus petite que celle de K_2.

. Les consonnes

La variance relative de l'aire est plutôt plus petite pour [ʃ ʒ] que pour

*C'est à dire exprimée par rapport à la moyenne.

*Le seul cas où il atteint cette valeur étant celui du locuteur qui présente un comportement de ses voyelles très similaire à celui des consonnes (cf. ABRY & al., 1979, p. 8).

[s z]; celle de la forme est tantôt en faveur de [s z], tantôt en faveur de [ʃ ʒ]. Pour toutes les consonnes, le rapport des variances relatives de S et K_2 est, contrairement à celui des voyelles, supérieur à 1. La variance de la forme des consonnes, tous contextes confondus, est donc relativement moindre que celle de leur aire.

Il reste à interpréter ces résultats en fonction du niveau acoustique. Il semblerait, dans une première approche, que l'aire aux lèvres étant plus importante pour maintenir les oppositions vocaliques, et ceci quelle que soit l'assimilation consonantique, il existe un certain contrôle - orienté vers la sortie - de la constance relative de ce paramètre. Pour les consonnes, qui ne semblent pas requérir une telle "maîtrise" au niveau des lèvres (les sibilantes restent suffisamment distinctes intrabuccalement), leur aire peut subir de fortes modifications contextuelles, leur forme articulatoire restant plus ou moins constante.

GEOMETRIE ET DEGRES DE LIBERTE

La modélisation de la géométrie des lèvres, pour les voyelles et les consonnes qui mettent en jeu la protrusion-rétraction, en retenant les trois paramètres S, K_2 et C - soit pratiquement A, B, K_1 et C - retient en fait les principaux degrés de liberté des lèvres. En effet, nous avons montré (ABRY & al., 1979), qu'il existait une relative indépendance de A et B pour les voyelles comme pour les consonnes. Rappelons aussi qu'il est tout à fait possible de définir dans le plan A/B, les distinctions retrouvées dans S/K_2, soit l'opposition [+ rond/- rond]: pour les voyelles, tous contextes confondus, et pour les consonnes seulement dans le contexte vocalique [- rond] (l'assimilation étant pratiquement totale dans le contexte [+ rond]). A et B permettent donc une bonne synthèse de la forme de l'orifice labial : ils définissent les arcs qui approximent les bords intérieurs des vermillons supérieur et inférieur (LINDBLOM & SUNDBERG, 1971) :

$$y = \pm\ B\ [1 - (2/B)^p |x|^p] \quad \text{avec} \quad p = K_1/1-K_1$$

Si l'on se fixe S, en se donnant A, on pourra ainsi mieux prendre en compte la pondération du produit A x B par K_1. En ce qui concerne la troisième dimension spatiale, A et C, sont généralement bien corrélés (négativement; mieux que

B et C) : la rétraction des lèvres entraînant "l'étirement" (sur A) et leur protrusion le contraire. La prédiction de C devenant pourtant relativement moins bonne lorsque la protrusion est importante, la synthèse pourra être améliorée pour les voyelles et les consonnes arrondies en se donnant la valeur de C.

REFERENCES

ABRY, C., BOË, L.J., DESCOUT, R., 1979, Voyelles labiales et voyelles labialisées en français. Etude labiographique; 9th Int. Congr. Phonetic Sci., Copenhague.

CHAFCOULOFF, M. & DI CRISTO, A., 1978, Les indices acoustiques et perceptuels des constrictives du français. Application à la synthèse; 9e JEP/GALF, pp. 69-81.

DESCOUT, R., BOË, L.J., ABRY, C., 1978, Labialité vocalique et consonantique en français. Premiers résultats; 9e JEP/GALF, pp. 179-189.

FROMKIN, V., 1964, Lips Positions in American English Vowels; Language & Speech, 7, pp. 215-225.

LADEFOGED, P., 1975, A Course in Phonetics; Harcourt Brace Jovanovich, New York.

LINDBLOM, B., 1965, Analysis of Labial Movement; QPSR/RIT 2, pp. 20-22.

LINDBLOM, B., 1965, Studies of Labial Articulation; QPSR/RIT 4, pp. 7-9.

LINDBLOM, B., 1965, Jaw-Dependance of Labial Parameters and a Measure of Labialization; QPSR/RIT 3, p. 12.

LINDBLOM, B., 1978, Phonetic Aspects of Linguistic Explanation; Studia Linguistica 32, I-II, pp. 137-153., trad. française in : Bulletin de l'Institut de Phonétique de Grenoble 7.

LINDBLOM, B.E.F., & SUNDBERG, J.E.F., 1971, Acoustical Consequences of Lip, Tongue, Jaw and Larynx Movement; JASA 50, pp. 1166-1179.

SCHANE, S.A., 1969, French Phonology and Morphology; M.I.T. Press, Cambridge.

Fig. 1

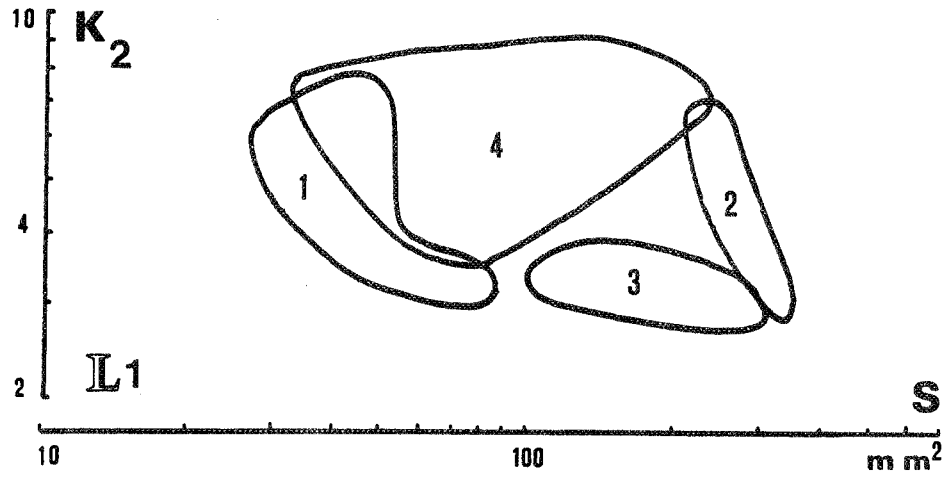


Fig. 2

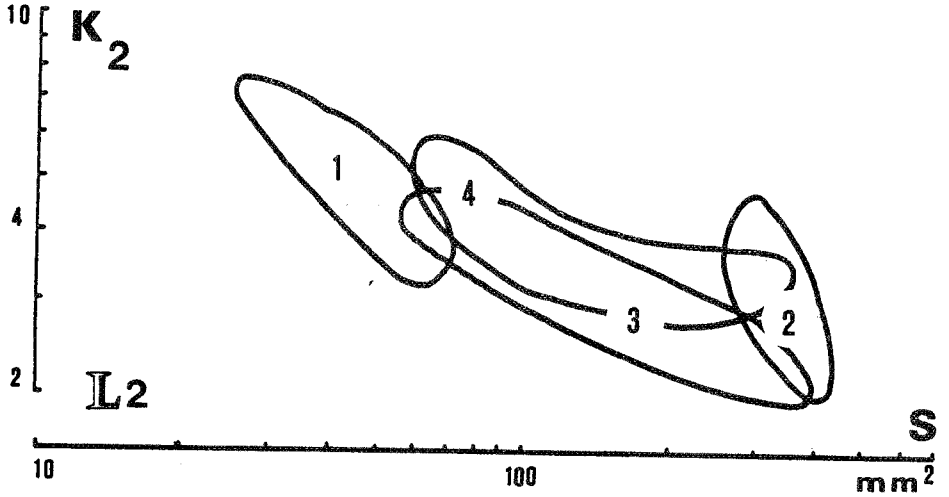


Fig. 3

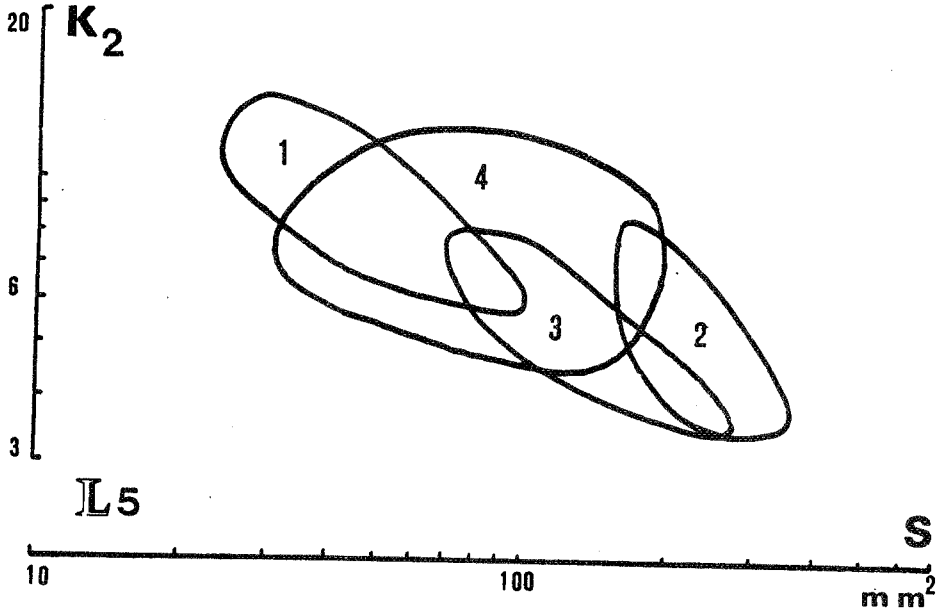


Fig. 4

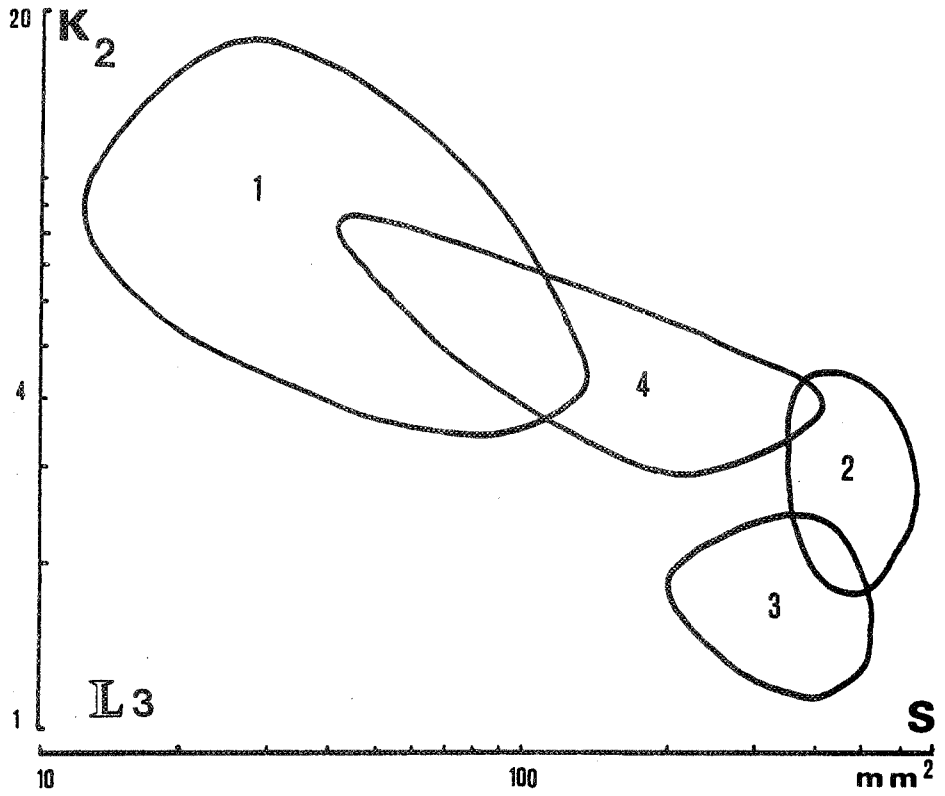
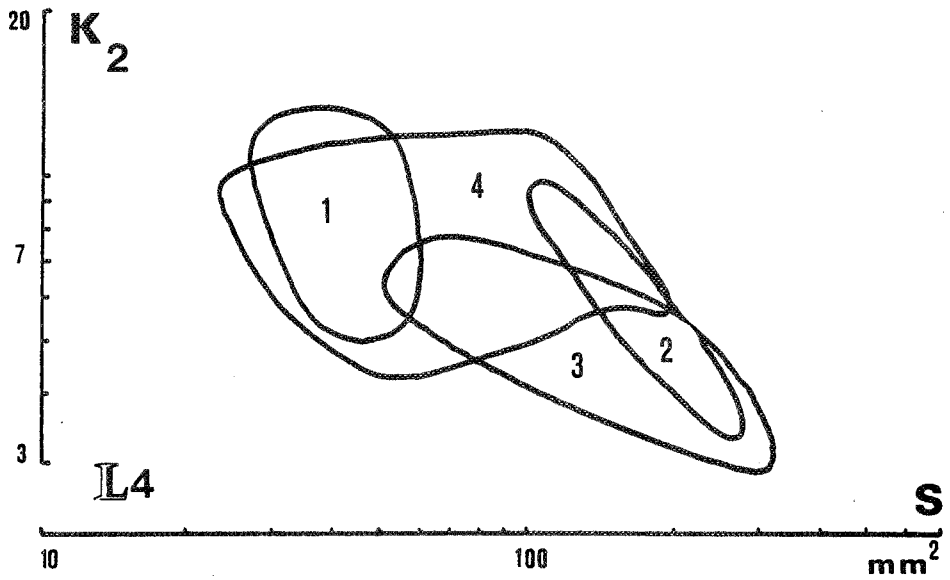


Fig. 5



Figures 1 à 5

Répartition des réalisations consonantiques et vocaliques [± rond] dans le plan S/K_2 pour les 5 locuteurs (L_1 à L_5).

Zones 1 : [y, ø] 3 : [ʃ, ʒ]

2 : [i, e] 4 : [s, z]

	L_1		L_2		L_3		L_4		L_5	
	i e	y ø	i e	y ø	i e	y ø	i e	y ø	i e	y ø
$\frac{\sigma}{m}$ (S)	0,20	0,38	0,16	0,26	0,18	0,50	0,29	0,40	0,22	0,44
$\frac{\sigma}{m}$ (K_2)	0,47	0,29	0,25	0,19	0,36	0,49	0,29	0,27	0,30	0,25
$\frac{\frac{\sigma}{m}(S)}{\frac{\sigma}{m}(K_2)}$	0,42	1,31	0,64	1,37	0,50	1,02	1,00	1,48	0,73	1,76

	L_1		L_2		L_3		L_4		L_5	
	s z	ʃ ʒ	s z	ʃ ʒ	s z	ʃ ʒ	s z	ʃ ʒ	s z	ʃ ʒ
$\frac{\sigma}{m}$ (S)	0,37	0,48	0,45	0,41	0,49	0,30	0,64	0,55	0,45	0,39
$\frac{\sigma}{m}$ (K_2)	0,21	0,20	0,22	0,26	0,29	0,19	0,25	0,30	0,28	0,28
$\frac{\frac{\sigma}{m}(S)}{\frac{\sigma}{m}(K_2)}$	1,76	2,40	2,04	1,57	1,68	1,57	2,55	1,83	1,60	1,39

TABLEAUX I et II - Ecart type relatif de l'aire S et de la forme K_2 pour les voyelles et les consonnes.

10èmes JOURNÉES D'ÉTUDE SUR LA PAROLE

GRENOBLE - 30 MAI - 1ER JUIN 1979

APPORTS ET LIMITES DES SYNTHETISEURS DE PAROLE DANS LE DOMAINE DE LA REHABILITATION DES ENFANTS DEFICIENTS AUDITIFS

S. BARTH D. GRENIER

Institut National de Jeunes Sourds de CHAMBERY
Centre d'Audiologie et d'Acoustique - B.P. 15 73160- COGNIN

RESUME

La réhabilitation des enfants déficients auditifs est soumise à la réalisation de certains impératifs pour laquelle, l'utilisation des possibilités des synthétiseurs de parole pourrait présenter un intérêt non négligeable. Nous résumerons les apports possibles de telles techniques et, à partir des résultats d'une expérimentation, nous essaierons aussi d'en cerner les limites actuelles.

CONTRIBUTIONS AND LIMITS OF SPEECH SYNTHETISERS IN THE DOMAIN OF DEAF CHILDRENS ' REHABILITATION

S. BARTH D. GRENIER

SUMMARY

The rehabilitation of deaf children is liable to some requirements for the satisfaction of which the use of speech synthetisers should offer a great range of possibilities. We shall summarize the possible contributions of these techniques and, from the results of an experimentation, we shall also try to define their present limits.

10èmes JOURNÉES D'ÉTUDE SUR LA PAROLE
GRENOBLE - 30 MAI - 1ER JUIN 1979

APPORTS ET LIMITES DES SYNTHETISEURS DE PAROLE DANS LE DOMAINE DE LA REHABILITATION DES ENFANTS DEFICIENTS AUDITIFS

S. BARTH D. GRENIER

INTRODUCTION

La déficience auditive prive, totalement ou partiellement, celui qui en est atteint, d'un canal sensoriel nécessaire à une activité propre à l'être humain : la communication parlée. De ce fait, elle peut être considérée comme un handicap social dont la gravité dépend du degré de dégradation des facultés auditives du sujet.

Depuis de nombreuses années, une certaine somme de connaissances, tant empiriques que théoriques, se sont accumulées sur ce problème. Si à l'heure actuelle, celui-ci n'a pas reçu de solution définitive, on a pu, néanmoins, cerner certains impératifs auxquels est soumise la réhabilitation des enfants sourds. Parmi ceux susceptibles de nous intéresser dans cet exposé, nous trouvons :

- Le test de la valeur sociale :

  . des restes auditifs qui se pratique au moyen d'épreuves de reconnaissance de mots issus de listes phonétiquement équilibrées (LAFON) ou faisant appel à la théorie des paires minimales (BERAHA).

  . de l'appareillage qui se différencie du précédent par l'utilisation d'appareils de correction auditive.

- L'éducation auditive qui a pour but d'apprendre à l'enfant à tirer partie de son audition résiduelle et de son appareil.

- L'apprentissage ou "démutisation" et l'entretien de la parole par un suivi orthophonique.

Dans les lignes qui suivent, nous nous intéresserons aux surdités dites de perception, qui concernent principalement les atteintes de l'oreille interne, pour lesquelles il n'existe pas encore, à notre connaissance, de techniques chirurgicales palliatives totalement fiables.

AUDIOMETRIE VOCALE

Qu'elle se fasse avec ou sans appareil de correction auditive, l'audiométrie vocale est essentielle pour évaluer un déficit de la communication dû à une surdité.

Lorsque l'on présente de la parole à un sujet souffrant d'une déficience auditive de perception, les traits acoustiques pertinents peuvent ne pas être perçus ou subir une dégradation importante. Divers facteurs influent sur ces "bruits" de transmission :

- la perte de sensibilité absolue

- la réduction du champ auditif sans que, parallèlement, soient améliorées les possibilités de discrimination dans ce champ restreint

- les modifications des seuils différentiels de fréquence d'amplitude et de temps ou (et) une croissance anormale de l'intensité sonore perçue par rapport à une certaine augmentation de l'intensité physique du signal (phénomène de recrutement)

- la fatigue auditive, entrainant une diminution de l'intensité perçue lorsque la durée du stimulus augmente

- les problèmes de masque fréquentiel et temporel.

La synthèse de la parole offre aux expérimentateurs de grandes possibilités en permettant de faire varier un certain nombre de paramètres et d'en tester la pertinence sur le plan de la perception. Notons, parmi les travaux déjà publiés, ceux de PICKETT et DANAHER (1973, 1975) PICKETT et MARTONY (1970), MARTONY (1974), FOURCIN (1975), BERAHA (1977). Certains d'entre eux étudient la discrimination de transitions formantiques de voyelles synthétiques, d'autres celles de mots. Ils engendrent deux constatations :

1) L'utilisation d'un synthétiseur de parole permet d'avoir des stimuli normalisés reproductibles en donnant une plus grande marge de variabilité que les systèmes de traitement analogique (filtres ...).

2) L'interprétation des résultats de tests de reconnaissance de mots synthétiques est parfois plus difficile que celle faite à partir de listes prononcées par un locuteur humain.

Pour argumenter ce deuxième point, nous avons effectué l'expérimentation suivante :

Sur une liste de 100 mots enregistrés grâce à l'amabilité des chercheurs du LIMSI (Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur - ORSAY), nous avons sélectionné 50 stimuli, de façon tout d'abord, à éviter les mots trop proches auditivement (ex. : cadeau et gâteau) et ensuite à être sûrs que tous les mots utilisés étaient connus des sujets testés. La liste obtenue a été présentée à 15 élèves de l'établissement atteints de surdité sévère (perte auditive moyenne comprise entre 70 et 90 dB calculée suivant les normes du Bureau International d'Audiophonologie) et âgés d'une dizaine d'années.

L'audition de celle-ci était précédée par une courte période d'adaptation à la voix de l'Icophone V au cours de laquelle nous avons proposé à l'enfant quelques petites phrases, des voyelles et les chiffres. Après chaque mot, le sujet disposait de 10 secondes pour transcrire ce qu'il avait perçu.

Dans un deuxième temps, nous avons fait entendre, dans un ordre différent, la même liste de mots, aux mêmes sujets, mais enregistrée cette fois avec un locuteur masculin dont la voix était inconnue des élèves.

Plusieurs mois séparaient les deux phases de l'expérimentation afin d'éliminer, dans une large mesure, l'apprentissage. Les stimuli étaient présentés par l'intermédiaire d'un magnétophone REVOX A77 équipé d'une bande SONY HL et d'une conque ELIPSON 35w.

Le niveau sonore d'émission était réglé pour que l'écoute soit jugée confortable par les sujets (en moyenne 70 dB SPL au niveau de leurs prothèses individuelles).

Le diagramme de la figure I positionne les mots de la liste dans le plan (pourcentage de reconnaissance ICOPHONE, pourcentage de reconnaissance voix humaine).

Le tableau II donne les résultats globaux, pour chaque sujet, en pourcentage de mots reconnus pour chacun des deux locuteurs.

On observe une nette supériorité de la voix humaine par rapport à la voix de synthèse. Cette observation est interprétable d'après les travaux de DANAHER, OBSBERGER et PICKETT (1973) qui montrent que la discrimination des transitions formantiques à des niveaux d'écoute confortable, est affectée, pour des surdités de perception :

1) par la région où apparait la transition du F2

2) l'amplitude relative de F1 par rapport à F2

3) la présence de la transition de F1

4) la proximité de F1 et F2

Ces auteurs insistent particulièrement sur les effets de masque produits par la présence de F1 (mis en évidence par la suppression de la transition de ce formant) difficilement prévisibles à partir des audiogrammes tonaux (mesure des pertes auditives en dB à 250, 500, 1000, 2000, 4000, 8000 Hz). Si l'on se rappelle l'allure assez plate du spectre de la voix de l'ICOPHONE V, les résultats que nous avons obtenus ne sont pas étonnants.

Toutefois nous devons souligner que les scores obtenus pour l'ICOPHONE V

sont loin d'être négligeables (environ 1 mot sur trois en moyenne contre 1 mot sur 2 pour la voix humaine), ce qui nous pousse à penser que ces expérimentations doivent se poursuivre. D'autre part, certains de nos sujets ont été surpris par le caractère peu naturel de la prosodie du synthétiseur. (rappelons que les enregistrements de l'ICOPHONE ont eu lieu en 1976)

Notons enfin, qu'après une période d'adaptation plus longue, les scores fournis par la voix de synthèse s'améliorent sans que l'on puisse atteindre ceux correspondant à la voix humaine.

En résumé, les méthodes d'audiométrie vocale peuvent s'affiner grâce aux synthétiseurs de parole si :

- la qualité de la voix obtenue se rapproche au mieux de celle de la voix humaine (équilibre spectral, schémas intonatifs et rythmiques ...)

- les paramètres de commande sont modifiables.

Il n'est pas utopique de penser qu'elles permettraient alors d'estimer au mieux les traitements des signaux qu'une prothèse auditive "sur mesure" aurait à faire afin d'assurer son rôle non pas d'une manière parfaite (cela nous semble irréalisable) mais de la façon la plus optimale possible.

EDUCATION AUDITIVE

En nous basant sur des résultats empiriques obtenus dans le domaine de l'appareillage prothétique, nous pouvons avancer que la faculté de discriminer des sons et notamment ceux de la parole croît, pour les enfants sourds, en fonction de la durée d'adaptation. Au bout de quelques mois d'exercices systématiques on arrive à une limite de ce pouvoir qui dépend du sujet (niveau intellectuel, caractéristiques audiologiques ...) et des possibilités de son appareil (BARTH 1977). L'apprentissage est donc capital.

Dans ce domaine, un système d'enseignement assisté par ordinateur possédant une unité de réponse vocale serait fort utile. Sans amoindrir les responsabilités des pédagogues spécialisés auxquels appartiendrait la mise en place des progressions, il faciliterait le travail individuel répétitif.

DEMUTISATION ET ORTHOPHONIE

Ces activités dont le but est d'acquérir et de conserver une parole correcte doivent être menées de front avec la précédente. Le rôle humain du rééducateur est ici prépondérant. Toutefois, un système qui permettrait de répéter en dehors de sa présence des mots, des phrases et de consolider l'acquisition des schémas mélodiques et rythmiques accroîtrait l'impact de son action. L'idéal serait la combinaison d'une unité de reconnaissance susceptible de comparer

certains paramètres d'une référence avec ce qu'a émis l'élève, d'une unité de synthèse fournissant les modèles et d'une unité de visualisation où l'aspect ludique ne serait pas absent (BARTH 1975, ADDA et BARTH 1979).

CONCLUSION

En quelques pages, nous avons voulu faire part de nos besoins en indiquant les écueils auxquels se heurtaient actuellement nos expérimentations. Ces derniers sont essentiellement liés au manque de naturel de la voix de synthèse utilisée et à l'impossibilité que nous avons d'agir sur les paramètres de celle-ci autrement que par des filtrages qui sont excessivement longs à déterminer (enregistrement sur bande).

Des accords de coopération en cours de discussion nous laissent entrevoir la possibilité de poursuivre notre travail sur ce problème capital.

REMERCIEMENTS

Nous tenons à remercier particulièrement MM. G. RENARD, J.S. LIENARD, D. TEIL, et J.J. MARIANI du LIMSI pour l'aide qu'ils nous ont apportée dans la réalisation de nos manipulations .

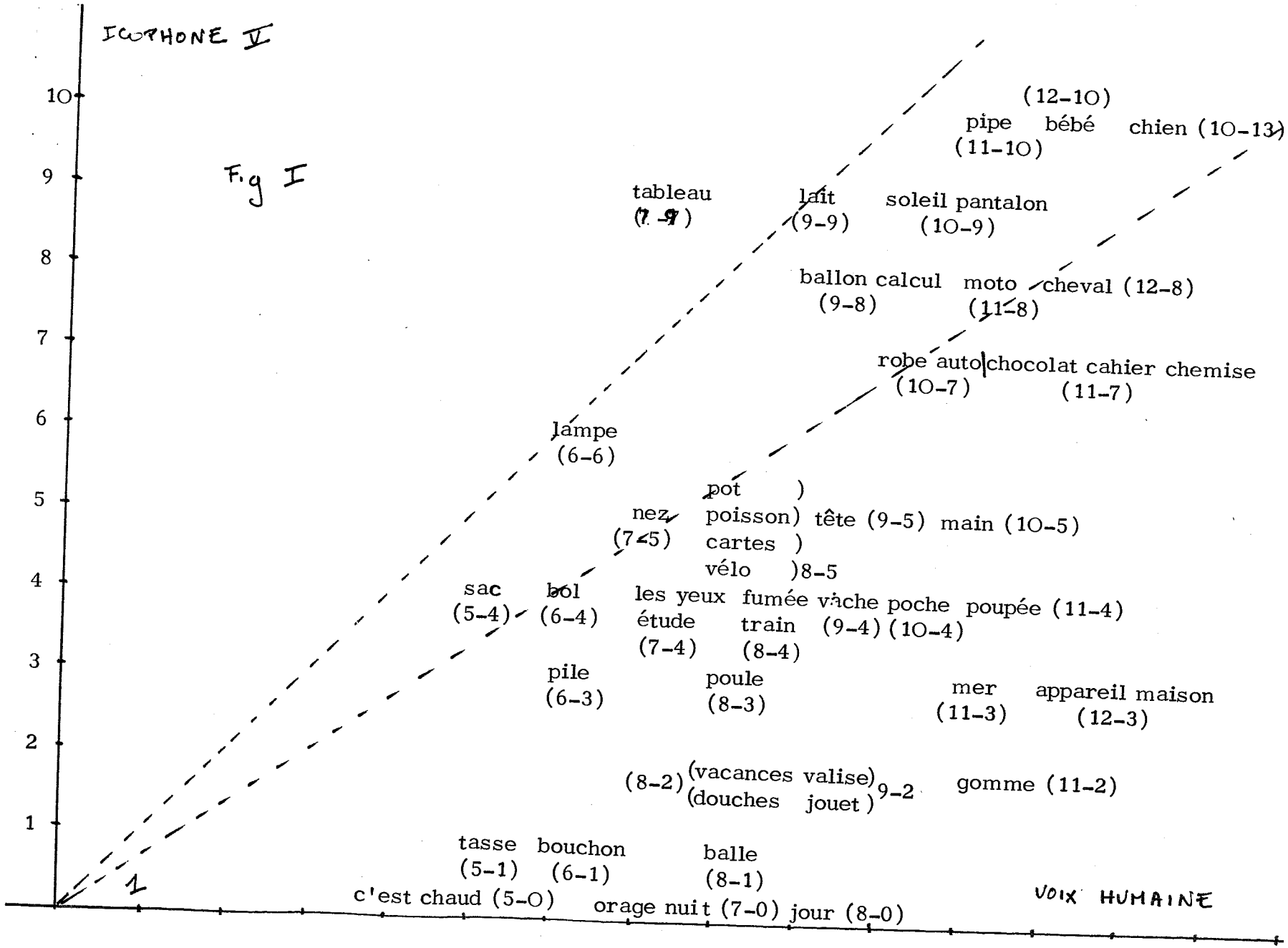

ICONPHONE IV
Fig I
10
9
8
7
6
5
4
3
2
1
(12-10)
pipe bébé chien (10-13)
(11-10)
tableau
(7-9)
lait
(9-9)
soleil pantalon
(10-9)
ballon calcul
(9-8)
moto cheval (12-8)
(11-8)
robe auto chocolat cahier chemise
(10-7)
(11-7)
lampe
(6-6)
pot)
nez poisson) tête (9-5) main (10-5)
(7-5) cartes)
vélo)8-5
sac bol les yeux fumée vache poche poupée (11-4)
(5-4) (6-4) étude train (9-4) (10-4)
(7-4) (8-4)
pile poule
(6-3) (8-3)
mer appareil maison
(11-3) (12-3)
(8-2) (vacances valise) 9-2 gomme (11-2)
(douches jouet)
tasse bouchon balle
(5-1) (6-1) (8-1)
c'est chaud (5-0) orage nuit (7-0) jour (8-0)
VOIX HUMAINE

TABLEAU II

SCORES GLOBAUX PAR SUJET

SUJET	% SCORE ICOPHONE	% SCORE VOIX HUMAINE
1	20	60
2	54	98
3	62	96
4	50	68
5	2	4
6	54	88
7	62	100
8	60	97
9	38	60
10	22	70
11	10	28
12	12	16
13	18	19
14	12	60
15	2	14

Score moyen Icophone : 31.8 %

Score moyen voix humaine : 58.5 %

BIBLIOGRAPHIE

A.J. FOURCIN et R. WRIGHT : Auditory patterning and speech production aids for the deaf - Department of phonetics and linguistics, University College London 1975

A. RISBERG : Hearing loss and auditory capacity - Research conference on speech processing aids for the deaf - Gallaudet College 1977

E. VILCHUR : Speech intelligibility in profound deafness : the effect of a severely reduced dynamic range of hearing - Fondation for hearing aid research - Woodstock 1974

J. MARTONY, A. RISBERG : Results of a rhyme test for speech audiometry International symposium on speech communication ability and profound deafness Stockholm 1970

E. DANAHER, M. OSBERGER, J. PICKETT : Discrimination of formant frequency transitions in synthetic vowels - Journal of speech and hearing research V 16 n° 3 September 1973

J.S. LIENARD : Analyse, synthèse et reconnaissance automatique de la parole - Thèse D.E. Paris 1972

D. TEIL : Conception et réalisation d'un terminal à réponse vocale. Thèse D.I. Paris 1975

S. BARTH : Application des procédés de reconnaissance automatique de la parole à l'aide aux déficients auditifs profonds - Thèse professorat - Ecole Nationale de la Santé Publique - Paris 1975

S. BARTH : La prothèse auditive en régime dynamique - Quelques aspects du traitement de l'information phonétique par les appareils de correction auditive - Rencontre européenne d' audioprothèse - Lyon 1978

G. ADDA, S. BARTH : La découverte du pouvoir de la parole par le petit enfant sourd - Colloque International Prélangage IV - Besançon 1979

S. BARTH, D. GRENIER : Etude comparée du rendement de la voix de synthèse et de la voix humaine dans le domaine de la transmission des informations phonétiques aux déficients auditifs - Journées d'Etudes des Instituts Nationaux de Jeunes Sourds - Paris 1977

10èmes JOURNÉES D'ÉTUDE SUR LA PAROLE

GRENOBLE - 30 MAI - 1ER JUIN 1979

SYNTHESE PAR REGLES DU FRANCAIS

S. CASTAN - J.Y. LATIL

Laboratoire CERFIA*- Université Paul Sabatier
118, Route de Narbonne - 31077 TOULOUSE CEDEX

RESUME

Nous décrivons dans cet article un système de synthèse vocale par règles, conçu pour un vocoder à canaux.

Les variations spectrales nécessaires à la production de la parole sont obtenues à partir d'approximations linéaires.

Ce système comporte des modules de génération prosodique permettant une production de phrases énonciatives.

Cette synthèse est actuellement commandée par une chaîne phonétique comportant des symboles de segmentation prosodique.

* CERFIA - Cybernétique des Entreprises, Reconnaissance des Formes, Intelligence Artificielle.

FRENCH SYNTHESIS BY RULES

S. CASTAN - J.Y. LATIL

SUMMARY

We describe a system of speech synthesis by rules, to be used with a channel vocoder. The phonemic generation is obtained by linear functions from a number of parameters such as : formant frequencies and amplitudes and energy in each channel of the vocoder, without any recording of natural speech.

The transition rules are obtained through the use of the time of transition, the point of the formant convergence (locus), and the generation of the transition is made by linear interpolation.

A prosodic information is given at different levels : phoneme, cluster of words and sentence.

This system implemented on a mini computer T 1600 SEMS occupies about 5 K-Words of core memories, and produces a better than worse speech in real time.

10èmes JOURNÉES D'ÉTUDE SUR LA PAROLE
GRENOBLE - 30 MAI - 1ER JUIN 1979

SYNTHESE PAR REGLES DU FRANCAIS

S. CASTAN - J.Y. LATIL

INTRODUCTION

Nous décrivons la réalisation d'un système de synthèse vocale par règles, utilisant comme organe de sortie, la partie synthétiseur d'un vocodeur à canaux.

Dans ce système, la génération de la parole en temps réel est obtenue par un ensemble de règles permettant la génération des phonèmes et de leurs transitions, sans utilisation d'élément de parole naturelle.

Une information prosodique a été ajoutée afin d'améliorer l'intelligibilité de la parole produite.

Les différentes approches possibles en synthèse de la parole sont de plusieurs types :

- Synthèse par enregistrement de messages,
- Synthèse par concaténation de mots enregistrés.

Ces deux types de méthodes fournissent en général de bons résultats (magnétophone digital), mais exigent une grande capacité mémoire et restent très limités.

- Synthèse par diphonèmes (CNET, LIMSI,...).

Cette méthode très utilisée, consiste à fabriquer la parole à partir de la combinaison des différents diphonèmes enregistrés. Nous remarquons que dans ce cas, le problème de la restitution des transitions entre phonèmes est résolu par l'enregistrement lui-même. Ce système permet une généralisation de la synthèse du français.

- Synthèse par juxtaposition de phonèmes.

Cette technique qui se rapproche de la méthode précédente pose le problème de la restitution des transitions et de l'existence de certains phonèmes isolés (Rodet, Mai 1974).

- Synthèse par règles (Rodet, 1977)

Parmi les différentes approches dans cette voie, les règles seront plus ou moins compliquées, selon que l'on utilisera ou non des éléments de parole naturelle, préalablement enregistrés (notamment pour les voyelles) et selon l'organe de sortie.

C'est ainsi que dans le système que nous proposons, nous n'utilisons au-

cun enregistrement de parole naturelle. L'organe de sortie est un vocoder réalisé au C.N.E.T.

PRINCIPE DE LA METHODE

Les parties stables et instables des phonèmes sont générées par des fonctions linéaires.

L'information prosodique est abordée au niveau du phonème, de son contexte immédiat, des groupes de mots et enfin au niveau de la phrase.

Génération des parties stables des phonèmes

La génération de ces formes est effectuée à partir d'un certain nombre de paramètres :

- une information formantique, indiquant les fréquences et l'amplitude relative des formants,
- une information de coloration du spectre (valeurs initiales des canaux) ; cette information est très utile pour la production des fricatives,
- une information de voisement indiquant simplement si le segment à générer est voisé ou non,
- des informations de durée et d'"énergie" que prendront les canaux vocoder en fin de segment généré.

Nous appelons "énergie" la valeur moyenne des canaux vocoder à un instant donné. La durée est exprimée en nombre d'échantillons.

A partir de ces informations, en appliquant une anamorphose d'amplitude sur le spectre initial, nous déterminons les échantillons vocoder aux points où l'énergie est fournie par les paramètres. La détermination des échantillons intermédiaires est obtenue par une interpolation linéaire.

Cette technique va permettre de générer toutes les voyelles et certaines consonnes ; plusieurs de ces segments mis bout à bout reconstitueront chaque voyelle. Pour les consonnes, en général, un seul segment très court (de l'ordre d'un échantillon) sera utilisé.

Pour éviter les discontinuités d'énergie qui pourraient se produire lors du passage d'un segment au suivant, nous effectuons un lissage par canal, des valeurs successives obtenues.

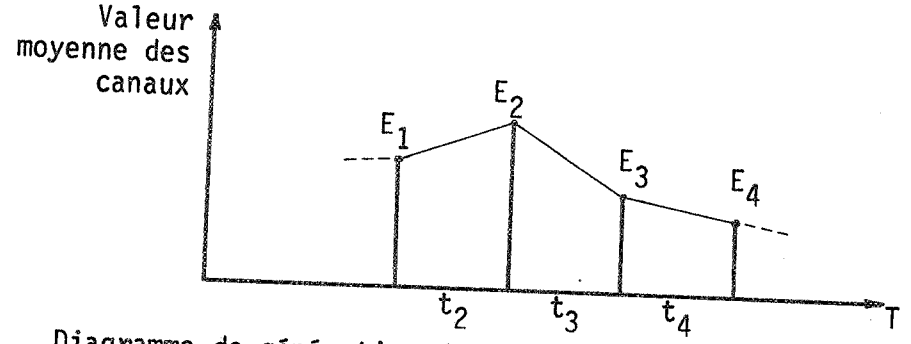


Diagramme de génération de plusieurs segments successifs

Génération des parties instables ou "transitions"

On sait, qu'au cours de la parole continue, les caractéristiques de la partie instable du spectre reliant deux phonèmes consécutifs, sont spécifiques de la combinaison de ces deux phonèmes.

Dans la langue française, on peut utiliser environ 34 phonèmes. Il est donc nécessaire de stocker 34 x 34, soit 1156 ensembles de paramètres (ou règles de transition), permettant de reproduire les caractéristiques de ces transitions.

En vue d'optimiser le système, nous avons réduit l'ensemble des informations nécessaires à la définition de chaque transition, aux paramètres qui nous ont paru les plus pertinents :

- Durée de la transition
- Valeurs des points, vers lesquels tendent chacun des formants du phonème précédant la transition. Ces points (fréquence et amplitude) sont définis par rapport aux formants du phonème qui suit la transition.

En utilisant ces informations, la transition est générée par interpolations linéaires :

- au niveau des variations formantiques en fréquence,
- au niveau des variations formantiques en amplitude,
- au niveau de l'énergie moyenne des échantillons au cours de la transition.

Cette méthode permet de calquer d'une manière imparfaite les variations spectrales de la parole naturelle.

Il est à remarquer que les variations simultanées des formants vers leur point de convergence, au cours de la transition, ne tiennent pas compte des éventuels retard et différence de vitesse, propres aux différents organes phonatoires pendant la production de la parole.

Les consonnes sont caractérisées par un spectre variable, influencé par les phonèmes adjacents.

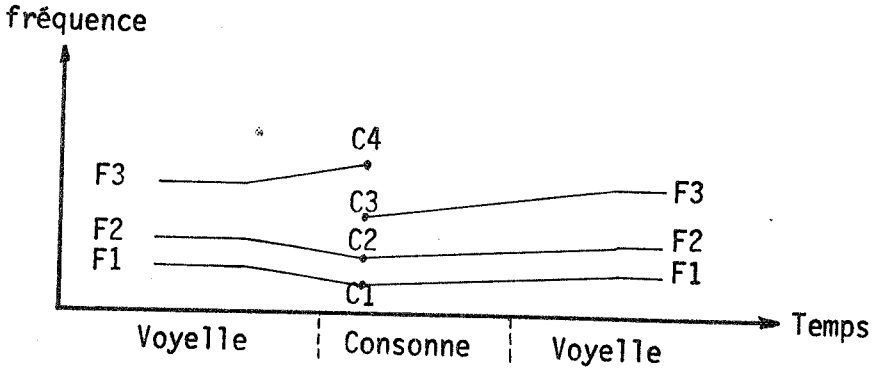
Pour générer les consonnes, nous définissons un segment stable plus ou moins bref. Ce segment va permettre de définir, pour chacune des transitions associant la consonne avec un phonème précédent, des points de convergence.

De la même manière, il permettra de définir des points de convergence, pour les transitions associant la consonne avec un phonème suivant ; ces points étant définis par rapport aux précédents.

Cette méthode permet ainsi de combler certaines lacunes de la théorie du locus (Delattre, 1966 ; Delattre, 1968), dans la mesure où les points de convergence d'une consonne peuvent être différents pour les formants du phonème précédant la consonne et ceux du phonème suivant.

Par contre, du fait de l'absence d'information sur la vitesse de variation individuelle des formants pendant les transitions, ces points de convergence sont toujours atteints ; ce qui n'est pas exact dans la parole naturelle

(Liberman, 1959 ; Coker, 1967 ; Liénard, 1977).



Exemple de génération des transitions Voyelles-Consonne-Voyelle dans le plan temps fréquence

Information prosodique

Pour améliorer la compréhensibilité de la parole obtenue, il a été nécessaire de générer une information prosodique :

- au niveau du phonème, hauteur relative de la fréquence du fondamental, niveau de l'énergie et durée relative du phonème (Dicristo A., Chafcouleff M. 1977).

- au niveau du contexte immédiat : influence de certaines consonnes voisées, sur le fondamental, l'énergie et la durée de la voyelle adjacente.

- au niveau de la phrase et des groupes de mots.

Les différents groupes de mots étant séparés par des marqueurs. Ces marqueurs sont insérés par l'utilisateur, après une analyse syntaxique et sémantique de la phrase.

Dans notre application, nous nous sommes limités à des phrases énonciatives.

Les contours prosodiques utilisés, sont tirés de schémas classiques que nous avons simplifiés (Choppy C., Liénard J.S., 1977 ; Emérard F., 1977 ; Caelen J., Maurang G., 1976).

Traitement du rythme

Au cours de la phrase, nous faisons décroître légèrement le rythme pour le diminuer brusquement sur la dernière voyelle.

Au niveau des groupes de mots, nous augmentons la durée de la dernière voyelle et nous marquons une pause à la fin de chaque groupe.

Traitement de la mélodie

Au cours de la phrase, la fréquence du fondamental croît pour atteindre un maximum à la fin du premier groupe de mots puis décroît jusqu'en fin du dernier groupe.

Au niveau des groupes de mots non finaux, la fréquence du fondamental atteint un maximum sur la dernière syllabe de chaque groupe.

Pour le groupe final, nous avons fixé le maximum sur la première syllabe de celui-ci.

Traitement de l'énergie

La caractéristique essentielle de la courbe utilisée est une chute de 4 db environ sur la dernière syllabe de la phrase.

Ce traitement prosodique relativement simple permet d'améliorer la compréhensibilité de la parole synthétique.

CONCLUSION

Cette méthode a permis de réaliser un système nécessitant une occupation mémoire relativement peu importante.

Les faibles dimensions de la table de paramètres, nécessaires à la génération des parties stables des phonèmes (800 mots environ), ainsi que la matrice renfermant les caractéristiques des transitions (1 k mots environ), leur permettent de résider en mémoire centrale.

L'ensemble de tous les modules utiles à la génération de notre synthèse, occupent environ 5 k mots mémoire.

Le peu de traitement à effectuer et la faible densité d'information à générer, permettent au système de travailler en temps réel.

Malgré les diverses approximations importantes utilisées, la voix synthétique obtenue est encore compréhensible par un auditeur relativement peu entraîné.

Cette synthèse peut parfaitement être utilisée dans le cadre d'un ensemble de dialogue homme-machine.

BIBLIOGRAPHIE

CAELEN, J., MAURANG, G., 1976, Fréquence intensité et durée : étude comparative des fonctions dans les phrases énonciatives simples et étendues. 7ème journées du Groupe "Communication parlée" Nancy.

CARRE, R., 1971, Contribution aux études sur l'analyse de la parole, rôle et importance des formants. Thèse d'état, Université de Grenoble.

C.N.E.T. Lannion, 1975,1976, Recherche/acoustique.

CHOPPY, C., LIENARD, J.S., 1977, Prosodie automatique pour la synthèse par diphonèmes. 8ème journées du Groupe "Communication parlée", Aix-en Provence.

COKER, C.H., 1967, Synthesis by rule from articulatory parameters. IEEE Conf. on speech communication and processing. Cambridge, Mass.

DELATTRE, P., 1966, Studies in French and comparative phonetics, London, the Hague, Paris : Mouton and Co.

DELATTRE, P., 1968, From acoustic aves to distinctive features, Phonetica 18.

DICRISTO, A., CHAFCOULOFF, M., 1977, Les faits micro-prosodiques du français : voyelles, consonnes, coarticulations. 8ème journées du groupe "Communication parlée" Aix-en-Provence.

DOURS, D., FACCA, R., PERENNOU, G., 1974, Analyse temporelle du signal de parole comparée à l'analyse fréquentielle du point de vue de la reconnaissance. 5ème journées du groupe "Communication parlée". Orsay.

EMERARD, F., 1977, Synthèse par diphones et traitement de la prosodie. Thèse de docteur de troisième cycle, Grenoble.

LIBERMAN, A.M., INGEMANN, F., LISKER, L., DELATTRE, P., COOPER, F.S. 1959, Minimal rules for synthesizing speech, JASA 31.

LIENARD, J.S., 1972, Analyse synthèse et reconnaissance automatique de la parole. Thèse d'état, Paris VI.

LIENARD, J.S., 1977, Les processus de la communication parlée. Editions Masson.

RODET, M.X., 1974, Synthèse de la parole. 5ème journées du groupe "Communication parlée. Orsay.

RODET, M.X., 1977, Analyse du signal vocal dans sa représentation amplitude temps, synthèse de la parole par règles. Thèse d'état, Paris VI.

ANNEXE

Caractéristiques de la partie synthèse du vocoder :

- nombre de canaux : 14
- nombre de valeurs de gain de chaque canal : 16
- pas d'incrémentation de ces valeurs : 3 db

Les fréquences des canaux sont indiquées dans le tableau suivant :

N° du canal	Fréquence centrale (Hz)	Largeur de chaque canal (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

Ce synthétiseur est connecté à un mini-calculateur T1600 SEMS, sur lequel est implanté notre système de synthèse.

10èmes JOURNÉES D'ÉTUDE SUR LA PAROLE

GRENOBLE - 30 MAI - 1ER JUIN 1979

ETUDE DES DUREES SPECIFIQUES DES VOYELLES ACCENTUEES DE MANIERE EMPHATIQUE AU MOYEN DE LA SYNTHESE.

D. DUEZ - R. CARRE -

D. DUEZ - Institut de Phonétique - AIX-EN-PROVENCE -

R. CARRE - Laboratoire de la Communication Parlée - GRENOBLE - E. R. A. au C.N.R.S. n° 366

RESUME

On a déjà montré que la durée des voyelles joue un rôle prépondérant dans la perception de l'emphase. Dans le présent rapport, nous étudions les seuils emphase/non emphase en fonction de la durée de la voyelle accentuée, selon la nature de cette voyelle (|i|, |a|, |ã|) et selon son contexte consonantique (constrictive voisée, occlusive non voisée). La durée de la voyelle a été modifiée en utilisant les possibilités offertes par un ensemble d'analyse synthèse à codage prédictif.

Les résultats obtenus montrent l'influence de la nature de la voyelle et l'influence de la consonne subséquente. Un allongement de 50 à 100 ms de la durée de la voyelle accentuée suffit à l'obtention de l'emphase.

STUDY OF SPECIFIC DURATION OF EMPHATIC VOWELS BY MEANS OF SYNTHESIS

D. DUEZ - Institut de Phonétique - AIX-EN-PROVENCE -
R. CARRE - Laboratoire de la Communication Parlée - GRENOBLE -
E.R.A. au C.N.R.S. n° 366

SUMMARY

In this paper, the perception of an emphatic level is studied according to the duration of accentuated vowels with different vowels ($|\dot{a}|$, $|\alpha|$, $|\tilde{\alpha}|$) and different adjacent consonants (voiced constrictive, non-voiced stops). Modifications in the duration of accentuated vowels were performed on a linear prediction vocoder.

The results show the influence of the nature of the vowel and the influence of the subsequent consonant. A lengthening of 50-100 ms of the accentuated vowel duration is enough to get emphatic level.

10èmes JOURNÉES D'ÉTUDE SUR LA PAROLE
GRENOBLE - 30 MAI - 1ER JUIN 1979

ETUDE DES DUREES SPECIFIQUES DES VOYELLES ACCENTUEES DE MANIERE EMPHATIQUE AU MOYEN DE LA SYNTHESE.

D. DUEZ - Institut de Phonétique - AIX-EN-PROVENCE -
R. CARRE - Laboratoire de la Communication Parlée - GRENOBLE -
E.R.A. au C.N.R.S. n° 366

On sait que la réalisation physique de l'accent final en français se manifeste par des modifications des paramètres acoustiques de la voyelle accentuée (durée, fréquence fondamentale, intensité). Un allongement particulièrement important de la durée de cette voyelle conduit à une impression d'emphase* (DUEZ D., 1978), le rôle de l'intensité et de la fréquence fondamentale étant alors limité.

On sait, par ailleurs, que la durée intrinsèque de la voyelle varie selon sa nature et selon son environnement consonantique (ROSSI et DI CRISTO, 1977, DI CRISTO, 1978).

Dans cette étude, nous avons voulu vérifier si la perception d'un niveau emphatique due à l'allongement des voyelles accentuées est influencée par la nature de ces voyelles et par leur environnement consonantique. Pour cela, nous avons enregistré plusieurs phrases prononcées par un même locuteur, comprenant des voyelles accentuées de nature différente et soumises à des environnements consonantiques différents. Ensuite, des modifications ont été opérées sur la durée de ces voyelles en utilisant les possibilités offertes par un vocoder à codage prédictif. Enfin, les échantillons obtenus ont fait l'objet de tests d'évaluation permettant de déterminer un niveau emphatique en fonction de différents allongements de ces voyelles.

* Nous entendons par emphase : mode d'expression comportant une certaine affectation.

I/ METHODE

1. Choix des phrases du corpus

Afin d'établir un corpus homogène, nous avons tenu compte de certaines influences :

- l'influence du débit, qui est souvent liée à la longueur de la phrase. *"The speed of the phrases depends on their length"* écrit FONAGY (1958).

- l'influence du nombre de syllabes par groupe rythmique et par phrases. Plus le groupe rythmique contient de syllabes, plus la durée de ses syllabes est brève (LINDBLOM, 1976)

- l'influence des pauses. La pause allonge la syllabe qui la précède (PIKE, 1972).

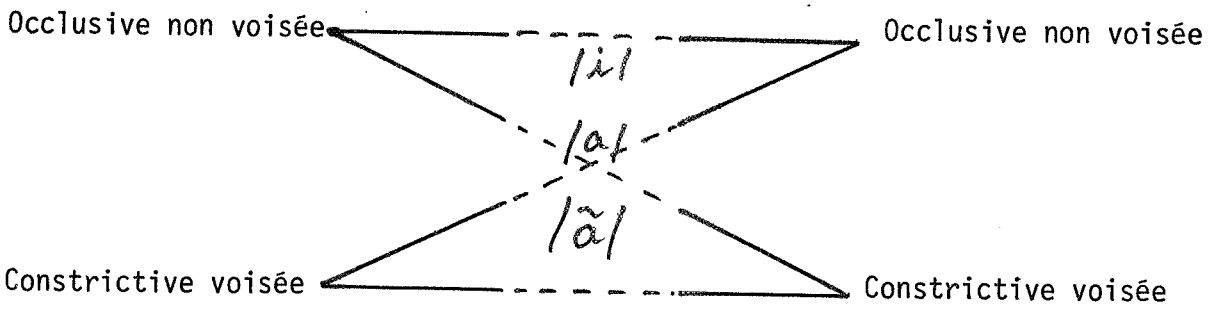
Nous avons donc choisi des phrases contenant un même nombre de syllabes (6), et composées de deux groupes rythmiques de 3 syllabes chacun, le 1er groupe rythmique étant suivi d'une pause.

La nature de la voyelle et l'environnement consonantique influencent la durée de la voyelle accentuée. Des résultats précis obtenus par DI CRISTO (1978) pour les voyelles du français, permettent de mesurer ces influences :

- Les voyelles basses sont 25 % plus longues que les voyelles hautes ; les voyelles nasales sont 73 % plus longues que les voyelles hautes. Les consonnes voisées précédant la voyelle accentuée l'allongent de 25 %.
L'effet de voisement des consonnes subséquentes est hautement significatif, puisque la voyelle accentuée subit un allongement de 100 % lorsque la consonne postvocalique est une constrictive, de 50 % lorsque cette consonne est une occlusive.

Afin de mettre en évidence ces influences, nous avons choisi comme voyelles accentuées, les voyelles [i], [a] et [ã], et comme consonnes subséquentes et précédentes, tantôt une occlusive non voisée, tantôt une constrictive voisée.

Nous avons établi les associations suivantes (Fig. 1).



- Figure 1 -

Douze phrases établies en fonction de ces critères ont été obtenues. Elles ont été lues par un même locuteur. La durée de la voyelle accentuée du premier groupe rythmique a été mesurée et est reportée ci-dessous, avec les phrases correspondantes :

Chez Edwige on mange bien	(260 ms)
Ce litige est mineur	(280 ms)
Cette otite me fait mal	(120 ms)
Partons vite dit sa mère	(140 ms)
Ces rivages me fascinent	(200 ms)
L'hermitage est très calme	(300 ms)
Ces patates sont très chères	(150 ms)
Ta cravate te va bien	(140 ms)
Il se venge dit son frère	(310 ms)
A Knutange on s'amuse	(370 ms)
Mon attente me pèse bien	(270 ms
La régente est bien morte.	(260 ms)

2. Expérimentation

L'expérimentation a été effectuée à l'aide d'un système d'analyse synthèse par codage prédictif. Ce système permet, en particulier, de modifier les différents paramètres suivants : durée, fréquence fondamentale et intensité. Les caractéristiques du vocoder sont les suivantes :

- l'analyse est effectuée à partir des coefficients d'autocorrélation. La fenêtre d'analyse est de 20 ms.
- 12 coefficients de prédiction sont calculés
- la fréquence fondamentale est détectée par filtrage inverse
- la synthèse est obtenue avec un filtre en échelle
- les 12 coefficients de prédiction sont transmis sur 16 bits tous les 25 ms.

Après analyse, les paramètres (fréquence fondamentale et amplitude) sont présentés sur un écran de visualisation. Chacun des paramètres peut être modifié à l'aide d'une table d'entrée graphique. Pour modifier la durée, le programme permet de retrancher ou d'ajouter des séquences de 25 ms.

3. Modifications

Tout d'abord, nous avons ramené les variations de fréquence fondamentale existant entre la voyelle accentuée étudiée et la voyelle inaccentuée précédente à un écart de 3 demi-tons (montée tonale), en tenant compte de la fréquence spécifique des voyelles, cet écart correspondant à une accentuation normale. On a opéré de la même façon pour les variations d'intensité. Les modifications apportées à la durée de la voyelle ont été faites en trois temps. D'abord, chacune des voyelles accentuées a été diminuée de 50 ms, le locuteur pouvant prononcer inconsciemment les échantillons avec une certaine emphase. La durée initiale de chacune des voyelles a ensuite été augmentée d'une séquence de 50 ms, puis d'une séquence de 100 ms.

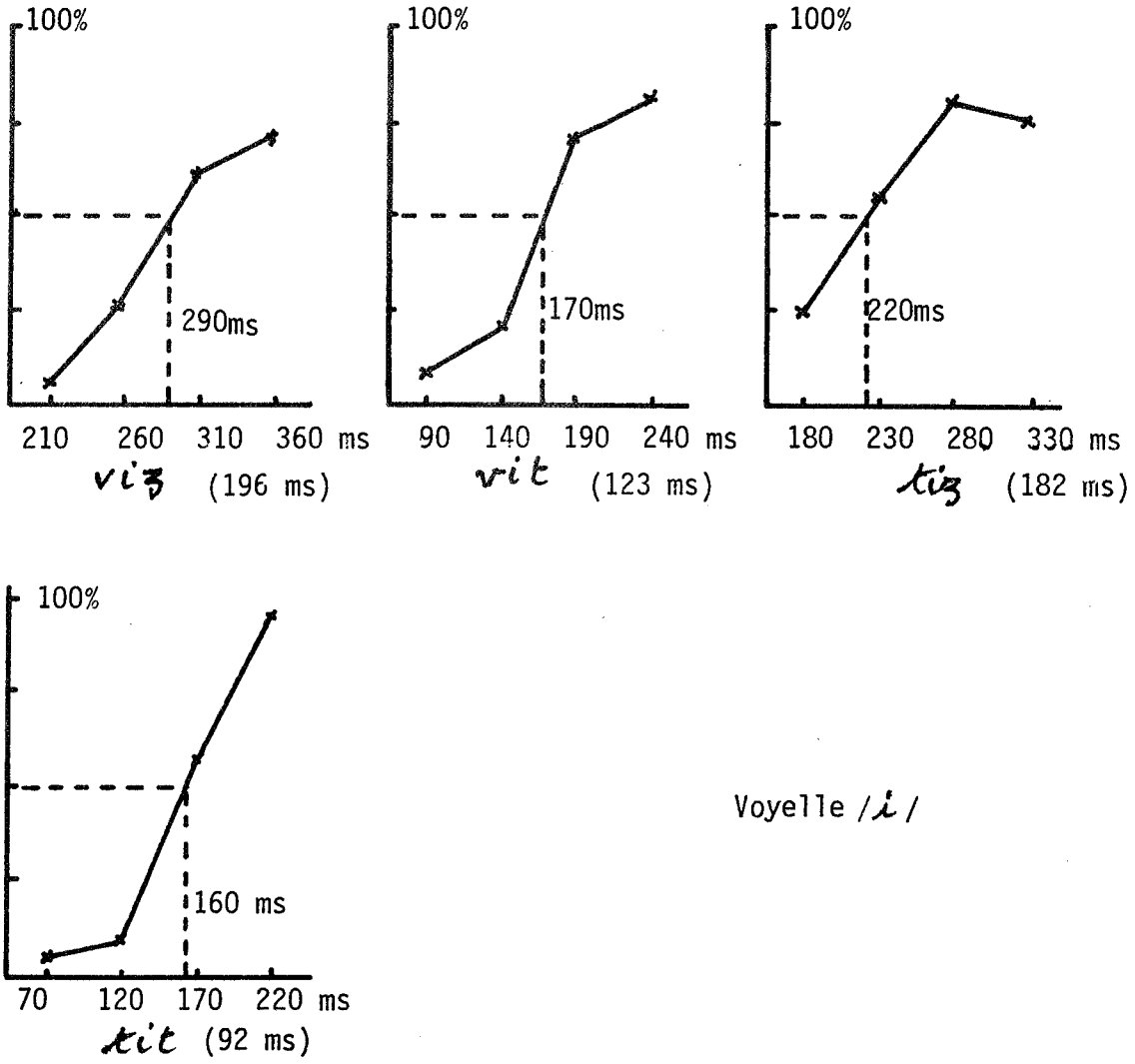


Figure IIa : Pourcentages d'emphase obtenus en fonction de la durée de la voyelle.
On a indiqué entre parenthèses les valeurs obtenues par DI CRISTO (1978) pour les voyelles accentuées correspondantes.

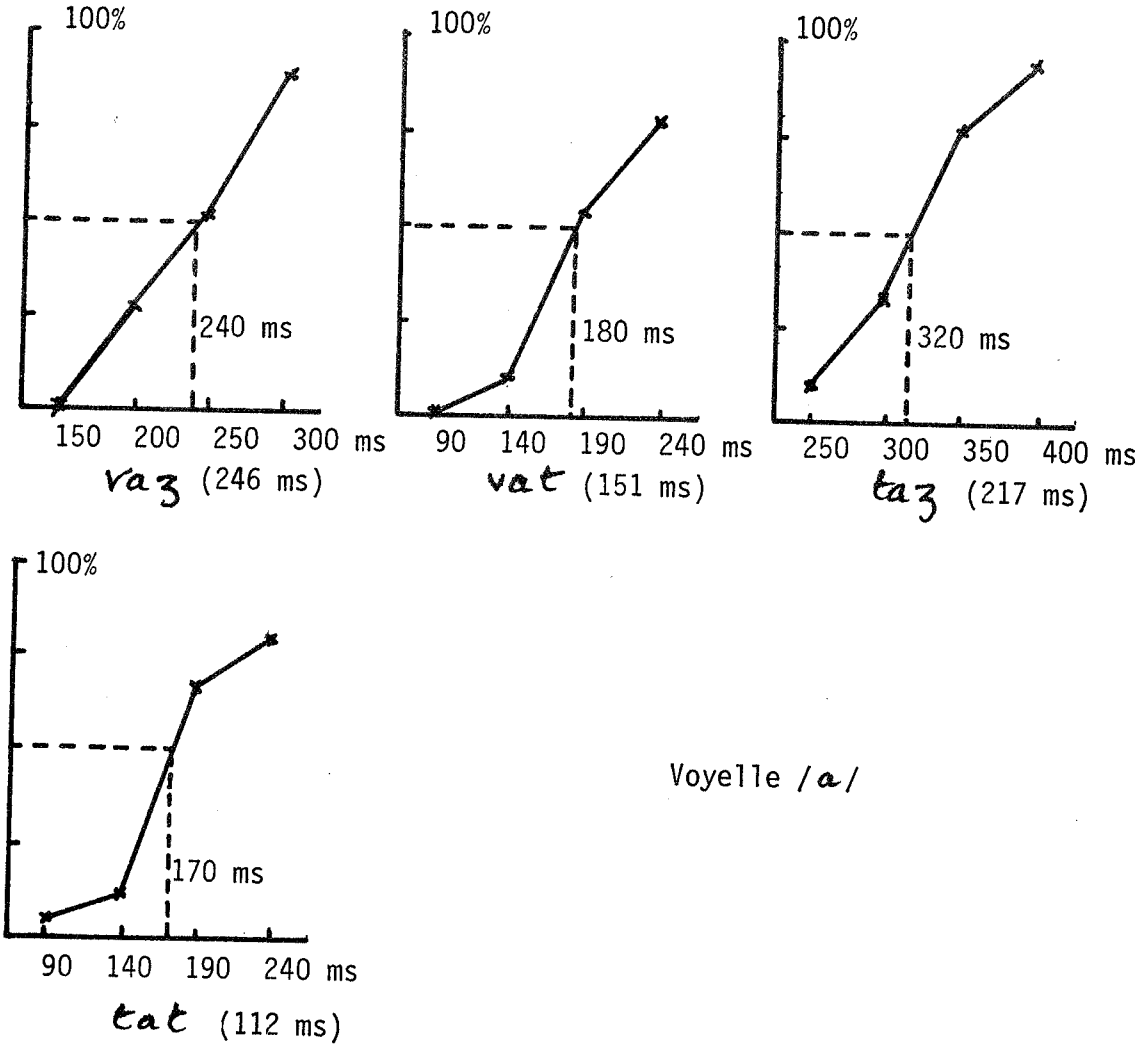


Figure IIb : Pourcentages d'emphase obtenus en fonction de la durée de la voyelle.
On a indiqué entre parenthèses les valeurs obtenues par DI CRISTO (1978) pour les voyelles accentuées correspondantes.

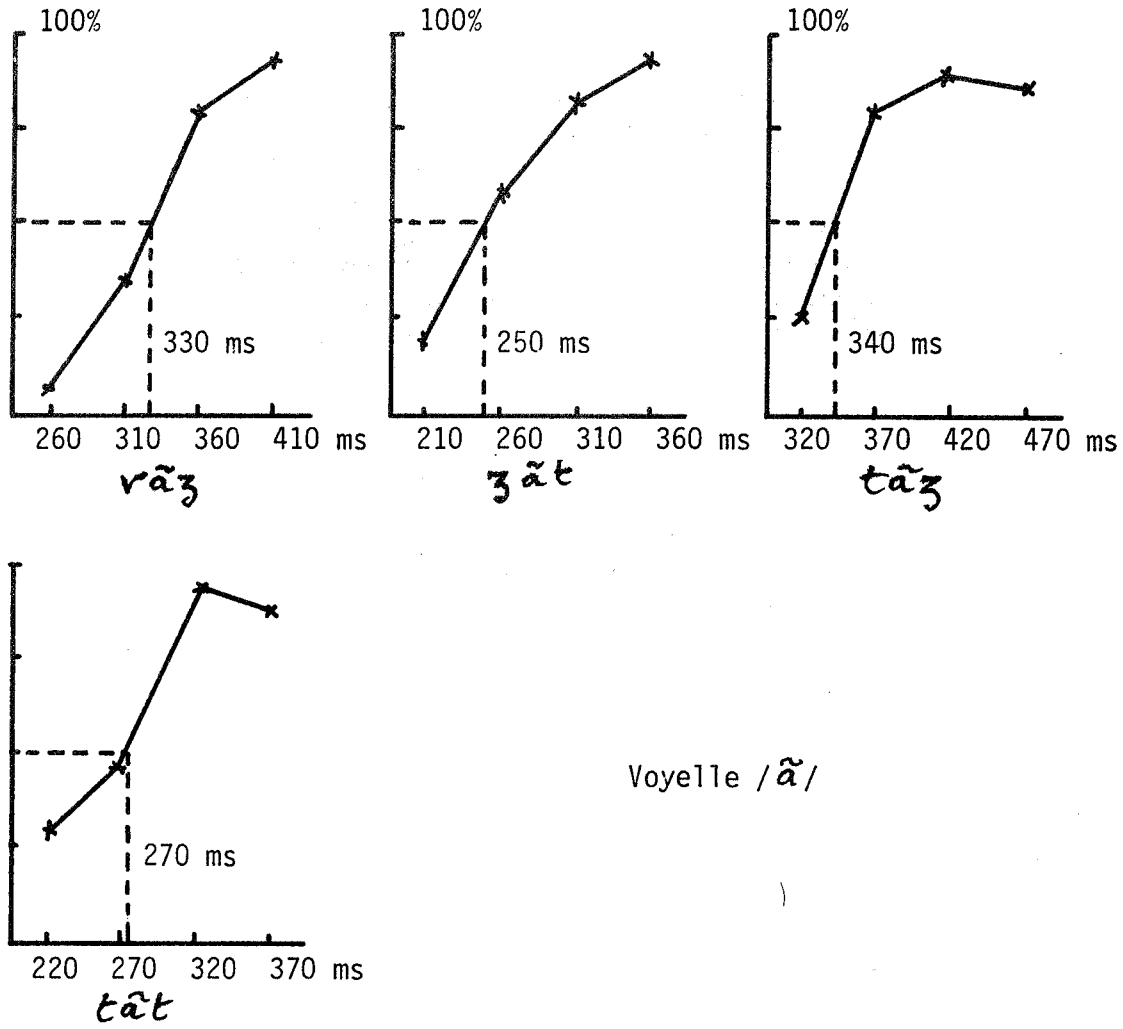


Figure IIc : Pourcentages d'emphase obtenus en fonction de la durée de la voyelle.

Ont ainsi été obtenus 48 échantillons -nous incluons les échantillons initiaux -.

4. Tests de perception.

Les 48 échantillons ont ensuite été enregistrés 10 fois, selon des ordres aléatoires différents.
L'ensemble a été présenté à 10 auditeurs (tous enseignants). Il leur a été demandé de préciser si l'échantillon était emphatique ou non.

II/ Résultats

L'ensemble des résultats est présenté à la figure II

III/ Discussion

Les différentes courbes montrent que le passage non emphase - emphase, se fait, quel que soit le contexte ou la nature de la voyelle, en 100 millisecondes environ.

Par ailleurs, on retrouve les tendances relevées par DI CRISTO (1978) en fonction de la consonne subséquente. Le seuil emphase - non emphase (à 50%) correspond à des voyelles de durée plus longue lorsque la subséquente est voisée.Les différences sont peu appréciables avec des consonnes précédentes différentes (voisée/non voisée).

Les écarts existant entre la voyelle haute et la voyelle basse sont peu marqués. Mais on retrouve un net allongement pour la voyelle nasale.

Si l'on cherche, comme DI CRISTO, à établir des pourcentages, on constate que le seuil obtenu pour la voyelle basse est 10% plus élevé que celui de la voyelle haute et que le seuil pour la voyelle nasale est 30% plus élevé que celui de la voyelle haute.

Lorsque la consonne subséquente est voisée, le seuil est 50% plus élevé que lorsque cette consonne est non voisée.

Ces pourcentages, par ailleurs moins élevés, ne correspondent pas bien à ceux donnés par DI CRISTO. L'emphase ne serait donc pas dû à un allongement de la voyelle proportionnel à la durée de cette voyelle dans un contexte déterminé, mais plutôt à un allongement, en valeur absolue, de la durée moyenne de la voyelle accentuée.

Cette hypothèse est confirmée par la première remarque que nous avons émise sur l'allure des courbes.

Une augmentation de 50 à 100 ms par rapport aux durées relevées par DI CRISTO pour les voyelles accentuées apporterait un caractère emphatique. Rappelons que pour des durées supérieures à 300 ms, ce qui est le cas de beaucoup de nos échantillons, notre faculté d'apprécier les durées est beaucoup moins fine que pour des valeurs inférieures (ROSSI 1971).

Enfin, par rapport aux résultats obtenus par DI CRISTO, certaines valeurs paraissent difficiles à interpréter (par exemple /vaʒ/)

Il nous parait nécessaire de réétudier l'ensemble de nos résultats en fonction de la durée de la pause entre les deux groupes rythmiques.

CONCLUSION

Les premiers résultats que nous avons obtenus montrent que la perception de l'emphase est liée à un allongement, en valeur absolue, de la durée moyenne de la voyelle accentuée.

Nos tests ont été éffectués à partir de phrases enregistrées par un même locuteur. Il ne semble pas nécessaire, pour étendre notre expérience, de changer de locuteur.

En effet, les auditeurs ont, plus sûrement, un rôle dans cette expérience. En revanche, l'étude de nouvelles phrases en tenant des pauses entre groupes rythmiques devrait être plus instructive.

BIBLIOGRAPHIE

DI CRISTO A.,De la microprosodie à l'intonosyntaxe. Thèse d'Etat. Aix en Provence, (1978)

DUEZ D., Essais sur la prosodie du discours politique. Thèse de 3è cycle. Paris III, (1978)

FONAGY I., Speed of utterance in phrases of different lengths. Language and Speech, I, 126-152, (1958)

LINDBLOM B et al., Durational patterns of Swedish Phonology. DO they reflect short term memory processes. Departement of Phonetics, Stokholm, (1976)

PIKE, General characteristics of intonation. Penguin Book, (1972)

ROSSI M et DI CRISTO A., Propositions pour un modèle d'analyse de l'intonation. VIIIè Journées d'Etudes de la Parole. Aix en Provence, 324-328, (1977)

ROSSI M., Le seuil différentiel de durée. Papers in Memory of Pierre Delattre, Mouton, La Hague, 435-450, (1971)

10èmes JOURNÉES D'ÉTUDE SUR LA PAROLE

GRENOBLE - 30 MAI - 1ER JUIN 1979

ESSAI D'EVALUATION DE L'INTELLIGIBILITE DES LIGNES D'ANNUAIRE EN PAROLE DE SYNTHESE

Françoise EMERARD et Patrick GRAILLOT

CNET - LANNION

RESUME

Dans le cadre d'un projet d'automatisation des Centres de renseignements, la possibilité d'utiliser la synthèse vocale est envisagée.

C'est dans cette perspective que se situe une expérience destinée à évaluer l'intelligibilité en synthèse par diphones des diverses informations contenues dans les lignes d'annuaire.

Une procédure a été élaborée pour tenter de simuler assez précisément une demande de renseignements.

Les résultats montrent que si le pourcentage de confusion sur les numéros de téléphone est pratiquement nul, l'intelligibilité de noms patronymiques n'excède 68% qu'à condition que soit réalisée une épellation. Les autres informations donnent des scores de reconnaissance compris entre 82% (adresses) et 96% (professions).

Intelligibility evaluation of synthesized speech for a telephone directory - assistance purpose.

Françoise EMERARD et Patrick GRAILLOT

CNET LANNION.

SUMMARY

This work is related to a program aiming at the automatisation of Telephone Information Centers : synthetic speech is considered for transmission of the information to be supplied.

At this stage, in order to test the intelligibility of speech synthesis by dyads, an experiment was carried with samples of various informations encountered in a traditional directory.

Procedures were established such as
- creating a normal psychological environment for the listeners,
- closely simulating real inquiries' conditions,
- obtaining, besides a normal intelligibility test, personal reaction to synthetic speech. These qualitative results are reported elseway (EMERARD & GRAILLOT, 1979).

Listeners were asked to dial on their domestic phone a number corresponding to the experimental automatic answering service. While listening to the recorded speech synthesized message, they were asked to write down on forms supplied what they understood.

101 listeners took part to the test. The material recorded includes 9 different messages : each message corresponds to 5 lines of a traditional directory. We introduced various spelling methods of names besides changes in content forms from one message to the other.

Names are correctly identified in 68% of cases when spelled letter by letter (a score of 91% is obtained for natural speech). When spelling is enhanced with cues (A as in Anatole...), the result jump to 94%.

For the remaining informations - first name, profession and address - we proceeded without any spelling : recognition scores were respectively 95%, 96% and 82%. Phone numbers stated by pair (44-56-92) as it is customary in France showed to be intelligible in more than 98% of the cases.

Those scores show that the intelligibility of the samples we synthesized is quite acceptable. However, the qualitative results point out that much is to be done in order of improving the naturelness of synthetic speech.

10èmes JOURNÉES D'ÉTUDE SUR LA PAROLE
GRENOBLE - 30 MAI - 1ER JUIN 1979

ESSAI D'EVALUATION DE L'INTELLIGIBILITE DES LIGNES D'ANNUAIRE EN PAROLE DE SYNTHESE

Françoise EMERARD et Patrick GRAILLOT

CNET - LANNION

INTRODUCTION

Cette expérience a été élaborée pour évaluer la qualité de la parole synthétique transmise par téléphone. Elle se situe dans le cadre d'un projet d'automatisation des Centres de Renseignements des Télécommunications développé au CNET. Les réponses que devra fournir un tel système sont essentiellement le numéro d'un abonné, sa profession et son adresse. Le matériel sonore utilisé a tout naturellement été constitué sur le modèle des lignes d'annuaire téléphonique.

La procédure de synthèse a été décrite par ailleurs (EMERARD, F., 1977), elle utilise comme éléments de parole, des diphones et comme analyseur-synthétiseur un vocodeur à canaux couplé à un ordinateur. Le stockage numérique d'environ 1200 diphones donne la possibilité de composer n'importe quel message de la langue française. Pour cette étude, le débit d'information est de 4800 bits/s.

Trois considérations ont orienté l'organisation de ce test :

*Créer pour les sujets des conditions psychologiques favorables :

- Nous n'avons pas fixé la liste des auditeurs. Chaque personne a décidé seule de participer ou non au test après avoir été avertie de la procédure envisagée.

- Les sujets n'ont pas eu à se déplacer pour répondre au test ; seule nécessité : se trouver dans un local disposant d'un téléphone.

- Enfin, le moment pour répondre au test n'est pas fixé par l'expérimentateur : les sujets ont la liberté de choix, dans les limites d'une journée. Nous verrons cependant que des contraintes matérielles réduisent cette liberté.

*Simuler au mieux les conditions réelles : quand un abonné désirera un renseignement, il décrochera son combiné, appelera le centre de renseignements, posera sa question et recevra la réponse en voix synthétique, par voie téléphonique sans forcément que soient réunies à ce moment-là les conditions optimales de réception téléphonique. C'est ce que nous avons voulu simuler. Pour cette raison, il n'était pas question que les sujets entendent de la parole synthétique dans une ambiance bien "aseptisée" (salle d'audiométrie ou chambre sourde).

.../...

Il nous a paru préférable qu'ils restent dans leur bureau avec l'environnement acoustique habituel même si cela entraîne en pratique l'impossibilité de contrôler le bruit ambiant au moment du test.

* Obtenir de la part des auditeurs non seulement une réponse précise au test d'intelligibilité mais aussi la formulation de leurs impressions sur la parole synthétique (critiques, suggestions...). Ces résultats sont présentés par ailleurs (EMERARD et GRAILLOT, 1979).

I - LE MATERIEL VOCAL, LES SUJETS, LA PROCEDURE DE TEST

1 - Le matériel vocal

Il est constitué de lignes d'annuaire téléphonique. Chacune d'entre elles comporte le nom patronymique (ou la raison sociale), le prénom, la profession, l'adresse (N°, nom de rue, localité), et le numéro de téléphone précédé de l'indicatif départemental.

Neuf listes de cinq lignes d'annuaire différentes ont été enregistrées :

Ces 9 listes ont été synthétisées à partir de diphones. Elles diffèrent par le mode d'épellation. En effet, pour certaines listes, le nom patronymique est épelé simplement, par exemple "DUPONT, je répète D,U,P,O,N,T" ; d'autres sont épelées avec le support d'une référence : "Dupont, je répète D comme Désiré, U comme Ursule etc...", une liste enfin a été épelée avec une référence partielle, uniquement pour certains sons connus pour leur fort taux de confusion : "Dupont, je répète D, U, P comme Pierre, O, N comme Noémie, T".

A des fins de comparaison, deux de ces listes ont été reprises :

- L'une a été enregistrée dans une pièce silencieuse par une locutrice équipée d'un combiné de téléphoniste actuellement utilisé dans les centres de renseignement et a servi de référence.

- L'autre a été obtenue à l'aide d'un vocodeur en analyse-synthèse directe (14 canaux couvrant la bande 300 - 3400 Hz). Pour ces deux listes, épellation simple.

Nous avons pensé que l'insertion de phrases prononcées en parole de synthèse et précédant chacune des lignes pouvait aider l'auditeur à mieux repérer les éléments du message. C'est pourquoi chaque ligne d'annuaire "synthétisée" a été présentée de la façon suivante :

"Les informations que vous avez demandées sont les suivantes : votre correspondant DUPONT, Jean, je répète D, U, P, O, N, T, profession : charcutier, domicilié 7, rue des Genêts Rouges, à Beaulieu, a pour indicatif le 32 suivi de 47, 29, 33".

Nous avons utilisé un système prosodique simple ne faisant appel qu'à un nombre limité de marqueurs. Les contours appliqués ici sont ceux qui, après différents essais, se sont révélés les plus agréables pour les auditeurs.

.../...

2 - Les auditeurs

Au total 101 auditeurs, tous agents du CNET, ont participé au test (23 femmes et 78 hommes). Parmi ceux-ci, 22% avaient déjà entendu de la parole de synthèse.

3 - Organisation pratique du test

* La tâche des auditeurs était la suivante :

Composer sur un cadran téléphonique un numéro donné, cet appel déclenchait le déroulement d'une bande enregistrée sur magnétophone (système répondeur).

Ecouter les cinq lignes d'annuaire et noter au fur et à mesure les résultats sur la feuille prévue à cet effet.

Répondre à un certain nombre de questions posées au verso de cette feuille.

* Le matériel utilisé : le système répondeur est composé d'une ligne téléphonique raccordée à un magnétophone télécommandable ; il est actionné par un organe de commande géré par un microprocesseur.

II - ANALYSE DES RESULTATS

242 feuilles de réponses nous sont parvenues (la plupart des auditeurs ont écouté plusieurs listes). Les statistiques que nous avons dressées ne portent que sur 1142 lignes d'annuaire car nous en avons éliminé une soixantaine qui étaient incomplètes.

Le tableau I donne les résultats pour les 5 présentations différentes des listes :

- Parole naturelle avec épellation simple,
- Vocodeur en analyse-synthèse avec épellation simple,
- Synthèse avec épellation des noms patronymiques de manière { simple / partielle / complète

Pour ce qui concerne les informations autres que l'épellation du nom de l'abonné, nous avons regroupé les scores obtenus en synthèse.

.../...

TYPES DE PRÉSENTATION	NB de LIGNES ÉCOUTÉES	NOM AVANT ÉPELLATION				PRÉNOMS	NOMS ÉPELÉS				PROFESSIONS	ADRESSES			COMMUNES			N°S ∅
		NB FAUTES					NB FAUTES					NB FAUTES			NB FAUTES			
		0	1	2	>2		0	1	2	>2		0	1	>1	0	1	>1	
EPELÉ SIMPLE	627	48%	21%	10%	21%	94 %	68%	19%	8%	5%	95 %	81%	9%	10%	80%	10%	10%	99,1%
VOCODEUR A/S	60	47%	18%	18%	17%	98%	65%	23%	10%	2%	98 %	96%	2%	2%	85%	5%	10%	98,5%
EPELÉ COMPLET	240	78%	8%	3%	11%	95 %	94%	6%	0 %	0%	99 %	88 %	7%	5%	86%	7%	7 %	100 %
EPELÉ PARTIEL	160	57%	13%	10%	20%	97%	79%	17%	3%	1%	95 %	75%	13%	12%	83%	8%	9%	93 %
NATUREL	55	71%	24%	3%	2%	100 %	91 %	7%	2%	—	100 %	99%	1%	—	93%	5%	2%	100 %
	1142																	
Σ SYNTHÈSE	1027	56,4	16,7	8,4	18,5	94,7					95,9	81,7	9,2	9,1	81,9	9	9,1	98,4

TABLEAU I : RESULTATS NUMERIQUES

1 - Analyse des confusions des sons lors de l'épellation des noms patronymiques :

Nous avons principalement regardé ici le cas des noms épelés de manière simple. Le tableau II présente la répartition des confusions en tenant compte de toutes les occurences possibles : la voyelle [ɛ] par exemple intervient dans l'émission de F, Z, S... ([ɛ f], [ɛ ɛ d], [ɛ s]...).

Globalement, le score de reconnaissance est de 95,2 %.

1 a - Les confusions sur les voyelles :

4265 voyelles ont été émises, elles ont donné lieu à 56 confusions. Le score de reconnaissance pour les voyelles est de 98,7 %. Ce sont les voyelles i et o qui provoquent le plus grand nombre d'erreurs, mais surtout dans les 43 confusions relevées lors de l'émission des voyelles [i, e, a, y, o,] 79% de celles-ci sont faites avec la seule voyelle [o] qui phonétiquement est une voyelle plutôt neutre, affaiblie et inaccentuée. Cette tendance laisse supposer qu'en synthèse et pour effectuer l'épellation, il faudra prévoir un traitement particulier qui réalise l'accentuation. Telles qu'elles existent actuellement, les voyelles de synthèse sont acceptables pour une utilisation dans la parole continue mais n'ont pas suffisamment de précision articulatoire pour être employées de façon isolée.

1 b - Les confusions sur les consonnes :

2711 consonnes ont été émises. Elles ont donné lieu à 276 erreurs, soit un pourcentage de reconnaissance de 89,8 %.

10% des erreurs proviennent d'une absence totale d'identification de la consonne : [ka] est perçu comme [a], [ʒi] est perçu comme [i], [z ɛ d] est confondu avec [ɛl, ɛn, ɛm]. Ces erreurs tiennent vraisemblablement aux méthodes de synthèse (explosion de k insuffisamment intense et durée trop élevée) et au mode de fonctionnement du vocodeur à canaux qui opère un choix binaire entre source d'excitation périodique et source de bruit : pour les constrictives voisées par exemple, seul le voisement est restitué.

90% des erreurs sont dues à la confusion d'une consonne avec une autre consonne.

- Les confusions qui mettent en jeu des oppositions minimales :

Opposition voisé/non voisé : la seule confusion de ce type concerne /p/. La totalité des erreurs (soit 28%) se fait avec /b/ . Le score de reconnaissance (72%) de cette consonne est l'un des plus faibles. Rappelons que /p/ est une consonne grave.

.../...

	i	e	ɛ	a	y	ø	o	p	t	k	b	d	g	f	s	ʃ	v	z	ʒ	m	n	l	R	non perçu	Emissions	Fautes
i	92	1			0,6	6,4																			390	31
e	0,1	99.9																							743	1
ɛ			100																						1519	0
a			0.2	99.6		0.2																			555	2
y					98	2																			254	5
ø	0.6				1.4	97.4	0.6																		506	13
o						1.3	98.7																		298	4
p								71.9			28.1														32	9
t									100																48	0
k										94.6														5.4	147	8
b											97.6	2.4													41	1
d												92.2								0.4	3.5	3.5	0.4		229	18
g													100												42	0
f														73.4	26.6										79	21
s														0.7	99.3										400	3
ʃ														0.8		99.2									123	1
v								0.6			0.6						98.8								170	2
z																		80.4						19.6	92	18
ʒ																	2.2		97.2					0.6	181	5
m																				84.9	8.0	7.1			112	17
n																				13.1	70.6	16.3			449	132
R																						100			223	0
l																				7.6	2.6	1.7	88.1		343	41

TABLEAU II : REPARTITION DES CONFUSIONS DES SONS

Opposition nasal/non nasal : elle concerne l'opposition n/l (ROSSI, 1971), /n/ , qui possède le plus faible score d'intelligibilité (70,6%), est confondu avec /l/ dans 16% des cas.

Opposition grave/aigu :

Scores de reconnaissance :

/m/ = 84,9 % ; 53% des confusions avec /n/

/n/ = 70,6 % ; 44,7 % des confusions avec /m/

/f/ = 73,4 % ; toutes les confusions sont faites avec /s/

/s/ = 99,3 % ; toutes les confusions sont faites avec /f/

/b/ = 97,6 % ; toutes les confusions sont faites avec /d/

D'autres confusions mettent en jeu des oppositions complexes :

/m/ + nasal , - aigu est confondu à 47,7 % avec /l/ - nasal, + aigu

/z/ + compact , + aigu est confondu à 80 % avec /v/ - compact, - aigu

/d/ - vocalique, - continu est confondu à 44,5% avec /l/ et aussi à 44,5% avec /n/

/r/ consonne polymorphe dont le taux de reconnaissance est de 88% est confondu avec /m/à 63%, avec /n/ à 22% et à 15% avec /l/

Tous ces résultats sont cohérents par rapport à une étude antérieure réalisant les projections des phonèmes sur des axes factoriels (GRAILLOT, 1974). Ils sont à rapprocher aussi d'autres études (PECKELS et ROSSI, 1971 ; CARTIER et ROSSI, 1973) portant sur les performances des vocodeurs au test de rime et qui montraient, entre autre, que la majorité des erreurs se rapportait à l'opposition grave/aigu. Ce que l'on observe, en définitive, c'est à quel point les confusions sont localisées (/ m, n, r / qui ne représentent que 34% de toutes les émissions consonantiques sont à l'origine de 69% de toutes les erreurs faites sur consonne), et comment un son peut être confondu avec un autre sans que la réciproque soit vraie ; de tels résultats sont très courants dans les matrices de confusion.

2 - Les prénoms

37 prénoms français courants et courts (une ou deux syllabes) étaient proposés dans les 45 lignes d'annuaire ; 14 d'entre eux ont donné lieu à une erreur faite par un ou plusieurs auditeurs. La reconnaissance globale de cette information est de 95% à la synthèse, 100% en naturel et 98% en analyse-synthèse.

3 - Les professions :

45 professions différentes étaient proposées. 9 d'entre elles ont fait l'objet d'erreurs.

.../...

La reconnaissance globale des professions est de 96% ; elle est de 98% en analyse-synthèse ; si l'on compare les résultats synthèse par diphones-transmission par analyse-synthèse sur la seule liste qui a servi à la transmission par vocodeur, on constate que le pourcentage d'erreur est identique.

4 - Les adresses :

Dans cette catégorie, nous avons réparti des dénominations connues et des adresses imprévisibles. C'est sans doute dans ce type d'information que l'on va obtenir des réponses qui révèlent largement le niveau socio-culturel des auditeurs. Autrement dit, quand l'information donnée fait référence à quelque chose de connu, il est assez aisé de la reconstituer à partir de quelques sons reconnus, au contraire si les sons entendus ne se calquent sur aucun acquis, la réponse donnée n'est au mieux qu'une approximation acoustique sans reconstruction sémantique. Les résultats semblent nous donner raison : en effet, sur les 45 adresses proposées, 20 (les plus connues) ne font l'objet d'aucune erreur. Malgré les taux d'erreurs importants sur certaines adresses, la reconnaissance globale de cette information - que les auditeurs s'accordent à considérer comme la plus difficile à déchiffrer - est quand même de 82%. Cependant, si l'on reprend la comparaison sur la liste "analyse-synthèse", on s'aperçoit une fois de plus que la synthèse par diphones obtient un score de reconnaissance (96%) identique à celui de la transmission par vocodeur.

5 - Les communes :

Comme précédemment, un phénomène de reconstitution s'effectue ici. En synthèse, l'intelligibilité pour cette information est de 82%. Synthèse par diphones et transmission par vocodeur donnent le même score de reconnaissance : 86%.

6 - Les numéros de téléphone :

Sur les 45 numéros écoutés par 208 personnes, 7 ont fait l'objet d'une erreur. Les scores de reconnaissance obtenus pour ce type d'information sont élevés (98%). Les 2/3 des erreurs sont dues à la confusion 509 $\rightarrow$ 69 ; les autres erreurs sont réparties : 20 $\rightarrow$ 1 ; 76 $\rightarrow$ 67 ; 76 $\rightarrow$ 65 ; 3 $\rightarrow$ 30. On notera que ces erreurs sont corrigées par une répétition chiffre par chiffre de tout le numéro (système anglo-saxon).

III- BILAN

1 - Bilan de l'expérience par rapport à la méthodologie utilisée :

Les nombreuses questions posées par les auditeurs sur les principes et les buts de la synthèse, la précision des remarques notées sur les feuilles de réponse et la fréquence avec laquelle le magnétophone était encore déclenché longtemps après que le test ait officiellement pris fin montrent l'intérêt des sujets pour ce test. Il faut noter que la participation - la motivation - d'un grand nombre de personnes (une centaine) vient vraisemblablement de ce que, pour une fois, tout en étant "acteur", elles pouvaient demeurer dans leur espace, choisir le moment de leur action, c'est-à-dire finalement maîtriser tous les paramètres de leur intervention.

.../...

Cette sensation de liberté est à notre sens l'une des raisons essentielles du nombre élevé des réponses. Une autre raison tient aussi à l'aspect ludique du test : les mêmes lignes d'annuaire étant en général écoutées le même jour par des personnes se connaissant, il n'était pas rare de surprendre des conversations pour échanger les réponses trouvées.

A côté de ces avantages, il reste que certains résultats, comme l'accoutumance, sont difficilement chiffrables au moyen de ce test. En effet, chaque liste étant différente pour des auditeurs différents, on ne maîtrise pas ce paramètre ; on aurait pu sans doute l'appréhender avec un corpus plus large, en faisant écouter des listes identiques à des groupes d'auditeurs différents et en établissant des pondérations. Lors d'une expérience précédente portant sur l'intelligibilité des noms patronymiques (SORIN et LARREUR, 1978) l'importance d'une accoutumance (amélioration de 20% sur le pourcentage d'intelligibilité entre la première et la troisième écoute) a été mise en évidence pour la synthèse par diphones.

On peut dire que tel qu'il a été précisément réalisé, ce test est complémentaire d'expériences réalisées en laboratoire qui, si elles permettent de contrôler mieux les paramètres, s'éloignent souvent des conditions réelles.

2 - Bilan de l'expérience par rapport à l'évaluation de la qualité de la parole synthétique :

* Le nom patronymique est reconnu de façon parfaite dans 68% des cas quand il est répété lettre par lettre (dans 91% des cas en parole naturelle) ; en cas d'épellation complète (A comme Anatole), on arrive à 94% de reconnaissance. En ce qui concerne les autres informations (prénom, profession, adresse), on obtient des scores globaux qui sont respectivement de 95%, 96% et 82%. Enfin pour les numéros de téléphone, le pourcentage d'intelligibilité est de 98,5%. Mais il convient d'introduire une réserve :

Le corpus proposé aux auditeurs ne permet d'obtenir véritablement une bonne idée de l'intelligibilité que sur les numéros de téléphone car tous les chiffres sont proposés ; pour toutes les autres informations, où les scores de reconnaissance sont très disparates selon la difficulté phonétique des lignes d'annuaire, il aurait fallu disposer d'un matériau sonore plus complet qui présente dans toutes les positions toutes les occurences phonémiques. Mais il reste que dans le cadre de cette application, le nom patronymique n'est pas l'information la plus intéressante à étudier puisque dans la plupart des cas, l'abonné recherche plutôt le numéro de téléphone d'un correspondant dont il connait le nom.

Le phénomène d'accoutumance, qui n'est pas négligeable, intervient dans cette expérience. Or dans les conditions réelles d'application (automatisation des Centres de Renseignements), l'accoutumance à la voix synthétique ne jouera que faiblement : la fréquence d'appel pour une même personne n'est pas telle qu'elle l'autorise.

* Les résultats qualitatifs - que nous n'abordons pas ici- obtenus par les réponses des auditeurs au questionnaire ont montré que pour de telles applications, si on arrive à des résultats suffisants en intelligibilité, il faudra s'attacher encore à améliorer le naturel de la voix synthétique.

.../...

CONCLUSION

Que penser en définitive de ce test si l'on se place dans l'optique générale de la mesure de l'intelligibilité de la parole artificielle ? A cause de la motivation et de l'intérêt que les auditeurs ont manifesté pour cette expérience et à cause de la cohérence des résultats, nous pensons que cette procédure, de par sa conception, est intéressante. En outre il est tout à fait envisageable :

- d'utiliser la même méthodologie à d'autres fins : tester par exemple le rôle de la prosodie sur l'intelligibilité de la parole de synthèse, évaluer l'intelligibilité des diphones avant stockage définitif ; il suffit simplement d'adapter le matériel sonore.

- de comparer des systèmes différents susceptibles de répondre à la même application. En l'espèce, où il s'agissait de tester un système en vue d'une application bien précise, il semble que l'on dispose d'éléments sérieux sur l'efficacité et l'opportunité de la mise en service d'un système de renseignement téléphonique avec synthèse vocale.

BIBLIOGRAPHIE

CARTIER, M., et ROSSI, M., Le test de diagnostic par paires minimales : mise en oeuvre et résultats ; Symposium : Intelligibilité de la parole. Liège, 191-208, 12-15 nov. 1973.

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

EMERARD, F., et GRAILLOT, P., Qualité de la parole synthétique transmise par téléphone, Institut de Phonétique de Grenoble, à paraître, 1979.

GRAILLOT, P., Projection, compression et reconstitution de données spectrales de parole, Bulletin de l'Institut de Phonétique de Grenoble, Vol. III, 52-71, 1974.

PECKELS, J.P., et ROSSI, M., Le test de diagnostic par paires minimales, adaptation au français du "Diagnostic Rhyme Test" de Voiers, Journées d'Etude sur la Parole, Aix-en-Provence, Bd-Bh4, 1971.

SORIN, C., et LARREUR, D., Intelligibilité de noms propres synthétiques, Rech./ Acoustique. CNET LANNION, Vol. IV, 111-128, 1977.

10èmes JOURNÉES D'ÉTUDE SUR LA PAROLE

GRENOBLE - 30 MAI - 1ER JUIN 1979

UN MODELE ARTICULATOIRE DE LA LANGUE AVEC DES COMPOSANTES LINEAIRES.

S. MAEDA

C.N.E.T. - LANNION

RESUME

Pour une description de la parole, la langue peut être considérée comme composée de systèmes mécaniques, indépendamment contrôlés tels que la mâchoire, l'articulateur du corps de la langue et les articulateurs dorsal et apical. Les composantes linéaires représentant de tels articulateurs (voir Fig.2) ont été déterminées par une analyse statistique des formes du conduit vocal (voir Fig. 1), faite à partir de films cinéradiographiques latéraux. 17 logatomes [pV$_1$ CV$_2$] (V$_1$, V$_2$ = [i, a, u] et C =[d,g]) lus par un locuteur féminin ont été analysés. La somme des quatre composantes linéaires (représentée par l'équation (1)), décrit de façon adéquate les contours de langue observés. L'équation linéaire peut donc être considérée comme un modèle articulatoire. L'avantage d'un tel modèle rédigé semble-t-il, dans sa capacité à décrire la dynamique du conduit vocal. En effet des résultats préliminaires (Voir fig. 4) indiquent que les mouvements des paramètres articulatoires pour les logatomes V$_1$CV$_2$ pouvaient être interprétés de façon rationnelle.

AN ARTICULATORY MODEL OF THE TONGUE WITH LINEAR COMPONENTS.

S. MAEDA.

SUMMARY

For the description of speech event, the tongue may be regarded as a composit of independently controllable mechanical systems, such as the jaw, tongue body, dorsal, and apical articulator. Linear components representing such articulators (see Figure 2) were determined by a statistical analysis of vocal tract shapes (see Figure 1) measured from the lateral x-ray films. A set of 17 [p V_1 C V_2] utterances (where V_1, V_2 = [i, a, u] and C= [d, g]) read by a female speaker was subjected to the analysis. The sum of the four components, as represented in Eq. (1), describes adequately the obserbed tongue shapes. The linear equation, therefore,can be regarded as an articulatory model. An advantage of such model is said to be in the ability of describing the dynamics of the vocal tract. Indeed, a preliminary result (shown in Figure 4) indicated that the vocal-tract dynamics during V_1CV_2 contexts were interpreted rationally in terms of the corresponding movements of the articulatory parameters.

10èmes JOURNÉES D'ÉTUDE SUR LA PAROLE
GRENOBLE - 30 MAI - 1ER JUIN 1979

UNE MODELE ARTICULATOIRE DE LA LANGUE AVEC DES COMPOSANTES LINEAIRES.

S. MAEDA.

INTRODUCTION

Dans une précédente publication (MAEDA, 1978) nous avons essayé d'appliquer l'analyse en composantes principales sur les mesures faites sur la coupe sagittale de la langue. Cette analyse a fourni la description la plus économe, d'un point de vue statistique, des mouvements de la langue. Toutefois, certaines des composantes principales sont apparues comme étant représentatives d'un ensemble d'articulateurs différents, ce qui présentait certaines difficultés d'interprétation. Le problème de base dans ce traitement purement mathématique réside dans le fait qu'il n'est pas possible d'utiliser une connaissance à priori du système articulatoire humain. C'est pourquoi, aboutir, par cette méthode, à une série de composantes physiquement interprétables relève plutôt de la chance.

Nous allons décrire dans cet article, une autre méthode statistique où l'analyse est guidée par une description qualitative, c'est-à-dire par un "modèle théorique" du système articulatoire.

D'après LINDBLOM et SUNDBERG (1971), les contours de la coupe sagittale moyenne de la langue sont déterminés par 3 composantes représentatives de la mâchoire, du corps de la langue, de la partie apicale de la langue. La composante "corps de la langue" est déterminée par deux paramètres concernant sa position et sa forme. Les deux paramètres semblent être reliés respectivement à l'articulateur du corps de la langue et à l'articulateur dorsal, suggéré par OEHMAN (1967). Il est donc raisonnable de supposer que les formes de la langue sont fonction des quatre systèmes mécaniques suivants, qui sont contrôlés indépendamment :

- un articulateur des mâchoires
- un articulateur du corps de la langue
- un articulateur dorsal
- un articulateur apical.

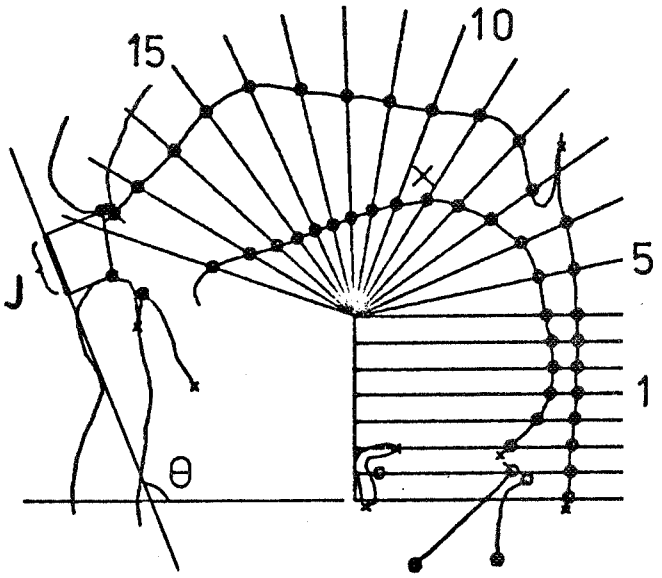
Comme l'influence des deux derniers articulateurs sur les contours de la langue est limitée respectivement aux régions dorsale et apicale, on peut supposer que dans la région laryngienne, les articulateurs de la mâchoire et du corps de la langue sont déterminants pour la spécification des contours. Dans la région dorsale, les deux articulateurs ci-dessus plus l'articulateur dorsal déterminent les formes de la langue. Dans la région apicale, les 4 articulateurs peuvent être déterminants, de façon égale, pour la description de la pointe de la langue. Cette description grossière du système articulatoire servira de guide dans notre analyse.

Ainsi, si le modèle factoriel résultant est formé des composantes linéaires qui représentent ces articulateurs individuels, la capacité du modèle à décrire les formes de la langue ne sera pas limitée aux seules formes soumises à l'analyse.

Autrement dit, le modèle pourra être utilisé comme modèle articulatoire général. Un tel modèle est sans doute bien adapté pour décrire en particulier la dynamique de la langue, si les articulateurs individuels peuvent être commandés par une loi simple.

I - UN MODÈLE GÉNÉRAL EN COMPOSANTES LINÉAIRES ET LA STRATÉGIE D'ANALYSE.

Les formes de la langue peuvent être mesurées et décrites dans un espace à une dimension en utilisant des coordonnées semi-polaires fixes par rapport au palais dur (Fig.1), (pour plus de détails, voir MAEDA, 1978). A un instant donné la forme de la langue est décrite par un vecteur (x_1, x_2..... x_p) p =16, dans lequel les variables x_j représentent les mesures dans ce système de coordonnées.



- Figure 1 -

Mesure des formes de la langue en utilisant des coordonnées semi-polaires repérées par rapport au palais dur, et mesure du paramètre mâchoire, J.

The measurement of the tongue shapes using the semi-polar coordinate fixed with respect to the hard palate , and of the jaw parameter, J.

Dans l'analyse statistique, les formes de la langue observées sont représentées par la somme des composantes linéaires comme suit :

$$\begin{bmatrix} z_1 \\ z_2 \\ \cdot \\ \cdot \\ z_p \end{bmatrix} = \begin{bmatrix} a_{11} \\ a_{21} \\ \cdot \\ \cdot \\ a_{p1} \end{bmatrix} f_1 + \begin{bmatrix} a_{12} \\ a_{22} \\ \cdot \\ \cdot \\ a_{p2} \end{bmatrix} f_2 + \ldots + \begin{bmatrix} a_{1p} \\ a_{2p} \\ \cdot \\ \cdot \\ a_{pp} \end{bmatrix} f_p \qquad (1) ,$$

où z_1 z_2.....z_p représentent des variables normées de la langue ayant une variance unitaire et une moyenne nulle, reliées aux variables originales par la relation :

$$x_j = r_j z_j + \bar{x}_j \qquad j = 1,2,\ldots\ldots, p \qquad (2) ,$$

où r_j et $\bar{x}_j$ correspondent respectivement au carré de l'écart-type et à la moyenne de x_j. Les coefficients a_{1j} , a_{2j}... a_{pj} sont les poids des différents facteurs et précisent le mode d'influence du j-ème facteur (normé) f_j, sur les formes de la langue.

Comme on l'a vu dans l'analyse en composantes principales, seules quelques composantes sont nécessaires pour reconstituer les formes originales de la langue, avec une bonne précision. Nous voulons déterminer le poids tel que chaque composante linéaire représente un seul articulateur. Dans ce cas, le modèle linéaire apparaît comme un modèle articulatoire et les facteurs correspondent aux paramètres articulatoires.

La procédure appliquée dans cette étude est appelée "modèle général de composantes linéaires" décrit par OVERALL (1962). Ce modèle permet d'analyser des données en utilisant dans un premier temps toute la connaissance a priori (obtenue dans notre analyse par la méthode diagonale), puis dans un deuxième temps de détecter les composantes inconnues par exemple par une analyse en composantes principales.

Fait important, plusieurs méthodes différentes peuvent être mélangées, pouvu que les facteurs calculés soient toujours orthogonaux deux à deux (c'est-à-dire, non corrélés). Ensuite, on voit facilement d'après l'équation (1) que la proportion de la variance des données, s_k, extraite (ou, comme nous le dirons dorénavant, "expliquée") par la k-ième composante, est donnée par la somme des poids élevés au carré :

$$s_k = \frac{1}{p} \sum_{i=1}^{p} a_{ik}^2 \qquad (3),$$

et que la proportion de la variance due aux N composantes c'est-à-dire, la proportion cumulée, est simplement la somme des s_k, soit $S_1 + S_2 + \ldots\ldots + s_N$. La proportion s_k correspond à une mesure de l'importance de la k-ième composante dans la description des données de la langue. On peut également noter que chaque poids élevé au carré ajK^2, indique l'importance du k-ième paramètre par rapport à la j-ième variable x_j. Ceci est une procédure simple et claire.

Décrivons maintenant la stratégie d'analyse. Si l'on possède un paramètre indiquant le degré d'une articulation particulière, les poids peuvent être estimés par intercorrélations entre ce paramètre et les variables mesurées de la langue. Sur la base des données radiographiques dont nous disposons, les paramètres ne peuvent toutefois être mesurés directement, sauf celui de l'articulateur de la mâchoire où la position de l'incisive inférieure centrale peut être considérée comme ce paramètre. C'est pourquoi nous commençons par estimer les intercorrélations entre les p mesures de la langue et le paramètre unique de la mâchoire, en formant la matrice R de corrélation, de dimension $(p + 1) \times (p + 1)$. Les poids pour le paramètre de mâchoire, ne sont rien d'autre que les coefficients de corrélation entre ce paramètre et les mesures de la langue, notés par le vecteur $a_1 = [a_{11}\ a_{21}, \ldots\ldots\ a_{p1}, 1]^T$. (L'unité en fin d'expression correspond au poids pour la mâchoire elle-même). Le trait remarquable de cette méthode est que l'influence de la composante mâchoire sur la matrice R de corrélation est éliminée par les soustractions suivantes :

$$R_1 = R - a_1\, a_1^T , \quad (4)$$

où R_1 représente la dispersion résiduelle (c'est-à-dire la matrice variance-covariance) de rang p. Cette procédure d'élimination après chaque détermination des poids assure l'orthogonalité mutuelle des facteurs.

Il faut noter que la matrice de corrélation après l'élimination, ne contient plus aucune influence de la composante mâchoire.

Ceci implique que l'une quelconque des mesures dans la région pharyngienne peut être considérée comme paramètre de la composante du corps de la langue, puisque les formes de la langue dans cette région sont déterminées essentiellement par les composantes "mâchoire" et "corps de la langue", conformément au modèle théorique décrit ci-dessus. Ainsi, en prenant une mesure de la langue dans cette région comme paramètre du corps de la langue, les poids peuvent être calculés à partir de la dispersion résiduelle R_1. De même l'influence de la composante du corps de la langue est éliminée de R_1, ce qui permet de déterminer le poids de la composante dorsale, en prenant une mesure quelconque de la langue dans la région dorsale comme paramètre dorsal. Enfin, le poids correspondant à la composante apicale peut être calculé après élimination de la composante dorsale.

Il n'est pas nécessaire, bien sûr, que les 4 composantes rendent compte complètement de la variance des données sur la langue ; des sources inconnues de variance peuvent exister. De telles composantes inconnues, si elles existent, peuvent être détectées en soumettant la dispersion résiduelle finale à une analyse en composantes principales.

II - L'ANALYSE ET SES RÉSULTATS

Environ 400 images représentant des coupes sagittales du conduit vocal ont été retracées manuellement à partir de films radiocinématographiques à 50 images/s correspondant à 17 séquences $[p\ V_1\ CV_2]$ ou $V_1, V_2 = [i, a, u]$ et $C = [d, g]$, lues par un locuteur féminin. Les tracés ont alors été traités par ordinateur pour effectuer une analyse statistique.

II.1 - Mesures de l'ouverture des mâchoires

Les mouvements de mâchoires se composent d'une rotation et d'un glissement. Le premier paraît dominant lors de l'articulation des voyelles. Divers auteurs (comme MERMELSTEIN, 1973 ; SHIRAI et HONDA, 1976) supposent une rotation de la mâchoire autour d'un axe fixe par rapport à une structure stationnaire comme le palais dur. Dans ce cas, le degré d'ouverture de la mâchoire peut être décrit par un paramètre unique, l'angle de rotation. Il n'est toutefois pas aussi évident de préciser comment l'axe de rotation peut être déterminé. Nous avons ici tenté d'utiliser une autre voie pour définir l'ouverture de la mâchoire, Si le mouvement de la mâchoire est mesuré à l'extrémité de l'incisive inférieure centrale, son amplitude sera très petite par rapport au rayon du mouvement de rotation. Le déplacement de l'incisive peut donc être approché par une ligne droite. Alors le paramètre mâchoire J peut être défini comme la projection de la distance entre les incisives centrales supérieure et inférieure sur une ligne droite, comme le montre la figure 1. Evidemment, "l'angle de vue" θ, c'est-à-dire l'angle entre la ligne droite et la ligne horizontale intervient sur la valeur de J. La valeur de θ a été déterminée de façon à ce que la proportion de la variance des données de la langue correspondant au paramètre J soit maximale.

II.2 - Composante de la mâchoire

Pour trouver la valeur otpimale de θ, nous avons calculé la proportion de la variance correspondant à la composante de la mâchoire en faisant varier de 75° à 155° par pas de 10°. La proportion de la variance a une valeur maximale de 0,44 pour $\theta = 115°$. Cela signifie que 44% de la variance dans les données de la langue est due à la composante de la mâchoire.

Le mouvement de la langue dû aux variations du paramètre mâchoire apparaît clairement dans la reconstitution des contours de la langue. Une forme de langue a été calculée à partir des équations (1) et (2). On a reporté sur un plan en coordonnées semi-polaires, sur la Fig.2 (a), les déviations des positions de la langue par rapport à la position moyenne, pour les valeurs + 1 et -1 du paramètre. La composante mâchoire semble déterminer, au moins en partie, la hauteur de la langue. On doit noter cependant qu'une élévation et un abaissement de la langue sont associées respectivement à un mouvement antérieur et postérieur dans la région pharygienne.

II.3 - Composante du corps de la langue

Après avoir éliminé l'influence de la composante mâchoire de la matrice R de corrélation, n'importe quelle mesure de la langue dans la région pharyngienne peut être considérée comme paramètre du corps de le langue. Il apparaît raisonnable, toutefois, de choisir la mesure à laquelle correspond le maximum de variance résiduelle par rapport aux données de la langue.

Prenant ces faits en considération, nous avons choisi la 5ième variable de la langue (voir fig.1 la position correspondante sur la coordonnée semi-polaire) comme paramètre du corps de la langue . L'effet de cette composante sur les formes de la langue est illustré sur la Fig.2 (b). Vingt sept pour cent de la variance est dû à cette composante. En comparaison avec l'influence de la composante mâchoire, l'influence de cette composante semble se réaliser dans le mouvement avant-arrière du corps de la langue.

II.4 - Composante dorsale

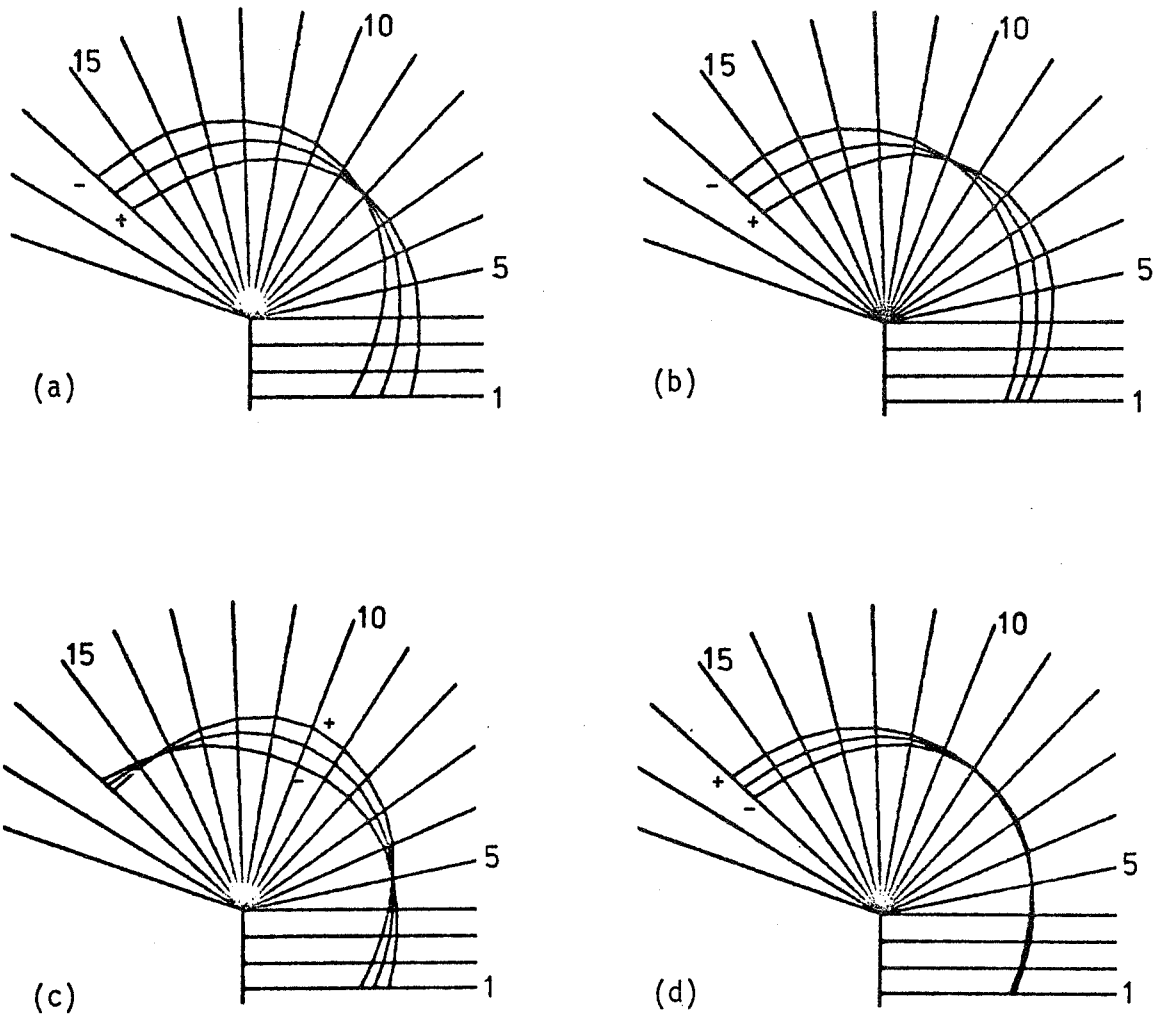
Après élimination de l'influence des composantes mâchoire et corps de la langue, les poids du facteur pour la composante dorsale (le 3ième terme de l'équation (1)) ont été déterminés en considérant la 9ème variable de la langue comme paramètre. La composante dorsal rend compte de 23 % de la variance. La contribution de cette composante dans les contours de la langue est portée sur la Fig. 4c). La déviation des contours indique une déformation du corps de la langue, dans le sens incurvé (marqué par +1) ou plat (marqué par -1). C'est pourquoi le paramètre dorsal modifie essentiellement la forme de la langue.

II.5 - Composante apicale

L'influence de la composante apicale sur les formes de la langue est montrée à la figure 2(d). La 15ième variable choisie est le paramètre apical. Nous avons trouvé que 5 % seulement de la variance est dû à cette composante. Bien que la proportion de la variance correspondante apparaisse relativement faible, dans la région du bout de la langue, sa contribution est significative.

II.6 - Composantes inconnues

Les quatre premières composantes, prises ensemble, rendent compte de 98 % de la variance et seulement 2 % de la variance peut être attribué à l'effet de sources inconnues. La dispersion finale résiduelle a été soumise à une analyse en composantes principales. La composante principale la plus significative qui rend compte de 11 % environ de la variance, semble améliorer la description de la région pharyngienne inférieure. Si l'on vise à décrire les voyelles, une telle composante n'est probablement pas essentielle. C'est pourquoi, si l'on n'a pas besoin d'une description très précise des formes de la langue, cette composante



- FIGURE 2 -

Effet des quatre composantes linéaires sur les contours de la langue: les composantes,

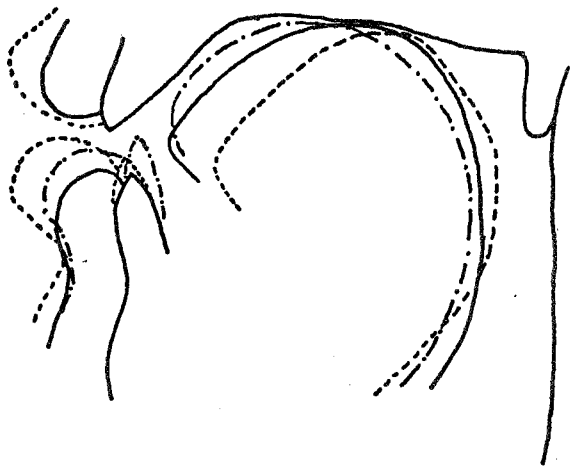
(a) "mâchoire", (b) "corps de la langue",
(c) "dorsal", (c) "apical".
Les symboles + ou - indiquent la valeur +1 ou -1 des paramétres normée. Les contours sans symboles représentent la position moyenne de la langue.

The effect of the four linear components upon the tongue shapes; in (a) the jaw component, in (b) the tongue-body component, in (c) the dorsal component, and in (d) the apical component. The symbols '+' and '-' indicate the value +1 and -1, respectively, of the standardized parameters. The contours without the symbols represent the mean tongue position.

peut être négligée. Dans ce cas, les contours de la langue peuvent être décrits par les quatre composantes linéaires: les composantes de la mâchoire, du corps de la langue, dorsale et apicale.

III - DISCUSSION.

Nous avons démontré qu'un modèle articulatoire linéaire de la langue pouvait être construit de façon systématique par une analyse statistique. Le modèle est simple et capable de décrire adéquatement les contours de langue observés. Jusqu'à quel point la dynamique de la langue peut-elle être décrite par un tel modèle? Nous ne sommes pas en mesure de répondre à cette question. Néanmoins, il peut être interessant de discuter de sa capacité à décrire la dynamique de la langue.



- Figure 3-

Formes du conduit vocal pendant l'occlusion de [g] dans trois contextes différents symmétriques: [paga] (——), [pugu] (-------), [pigi] (—·—·—).

Vocal tract shapes of [g]-closure in three symmetric vowel contexts: [paga] (——), [pugu] (-------), [pigi] (—·—·—).

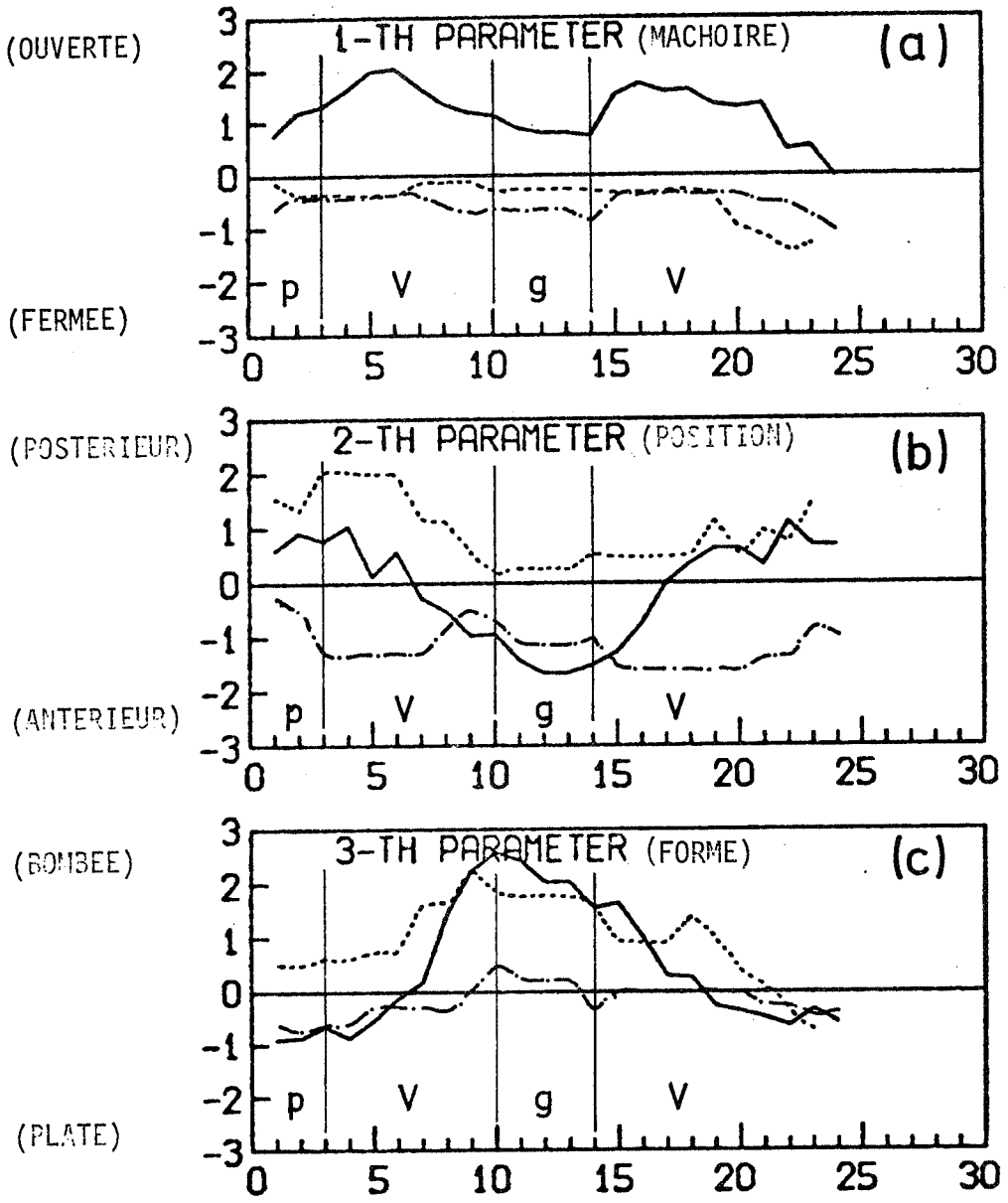
On peut noter que la position de l'occlusion dépend du contexte.

Notice that the place of closure differs depending on the context.

Nous avons montré à la Fig.3 trois contours de langue pour la même consonne [g], extraite de trois contextes phonétiques différents, [paga] (trait plein), [pugu] (trait pointillé) et [pigi] (autre trait). Öhman (1967) a noté une différence semblable entre les points d'articulation de [g] pour un locuteur suédois. Cette différence est considérée comme étant due au phénomène de coarticulation. Ce dernier peut être étudié directement par l'étude des variations dans les contours de langue. Cependant, la représentation de la dynamique en termes de paramètres articulatoires offre, semble-t-il, un moyen de parvenir à une compréhension plus approfondie de la manière d'articuler.

Les figures 4a, 4b et 4c indiquent les variations des trois premiers paramètres représentant respectivement la position de la mâchoire, la position antérieure-postérieure de la langue et le contour dorsal. Dans chaque figure, le trait plein correspond à [paga], le trait pointillé à [pugu] et les tirets à [pigi] (comme sur la figure 3).

Il est évident que le point d'articulation de la consonne [g] est différent. Dans le cas de [paga], la mâchoire est moins fermée que dans les cas de [pugu] et de [pigi]. Un mouvement important vers l'avant et une courbure considérable du corps de la langue semblent compenser le fait que la mâchoire est moins fermée. On peut noter que la position de la langue pour [g] dans [paga] est



- FIGURE 4 -

Valeurs normées des trois premiers paramétres articulatoires dans les images successives du film cinéradiographique: [paga] (———); [pugu] (-------); [pigi] (—·—·—).

The standardized values of the first three articulatory parameters in the successive x-ray film frames of the utterances: [paga] (———); [pugu] (-------); [pigi] (—·—·—).

plus antérieure que pour les autres [g] (voir Fig.4b) et que le degré de courbure est le plus important (voir Fig.4c). Ce fait est naturel dans le sens que le mouvement vers l'avant de la masse de la langue élève effectivement la position de la langue (comme le montre la figure 4c), mais un supplément de courbure est nécessaire pour permettre à l'occlusion de se placer à un endroit approprié. Dans le cas de [pugu] , la position de la mâchoire est quasiment fixe. La fermeture de [g] est effectuée, semble-t-il, par un mouvement vers l'avant de la masse de la langue associé à une courbure. Par contraste, pour [pigi] , la mâchoire est plus fermée, bien que la position de la mâchoire durant la voyelle [i] soit la même que pour [u] . On peut fournir l'explication suivante: comme le lieu d'articulation pour [i] est antérieur à celui de [g] , il est nécessaire de réaliser un mouvement vers l'arrière (comme illustré sur la Fig.4b). Ce mouvement vers l'arrière résulterait, cependant, en un abaissement de la masse de la langue (voir Fig.2b). Pour compenser cet abaissement, la mâchoire doit être fermée. On peut noter, dans le cas de [pugu] qu'un tel mouvement compensatoire n'est pas nécessaire puisqu'on doit mouvoir la masse de la langue vers l'avant à partir d'une position très postérieure.

Il apparait donc que les mouvements des paramètres articulatoires peuvent être interprétés de façon satisfaisante. Une investigation plus poussée de la dynamique du conduit vocal en termes de paramètres articulatoires doit permettre de redécouvrir des principes tel que le principe "d'économie d'effort physiologique" sous-jacent à l'organisation de la dynamique articulatoire.

BIBLIOGRAPHIE

LINDBLOM, B.E.F. and SUNDBERG, J.E.F., 1971, "Acoustical Consequences of Lip, Tongue, Jaw, and Larynx Movement," J. Acoust. Soc. Amer. 50, 1166-1179.

MAEDA, S., 1978, "Une analyse statistique sur les positions de la langue : étude préliminaire sur les voyelles françaises," 9ièmes Journées d'Etude sur la Parole, Lannion, (G.A.L.F.), 191-199.

MERMELSTEIN, P., 1973, "Numerical Model of Coarticulation," J. Acoust. Soc. Amer. 41, 310-320.

SHIRAI, K. and HONDA, M., 1976, "Estimation of Articulatory Motion," in Dynamic Aspects of Speech Production, ed. M. SAWASHIMA and F.S. COOPER, University of Tokyo Press, Tokyo.

OVERALL, J.E., 1962, "Orthogonal Factors and Uncorrelated Factor Scores," Psychological Report, 10, 651-662.

Je remercie Madame Pela SIMON et son collègue Monsieur Gilbert BROCK de l'Université de STRASBOURG ainsi que le Docteur François WOLFF du Centre Médico-Chirurgical et Obstétrical de la Sécurité Sociale à STRASBOURG pour nous avoir permis de réaliser les radiocinématographies. Mademoiselle Michèle DUVERNEUIL de l'Université de GRENOBLE a accepté aimablement de servir de sujet.

10èmes JOURNÉES D'ÉTUDE SUR LA PAROLE

GRENOBLE - 30 MAI - 1ER JUIN 1979

UN OUTIL CONVERSATIONNEL POUR L'ANALYSE ET LA SYNTHESE DE LA PAROLE PAR PREDICTION LINEAIRE.

Jean SAMAKE, Jean-Paul HATON

Equipe de Traitement du Signal et Reconnaissance des Formes
Centre de Recherche en Informatique de Nancy
C.O. 140 54037 NANCY Cédex

RESUME Cet article présente les activités en synthèse de parole effectuées au CRIN en vue de la réalisation d'un organe de réponse vocale pour l'aspect dialogue du système de compréhension de la parole MYRTILLE II.

L'organe de synthèse utilise la technique de prédiction linéaire (autocorrélation). Sa structure, classique, est constituée de cellules du second ordre en série.

La stratégie de synthèse revient à produire des syllabes VCV à partir de la synthèse de phonèmes et de règles d'interpolation.

Nous avons développé un outil expérimental permettant de synthétiser un message écrit donné sous forme phonétique. Cet outil comporte en particulier un module conversationnel permettant de modifier dynamiquement soit les paramètres d'excitation de l'organe de synthèse (y compris la mélodie), soit les paramètres des éléments phonétiques stockés dans le dictionnaire.

UN OUTIL CONVERSATIONNEL POUR L'ANALYSE ET LA SYNTHESE DE LA PAROLE PAR PREDICTION LINEAIRE.

Jean SAMAKE, Jean-Paul HATON

SUMMARY This paper presents research in the field of speech synthesis carried out in our laboratory. The goal is to build a vocal response unit to be incorporated in MYRTILLE II Speech Understanding System in order to make an actual evaluation of the dialog in this system.

The analyzer uses the autocorrelation method of linear predictive coding. The synthesizer is classically made of second order cells connected serially. The synthesis strategy consists of producing VCV syllables from phoneme synthesis and interpolation rules.

Around this synthesizer we have developed an experimental tool which makes it possible to synthesize a written message given in phonetic transcription. This tool enables the user to interactively and dynamically modify either the control parameters of the synthesizer (including pitch contours) or the phonetic parameters stored in the dictionnary.

10èmes JOURNÉES D'ÉTUDE SUR LA PAROLE
GRENOBLE - 30 MAI - 1ER JUIN 1979

UN OUTIL CONVERSATIONNEL POUR L'ANALYSE ET LA SYNTHESE DE LA PAROLE PAR PREDICTION LINEAIRE.

Jean SAMAKE, Jean-Paul HATON

INTRODUCTION

Cet article présente le travail effectué dans notre équipe en synthèse de la parole en vue de la réalisation d'une unité à réponse vocale. Cette unité est destinée à être intégrée dans la partie dialogue du système de compréhension automatique de la parole MYRTILLE II en cours de réalisation.

Jusqu'à présent le dialogue était simulé (par exemple dans le système MYRTILLE I), une étude vraiment complète des réactions à la fois du locuteur et du système nécessite la mise en place d'une procédure de vrai dialogue oral.

La technique choisie peut être résumée comme une synthèse par phonèmes utilisant une analyse par prédiction linéaire (méthode d'autocorrélation).

Dans un premier paragraphe le système de synthèse par règles est décrit, ainsi que la stratégie permettant de passer d'un message écrit au signal acoustique.

Nous présentons ensuite un outil expérimental qui a été réalisé pour aider à la mise au point d'un tel système de synthèse. Il s'agit d'un ensemble de programmes conversationnels permettant de modifier dynamiquement (et par suite de constater immédiatement les changements obtenus) les paramètres de commande du synthétiseur et les paramètres obtenus lors de la phase d'analyse des phonèmes.

PRINCIPE UTILISE

La synthèse de la parole par règles nécessite dans notre système la mise en oeuvre de trois modules :

- un synthétiseur, jouant le rôle d'organe vocal par simulation du système de phonation.
- un stratège qui assure la transformation d'un ensemble composé d'une chaîne phonétique et des paramètres phonologiques associés en paramètres de commande du synthétiseur.

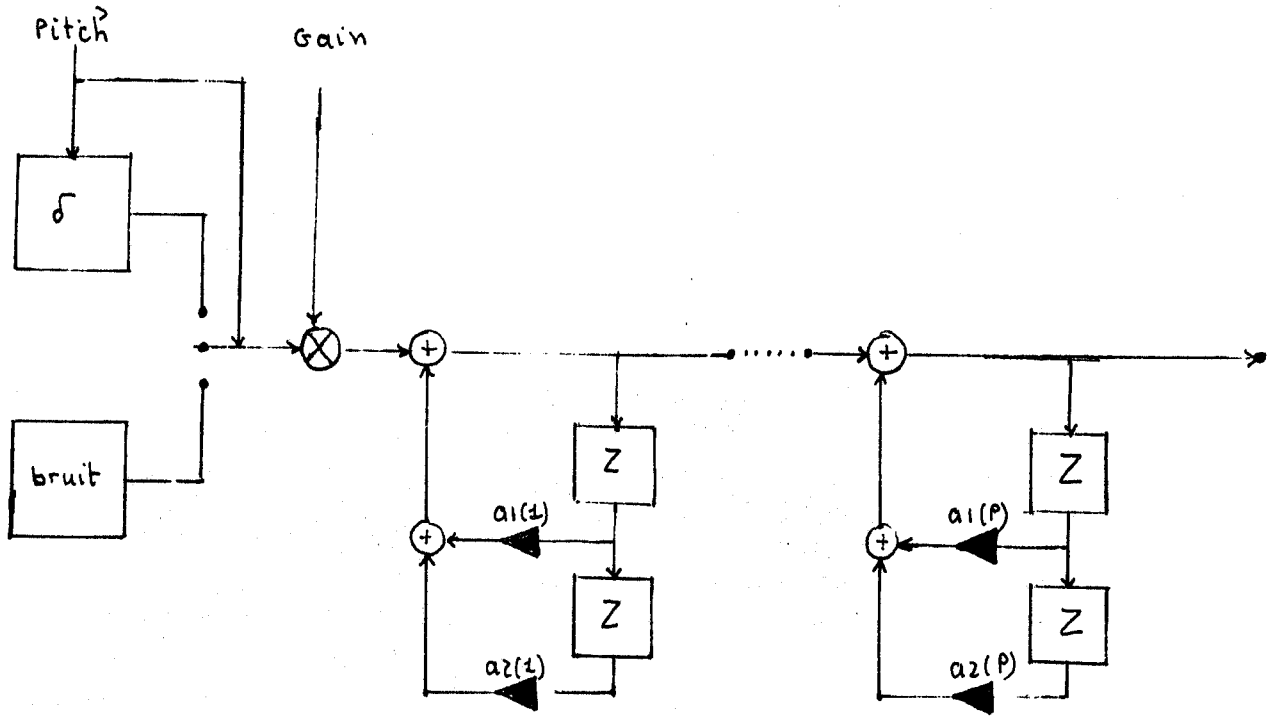


Fig. 1 SYNTHETISEUR

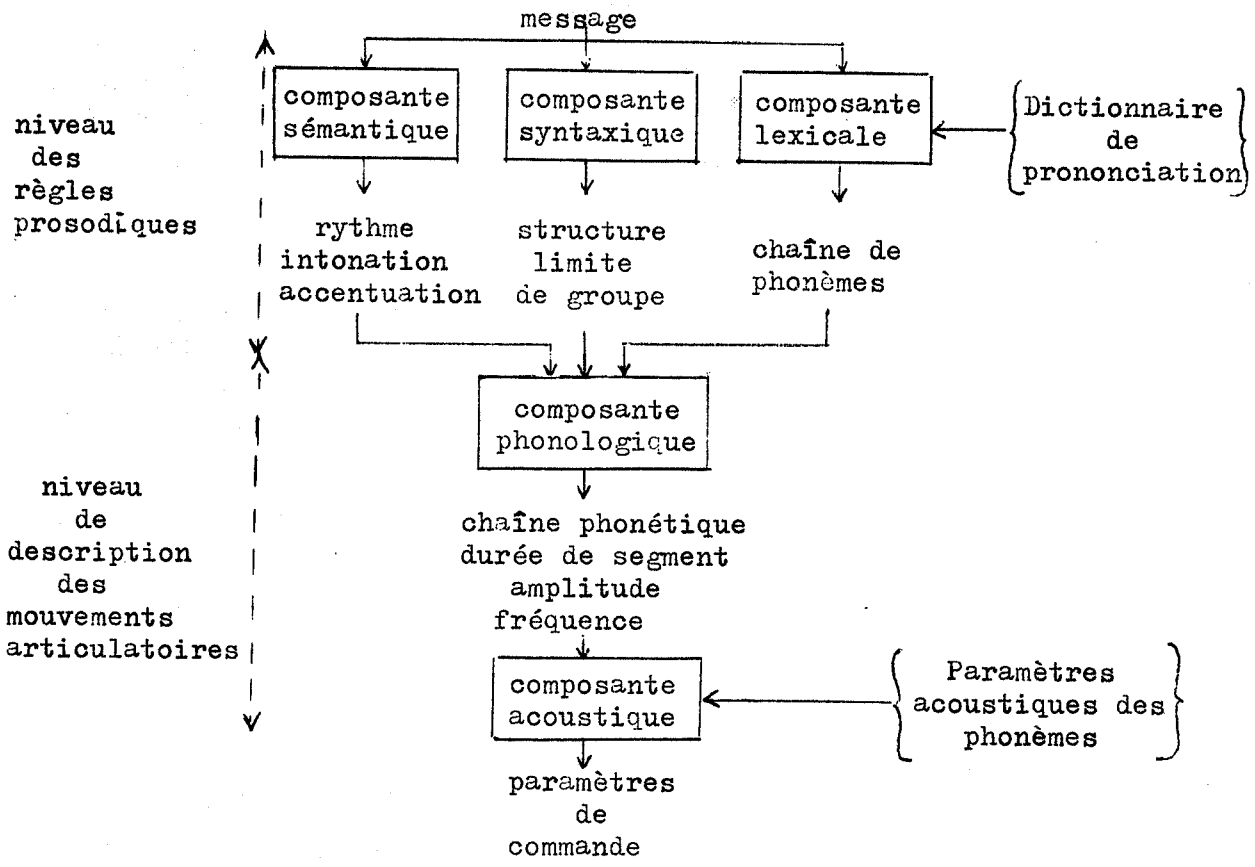


Fig. 2 GENERATEUR DE PHRASES

- un générateur de phrases chargé d'effectuer d'une part la transcription graphème-phonème du message (ce module n'existe pas dans le système, le message est supposé transcrit phonétiquement), et, d'autre part, la spécification des paramètres prosodiques du message en termes de durée, intensité et mélodie. Le schéma général est donné fig. 2.

1) Le synthétiseur

On peut montrer(SAMAKE,1977)que la fonction de transfert du modèle de système de phonation peut se réduire à un modèle tout-pôle. Notre synthétiseur est ainsi constitué de filtres digitaux du second ordre, en série, comportant uniquement des pôles. Ce système est excité soit par un générateur d'impulsion dans le cas d'un son voisé, soit par un générateur de bruit pour les sons non-voisés (fig. 1).

Les paramètres a_1 et a_2 de chacun des 3 premiers filtres ne dépendent que des fréquences et largeurs de bandes des formants. L'énergie du signal synthétique est fixée par le gain G. Le rôle du 4ème filtre consiste à équilibrer la répartition spectrale (RABINER, 1970).

Ce modèle diffère des modèles comportant deux branches (une pour les sons voisés, l'autre pour les sons non-voisés). Les résultats obtenus en analyse-synthèse fondée sur la prédiction linéaire ont montré qu'un modèle à branche unique tel que le nôtre, et ne comportant que des pôles, peut simuler de manière acceptable les différents sons du langage.

2) La stratégie de synthèse

La stratégie permet de préciser les règles définissant la suite des mots et les caractéristiques acoustiques des phonèmes, stockées dans un dictionnaire.

Les voyelles sont bien caractérisées par leurs trois premiers formants. Nous avons choisi de représenter les consonnes par une fonction linéaire de la voyelle adjacente, les paramètres de cette fonction dépendant de la nature de la consonne elle-même.

Le rôle des transitions étant fondamental, en particulier pour la perception des consonnes, nous assurons les différentes transitions au sein des groupes syllabiques VCV avec les choix suivants :

- les caractéristiques des transitions (vitesse du mouvement, retard) sont fonction de la consonne seulement
- les interpolations utilisent toutes une fonction de type exponentiel, sauf pour les largeurs de bandes formantiques qui sont interpolées linéairement.

RODET (RODET, 1977) utilise les mêmes groupes VCV mais avec des fonctions d'interpolation différentes.

IMPLANTATION DU SYSTEME

Le système de synthèse est implanté sur un minicalculateur SEMS Mitra 125. Il a été conçu comme un système conversationnel permettant de modifier à volonté, et de façon dynamique, les différents paramètres pour améliorer la qualité de la parole synthétisée.

Le schéma général du système est donné fig. 3. Le fonctionnement peut être observé par affichage des paramètres sur écrans alphanumérique ou graphique et sur imprimante. La modification de paramètres peut être effectuée à partir des claviers ou par tablette graphique (évolutions temporelles).

On distingue deux modes de fonctionnement :

- automatique : le système produit le message sonore correspondant aux phrases données en entier. Dans ce cas le superviseur SINVOC ne sert que de relais entre les différents organes
- semi-automatique : il est alors possible d'interrompre le fonctionnement et d'intervenir sur les paramètres par l'intermédiaire de SINVOC. Le message à synthétiser est transmis par SINVOC à SYLLABE qui effectue le découpage en groupes syllabiques. Les associations VCV ou VCCV sont ensuite traitées séparément pour assurer dans les différents cas les mouvements formantiques. Le module SORTIE permet de visualiser graphiquement l'évolution des paramètres (mouvements fréquentiels, intensité, mélodie)

En cas d'interruption par l'utilisateur, le superviseur SINVOC reprend la main. Trois types d'actions peuvent alors être lancées.

- modifications de paramètres (assurées par le module CHANGE) portant :
  . sur le dictionnaire des éléments phonétiques : CALCUL peut alors rejeter les modifications proposées si celles-ci ne respectent pas certains critères mathématiques.

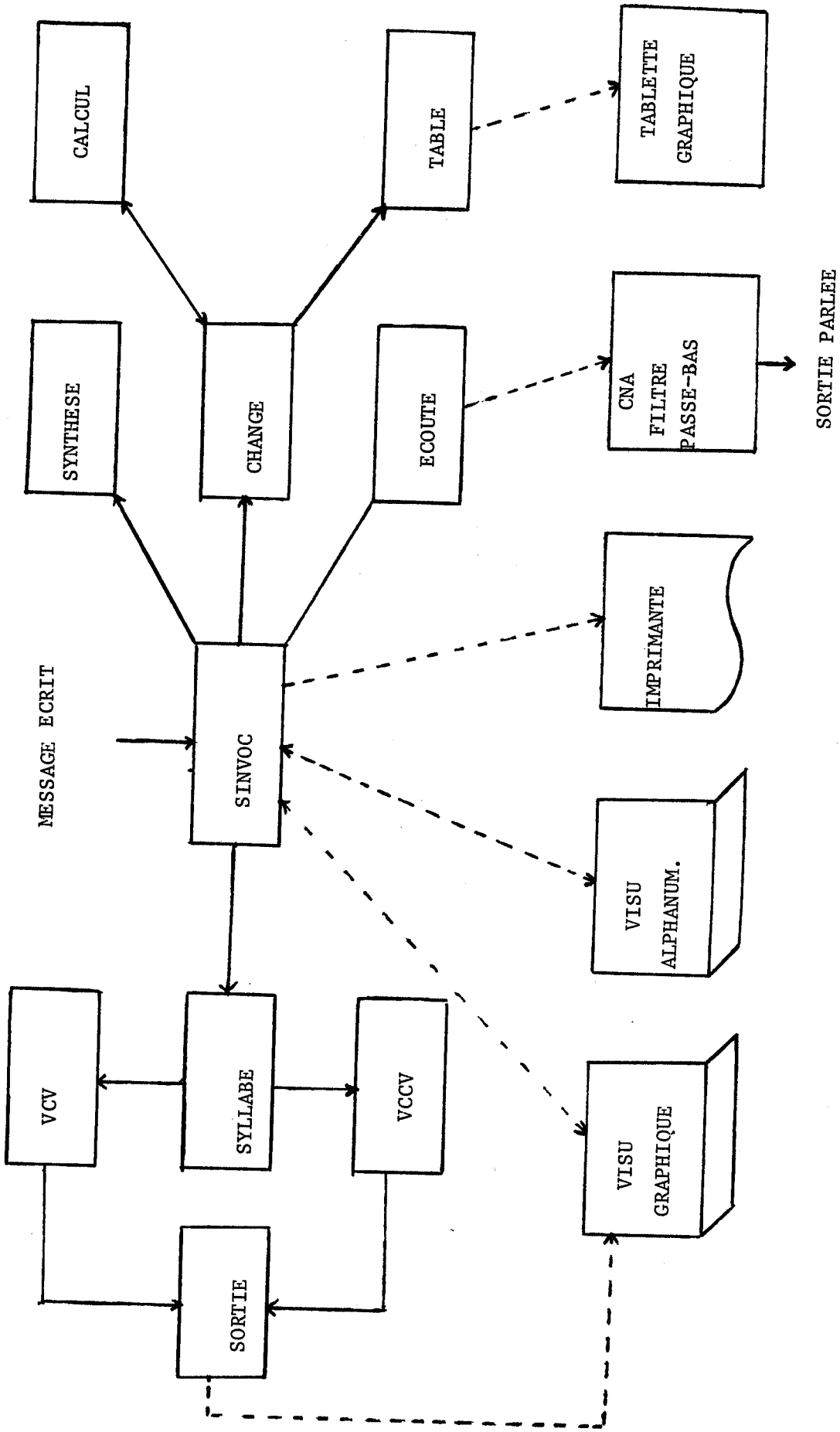


FIG. 3 Schéma d'ensemble

. sur les paramètres acoustiques (fréquence, intensité, mélodie) soit par envoi de valeurs à partir d'un clavier, soit par un tracé d'un nouveau contour de mélodie ou d'intensité à l'aide d'une tablette graphique.

- construction du signal sonore
  . par le module SYNTHESE, la synthèse est réalisée à l'aide des paramètres acoustiques définis
  . par le module ECOUTE, l'envoi du signal synthétique est effectué à travers un convertiseur numérique analogique.
- retour-arrière en un point quelconque du message, par exemple pour relancer un traitement depuis ce point.

La détermination des différentes zones du message est effectuée en positionnant les curseurs sur l'écran graphique. En cas d'une erreur quelconque, SINVOC émet un message sur l'écran alphanumérique, alors le traitement est soit abandonné, soit relancé au choix de l'opérateur.

CONCLUSION

Nous avons décrit un système de synthèse par règles utilisant un synthétiseur composé de quatre filtres digitaux en série, avec des pôles pour simuler aussi bien les sons voisés que les fricatives non voisées et sans zéro pour simuler les nasales, ceci en se fondant sur l'hypothèse que les zéros peuvent être approximés par des pôles.

Cette particularité a nécessité de définir des règles régissant les variations des largeurs de bandes des formants, en particulier pour mieux préserver les effets de nasalité. Si l'importance des pôles qui représentent les fréquences à maximum d'énergie est essentielle, il nous est apparu nécessaire d'exploiter le rôle des autres pôles pour équilibrer la répartition spectrale, en particulier dans le cas des fricatives.

En fait, nous n'avons pas abordé tous les aspects du système, une poursuite concerne l'extraction des informations d'ordre prosodique afin de mieux garantir le naturel de la voix. Au stage actuel de l'étude il n'est pas encore possible de se prononcer sur la validité de cette approche comparée à des approches plus classiques.

BIBLIOGRAPHIE

L.RABINER : "System for Automatic formant analysis of voiced speech". JASA, Vol. 47, pp 634-648, 1970.

X.RODET : "Analyse du signal vocal dans sa représentation amplitude-temps et synthèse de la parole par règles". Thèse d'Etat, Université de Paris VI, 1977.

J.SAMAKE : Rapport de DEA, Université de Nancy I, 1977.

THEME I (c)

SYNTHESE DE LA PAROLE

Synthèse à partir de textes

10èmes JOURNÉES D'ÉTUDE SUR LA PAROLE
GRENOBLE - 30 MAI - 1ER JUIN 1979

Pour une meilleure formalisation de la conversion automatique graphème-phonème.

S Nina Catach, V. Meissonnier
Domiciliation : CNRS - HESO, 27, rue Paul Bert 94200 IVRY sur Seine
CNRS - LISH, 54 Boulevard Raspail 75007 PARIS

Résumé - Ce que l'on appelle jusqu'ici "synthèse par règles" du français est trop souvent constitué d'un mélange d'unités de taille différente, ne formant pas système, de "règles" phonétiques et graphiques et d'approches empiriques peu ou mal délimitées. Les auteurs pensent qu'une systématisation des correspondances grapho-phoniques, fondée en particulier sur une étude statistique du rendement des règles, une hiérarchisation fonctionnelle des équivalences et des lois de position en contexte proche, est possible et nécessaire. Elles peuvent non seulement améliorer les résultats de conversion, mais apporter une lumière nouvelle sur la phonétisation elle-même. Pour cela il faut que les correspondances soient au maximum regroupées suivant des critères clairs, en règles de réécriture stables. Plus les unités seront petites et formant système, plus leur rendement sera étendu et leur accès facile. Cependant, il reste indispensable de travailler à plusieurs niveaux : niveau d'une table d'équivalences de base (T.E.B.), niveau des règles (R.) et des lois de position (L.D.P.), mais aussi niveau syllabique (SYLL.), niveau morphologique étroit (MORPH, Liaisons), niveau lexical (mots les plus fréquents codés, exceptions, EXC.) et parfois même syntaxique (SYNT.,pour les ambiguïtés). Le niveau syllabique revêt parfois une importance particulière, et on y insistera ici. Un exemple est donné de formalisation des règles dans un cas particulièrement difficile : celui du E.

Pour une meilleure formalisation de la conversion automatique graphème-phonème

s : Nina Catach, V. Meissonnier

SUMMARY What is generally called "synthesis by rules" of the French language, in the framework of the grapheme-phoneme automatic conversion, is too often a miscellaneous set of units of different lenght, that cannot make up a system, and of phonetic and graphic empirical "rules", little or ill defined. The authors think that a systematization of the graphemic-phonetic correlations, principally based on a statistical study of the efficiency of the rules, a functional ranking of the equivalences and of the position rules in a near-context, is possible and necessary. They can not only improve the results of conversion, but also bring a new insight into the phonetization itself. In order to do so, the correlations should be classified into stable rewriting rules, according to clear criteria. The smaller the units will be and the more integrated into a system, the broader will be their efficiency and the easier their apprehension. It is nevertheless necessary to work at several levels : 1- basic equivalences table (T.E.B.) ; 2- rules (R) and position laws (L.D.P.) ; 3- syllabic level (SYLL) ; 4- narrow morphological level (MORPH, liaisons) ; 5- lexical leval (the most frequent coded words, and exceptions,EXC.) ; 6- sometimes even syntactic level (SYNT, for ambiguities). After having presented the importance of such a formalization and hierarchization of the rules, the authors described shortly their own view of a program of automatic phonetization, particularly insisting on the interest of a previous stage of automatic syllabization. This syllabization allows to better circumscribe the graphemes and their phonetic correspondance. Four examples are given : the groups consonant+liquid, the groups I, U, OU+vowel, the nasal vowels, the open and closed vowels. Then an example is given of formalization of rules in a particularly difficult case : that of E, which requires at least 9 previous rules and 3 general rules of conversion. The efficiency of the rules is particularly interesting if one takes as a starting point the phonic system with eleven vowels to be found in the south of France.

10èmes JOURNEES D'ETUDE SUR LA PAROLE

G R E N O B L E - 30 MAI - 1er JUIN 1979

Titre : Pour une meilleure formalisation de la conversion automatique graphème-phonème

Auteurs :Nina Catach, V. Meissonnier

La plupart du temps, ce que l'on appelle la "formalisation par règles" pour le passage graphème-phonème n'est pas faite de véritables "règles". Selon les cas, on transcrit intégralement des mots, ou des parties de mots, ou des graphèmes, ou des lettres, au fur et à mesure des découvertes et des erreurs constatées, le tout reposant sur deux postulats paradoxalement contradictoires :

- il n'y a pas de véritables "règles" générales possibles pour établir en français des équivalences simples entre graphèmes et phonèmes.
- n'importe quelle personne peut établir ces "règles" (il existe des dizaines de programmes de phonétisation) pourvu qu'il s'agisse de sa langue maternelle.

Or si chacun peut les connaître et les appliquer, c'est sans doute que certaines règles existent et elles mériteraient, par conséquent, d'être mieux analysées, et définies une fois pour toutes.

D'autre part, on pense que cela n'en vaut pas la peine, car les résultats enregistrent en définitive un faible pourcentage d'erreurs, lesquelles existeront de toutes façons : un gros effort se solderait par une diminution d'erreurs d'un à deux pour cent.

Nous voudrions montrer qu'à notre avis, un tel raisonnement est erroné :

1) Il existe pour la plupart des unités graphiques des règles d'équivalences stables, dont on peut constituer une Table(Table d'équivalence de base, T.E.B.).

2) Plus ces unités seront petites, stables et en petit nombre, plus leur portée sera grande et leur accès facile.

3) Les rectifications et les listes d'exceptions deviennent possibles dans la clarté et on ne risque pas de donner des ordres contradictoires.

4) Le programme peut être compris et utilisé par un grand nombre d'utilisateurs, ce qui n'est pas le cas actuellement.

5) A partir d'une bonne formalisation graphémologique faite pour le français, et seulement dans ce cas, on pourrait étendre cette formalisation à d'autres langues et tenter d'établir un réseau d'équivalences qui est théoriquement possible, au moins pour les phonèmes communs. La synthèse faite à partir de mots ou de lettres rend ces équivalences impossibles.

6) Enfin, une telle formalisation permettrait une véritable analyse linguistique des deux chaînes, et amènerait sans doute des découvertes plus générales sur les rapports entre la phonologie, la morphologie et la syntaxe, avec éventuellement des retombées pour l'amélioration du processus inverse, en cas de conversion phonème-graphème.

Nous donnerons par conséquent ici :
- un bref rappel de nos hypothèses de départ.
- une description succincte de notre projet de phonétisation automatique.
- quelques exemples de l'avantage d'une syllabisation préalable.
- un exemple de formalisation des règles dans un cas particulièrement difficile, et nécessitant obligatoirement un sous-programme, celui du E.

I HYPOTHESES DE DEPART

Nos recherches précédentes (N.C. 1973, 1974) nous avaient amenés à dégager une hiérarchisation statistique des unités graphiques (ou graphèmes) de plus fort rendement, dans les textes et en lexique,et nous avaient permis d'établir une table d'équivalences de base (ou T.E.B., voir document I), ordonnant ces unités en sous-ensembles formant système. Il était ainsi apparu que, contrairement à ce qu'on pensait, un très petit nombre d'éléments graphiques (que nous avons fixés à 33, appelés archi-graphèmes) présentaient en français un rendement couvrant 80 à 85% de toute chaîne parlée. Douze à quinze éléments supplémentaires entrant en combinaison ou en commutation avec eux suffisaient à assurer l'essentiel de leur fonctionnement en discours réel (lois de position, L.D.P.) Avec 100 à 150 repérages des éléments situés directement à gauche et surtout à droite des graphèmes, on pouvait déjà établir les bases de n'importe

quel programme de phonétisation.

Trois catégories de positions aident essentiellement le lecteur à s'y retrouver : la catégorie "voyelles", la catégorie "consonnes" et la catégorie "blancs". Il fallait donc utiliser ces trois catégories dans notre programme de phonétisation.

II DESCRIPTION SUCCINCTE DU PROGRAMME DE PHONETISATION

Dès le début, il nous est apparu que dix conditions étaient nécessaires pour améliorer les programmes actuels (N.C., 1977):

1) meilleure mise au point des répertoires de base, phonique et graphique ;
2) meilleure connaissance des règles d'assemblage ;
3) meilleure délimitation et meilleur déchiffrage des graphèmes ;
4) utilisation des lettres muettes finales comme moyen de lutte contre les ambiguïtés (codification automatique) ;
5) inventaire des cas de liaison les plus fréquents ;
6) syllabisation automatique à disposition ;
7) meilleure phonétisation ;
8) utilisation critique des listes de fréquence, choix sélectif de listes de base à plusieurs niveaux, avec formes fléchies codées ;
9) traitement du reliquat d'ambiguïtés à l'aide des détecteurs morphologiques (déterminants et pronoms) ;
10) seulement en cas d'échec, analyse syntaxique.

Nous avons donc conçu notre programme en un certain nombre de modules d'accès direct et permettant au noyau principal (la T.E.B.) de jouer pleinement son rôle :

I - Traitement des liaisons et programmes spéciaux.
II - Découpage syllabique
III - Suppression des fins de mots muettes
IV - Phonétisation
V - Comparaison des deux chaînes et comptages.

(voir plus loin la présentation du programme et les documents deV.Meissonnier)

III LA SYLLABISATION

Je voudrais insister tout particulièrement sur les aspects simplificateurs que peut présenter un découpage syllabique préalable du texte à transcrire.

Contrairement à ce qui se passe en anglais (coupures étymologiques et sémantiques) la syllabation graphique française peut s'automatiser assez facilement. Nous avons mis au point un programme de syllabisation qui fonctionne bien et qui se résume, mise à part une courte liste d'exceptions, en quatre à cinq règles.

Cette syllabisation préalable permet de traiter en séries et à un premier stade un certain nombre de problèmes généraux qui se posent en français en ce qui concerne la répartition des lettres entre les différentes unités graphiques : c'est en réalité à l'intérieur de la syllabe qu'en règle générale les graphèmes dont j'ai parlé plus haut existent, avec leurs correspondants stables à l'oral. J'en donnerai ici quatre exemples :

1. Les groupes consonne + liquide : Ils font toujours partie de la même syllabe. Conséquence phonique : apparition d'un i supplémentaire devant le yod de cri-er, pli-er, peupli-er. Problème réglé une fois pour toutes.

2. Les groupes I, U, OU + voyelle phonique : Ils font toujours partie de la même syllabe. Conséquence à l'oral : la voyelle se transforme en semi-voyelle : pie/pied, lu/lui, loup/louis, loi. Dans toutes les autres conditions (devant consonne ou blanc) on a les voyelles I, U, OU (piste, su, sou).

3. Les voyelles nasales : Nous avons 22 graphèmes de voyelles nasales. Cela vaut la peine de les traiter d'un coup. Soit $\hat{V}\hat{N}$ = voyelle nasale, $\hat{C}\hat{N}$ consonne nasale, V voyelle, C consonne, # séparation de syllabe, + suite de lettres. Deux positions se présentent : devant voyelle/ autres positions :

Règle I : devant voyelle :

$V+\hat{C}\hat{N}+V \rightarrow V\#\hat{C}\hat{N}+V$ ex. â/ne, â/nerie, â/non.

Règle II : autres positions (consonne ou blanc) :

$V+\widehat{CN} \rightarrow \widehat{VN}$ = ex. an, an/cre.

Reste le problème de la consonne double nasale, et quelques autres problèmes de préfixes, le tout pouvant aussi se régler en série.

4. Les voyelles ouvertes et voyelles fermées :Autre problème plus important encore peut-être, et, notons-le, presque entièrement dépendant de la structure de la syllabe : en règle générale, à moins d'être notées par un graphème spécifique (accents, ai, au en particulier), les voyelles françaises sont fermées en syllabe ouverte, ouvertes en syllabe fermée. Une syllabe est fermée par une consonne, phonique ou graphique. Une fois le texte syllabisé (et en dehors des finales de mots qui posent des problèmes particuliers), on aura donc en principe :

Règle I : devant consonne :

V+C→V ouverte ex. res/ter, por/ter

Règle II : autres positions (blanc, finale de mot) :

V → V fermée ex. au/to, au/to/mo/bile

Bien sûr il y a quelques problèmes, en particulier celui de la consonne (ou de la double consonne) suivie de e caduc, qui se ramène à la Règle n° I :

V(+C)+C+e caduc → V+C ex. note, botte, abreuve, presse.

Rentabilité de la règle : 61% pour E ouvert, 75% pour 0,93% pour EU, ces trois voyelles couvrant environ 6% de l'ensemble des phonèmes du français. Cependant, pour E, comme on sait, de terribles problèmes se posent, en particulier à la finale, dont nous allons dire un mot ci-dessous.

Auparavant, nous noterons précisément qu'un autre avantage du découpage syllabique est de pouvoir traiter spécifiquement de la dernière syllabe du mot (lettres muettes) et de la première syllabe du mot (préfixes) deux situations qui peuvent modifier considérablement les règles de conversion générales.

Une remarque au passage : nous nous faciliterons beaucoup la tâche si nous décidions une fois pour toutes de procéder à la synthèse de la parole en nous fondant sur le système phonique à onze voyelles du sud de la France, (qui ne connaît pratiquement pas d'exceptions aux alternances combinatoires à l'intérieur de la syllabe)et non sur le système entièrement dépassé, à notre avis, de 16 voyelles décrit dans les manuels.

IV UN EXEMPLE DE FORMALISATION PARTICULIEREMENT DIFFICILE : LE CAS DU E

Prenons le E, qui, lorsqu'il n'est pas accentué, peut correspondre à e fermé (dans nez) e ouvert (dans net) ou e caduc ou instable (fenêtre), sans parler du e entièrement graphique de lycée, émue ou prie.

On ne me croirait pas si je disais qu'avec ma T.E.B., je résous tous les problèmes des voyelles notées par E, qui sont au coeur de tout le fonctionnement phonique et graphique du français.

On peut cependant, même dans ce cas extrême, tenter d'ordonner les filtres nécessaires à la mise au clair des lois générales (document II).

Ces filtres sont au moins au nombre de neuf, et sont suivis de trois lois générales.

I . Détermination et mise en mémoire des graphèmes à distinguer de E (EI, EIN, EAU, etc.) ;

II. Traitement des liaisons (en ES) ;

III. Traitement de certains monosyllabes (mes, tes, ses, etc.) ;

IV. Traitement des exceptions (noms propres, emprunts, mots aberrants, groupes que, gue, ge, etc.) ;

V . Syllabisation (valable pour l'ensemble des textes) ;

VI. Traitement des finales (voyelle E, voyelle ES, E tréma, ES tréma, E (cons.) cons. E, ESSE, groupes de consonnes finales, finales en ENT, etc.);

VII. Traitement des syllabes initiales (préfixes E-, DE-, PRE-, RE-, groupe cons. ESS-, etc.) ;

VIII. Traitement de EN (préfixes EN-, EM-, groupes ENN-, EMM-, etc.) ;

IX. Traitement des cas spéciaux (EILL, EIL, EY, AY, etc.).

Ces neuf filtres permettent d'établir ensuite une nette répartition des différents E :

X . E accentué (aigu, grave, circonflexe, tréma) ;

XI. E consonne→E ouvert (règle syllabique) ;

XII. E caduc (toujours en fin de syllabe ou en fin de mot, avec règles correspondantes).

Pour finir, je voudrais dire que le passage graphème-phonème en vue de la synthèse de la parole n'est qu'une des applications possibles d'une étude systématique des correspondances des deux chaînes. Il s'agit de deux canaux de communication universels sinon autonomes ou différents, du moins

distincts. La chaîne graphique connaît en somme deux sortes de lois : ses lois de correspondance avec la chaîne phonique et ses lois propres. La solution de nos problèmes se trouvent à l'intersection des deux.

C'est important non seulement pour les linguistes et les acousticiens, mais aussi pour les psychologues, les pédagogues, les sociologues et en général tous ceux qui s'intéressent aux problèmes de lecture et de communication, y compris, évidemment, aux problèmes de télécommunication. Cela vaut la peine, par conséquent, de tenter de dépasser les procédés plus ou moins empiriques utilisés jusqu'à présent.

V PRESENTATION DU PROGRAMME

(V. Meissonnier)

Le travail que Mme Catach souhaitait automatiser concernait la phonétisation du Français. Elle envisageait pour cela quatre phases.

1) Codification des fins de mots muettes : par rapport aux règles générales de phonétisation, les fins de mots constituent une classe de traitements particuliers. Les finales muettes sont supprimées :

Exemple : radis, oasis) le "s" final (ne se prononce pas / se prononce.

2) Traitement des liaisons entre les mots avec une syntaxe simple : en utilisant des listes de mots avec leurs catégories syntaxiques, un ensemble réduit de règles syntaxiques s'appliquant à des suites de 2 ou 3 occurrences peuvent permettre de déterminer s'il y a ou non liaison :

Exemple : les enfants
(liaison)

La règle (article + syntagme nominal) rend la liaison obligatoire.

3) Syllabisation des mots : c'est à partir de la suite des lettres qui constituent un mot, la mise au point d'un ensemble de règles permettant leur regroupement en syllabes :

Exemple : TRUAND ⟶ TRU AND
ARTISTE ⟶ AR TIS TE

4) Phonétisation : à partir des résultats des trois étapes précédentes, par un ensemble de règles (chaîne graphique → chaîne phonétique), c'est la production de la chaîne phonétique à partir de la chaîne écrite.

Exemple : 'EAU' → donne le phonème 'o'.

Chaque phase doit fournir des informations linguistiques sur la structure de l'orthographe dans son interaction avec la prononciation.

L'étude de ce programme de travail ne pouvait se faire sans prendre connaissance de ce qui se fait ailleurs, en particulier au CNET et au LIMSI. En collaboration avec M. Prouts (LIMSI), nous avons réalisé un programme d'édition des règles de phonétisation, un programme de calcul des fréquences d'utilisation de ces règles. Un module de comptage des phonèmes est en cours de réalisation (PHONSTAT).

Le programme de syllabisation des mots est réalisé. Il fonctionne bien aussi dans une version simple qu'avec un module de comptage (SYLLAB, SYLSTAT).

Les étapes 1) et 2) nous ont semblé moins prioritaires. Pour avoir sur le texte, des informations relatives à ces étapes, nous avons réalisé un programme de comptage qui fonctionne à partir de textes déjà codés manuellement (Documents III, IV, V, VI et annexes).

Ouvrages cités

N. Catach, 1973, "La structure graphique du français", La Recherche, n° 39; 1974, Table ronde internationale CNRS sur la structure de l'orthographe française, Klincksieck ; 1977, "Conversion automatique graphème-phonème, directions de recherche et résultats", Ambiguïtés de la langue écrite, CNRS, Paris.

10èmes JOURNÉES D'ÉTUDE SUR LA PAROLE

GRENOBLE - 30 MAI - 1ER JUIN 1979

LA PONCTUATION, INDICATEUR PROSODIQUE POUR LA SYNTHESE A PARTIR DU TEXTE : ETUDE DE LA VIRGULE.

Christine CHOPPY

8, Résidence de Chevreuse
91 400 ORSAY

RESUME

La ponctuation, qui est un élément important pour tout traitement de textes fournit des indications prosodiques pour la synthèse vocale à partir du texte. Le cas de la virgule est complexe car elle est employée dans des contextes grammaticaux variés, et important car c'est la ponctuation la plus fréquente.

La virgule est généralement associée à une montée du Fo et à une pause ; cette étude met en évidence d'autres schémas prosodiques tels que : chute du Fo et pas ou peu de pause. Pour les propositions incises nous observons une Fo inférieure au reste de la phrase. Nous nous interrogeons d'autre part sur la valeur des différentes virgules à l'aide d'une expérience de ponctuation.

Selon les types d'application, il peut être souhaitable d'incorporer ces nouveaux éléments de prosodie pour le traitement de la parole par ordinateur.

PROSODIC FEATURES CONNECTED WITH PUNCTUATIONS FOR SPEECH SYNTHESIS : SOME CASES WITH COMMAS

Christine CHOPPY

SUMMARY

Punctuation is of primary interst for natural language processing, and in particular, it provides cues for prosodics in automatic speech synthesis. Commas are both the most frequent punctuation and a very complex one since they are used in various grammatical contexts. However, commas are generally associated with a preceeding Fo rise and a pause. Investigating further, we found cases where commas are associated with a Fo fall and no pause (or a very short one) : those commas are located before an additional comment at the end of a sentence (like in : "You can just come, if you feel like it."), or at "emphatic pronouns" (which are frequent in idiomatic French). We also found (1) a lower Fo (Fig. 1, 2, 3) on incidental clauses (clauses inserted in sentences where one is reporting someone else's words, e. g. : "I suggest we have a party, Jacqueline said, and invite all our friends."), and (2) that incidental clauses are preceded by a short pause and followed by a long pause.

To investigate the meaning of those commas, we asked subjects to fill in commas in a corpus where commas had been withdrawn : commas were found to be perceived as more important at clause boundaries than at emphatic pronouns.

These results may be useful in speech processing and in particular in speech synthesis e. g. as used in a reading machine for the blinds.

10èmes JOURNÉES D'ÉTUDE SUR LA PAROLE
GRENOBLE - 30 MAI - 1er JUIN 1979

LA PONCTUATION, INDICATEUR PROSODIQUE POUR LA SYNTHESE A PARTIR DU TEXTE : ETUDE DE LA VIRGULE.

Christine CHOPPY
8, Résidence de Chevreuse
91 400 - ORSAY

INTRODUCTION

Dans les travaux de synthèse de la parole à partir du texte, la ponctuation est une source d'information facilement accessible concernant la syntaxe d'une phrase. Elle est également utile pour tirer des indications sur la prosodie souhaitable pour la synthèse vocale du texte entré sur ordinateur ; en effet, si la ponctuation est un phénomène de l'écrit et non de l'oral, les exercices de lecture à haute voix ou de prise sous la dictée nous conduisent à établir des correspondances entre la ponctuation et sa réalisation orale sous forme de variations prosodiques.

Si le point a généralement le rôle de terminer une phrase, la virgule a des rôles multiples (M. GREVISSE, 1969)

- dans une proposition : séparation d'éléments semblables, d'un élément à valeur explicative, ou pour isoler des mots formant pléonasme ou répétition (ex. : pronoms de renforcement ou d'emphase (C. BLANCHE-BENVENSITE, 1973))...

- dans un groupe de propositions : pour séparer des propositions de même nature ou introduites par des conjonctions.

Des études sur le roman contemporain (C. GRUAZ, 1977 ; L. PASQUES, 1977) montrent que la virgule est la ponctuation la plus fréquente. On peut toutefois remarquer que l'emploi de ponctuations varie beaucoup individuellement.

Les indications prosodiques admises couramment en synthèse de la parole (J. VAISSIERE, 197. ; F. EMERARD, 1977 ; C. CHOPPY, 1977) sont, respectivement, que le point est précédé d'une chute marquée de Fo (fréquence du fondamental), d'un allongement de durée sur la dernière voyelle et est noté par une pause, et que la virgule est précédée d'une montée de Fo , d'un allongement de durée sur la dernière voyelle, et est notée par une pause moins longue que pour le point. Un seul schéma prosodique, donc, pour la virgule, alors qu'il s'agit d'une ponctuation à la fois fréquente et complexe.

Dans cette première étude nous nous sommes proposés de chercher s'il existait d'autres schémas, d'autres indications prosodiques fournies par la virgule et si une correspondance pouvait être établie entre ces nouvelles indications et des rôles particuliers de la virgule. Notre étude est donc guidée, non pas par la description exhaustive des réalisations prosodiques de la virgule selon ses rôles, mais par la recherche de nouveaux schémas prosodiques.

Une étude préliminaire nous a permis d'observer de nouveaux schémas : (i) des virgules précédées d'une chute de Fo, et, soit marquées par une pause très courtes, soit non marquées par une pause, (ii) pour les propositions incises dans une phrase, une Fo plus basse sur ces incises. Nous avons alors étudié ces phénomènes de façon systématique pour dix locuteurs, hommes et femmes, qui ont lu un même texte à voix haute. Les locuteurs ont été enregistrés en chambre insonorisée. Les paramètres prosodiques - Fo, durée et pause - ont été mesurés à l'aide du programme FPRD (W. HENKE, 1976) sur ordinateur PDP9.

I - ROLE DE LA PONCTUATION.

Nous avons observé que les virgules associées au schéma prosodique (i) séparent (a) des propositions en fin de phrase qui jouent le rôle d'un commentaire additionnel (ex. : "Il est dans son bureau, je pense."), qui ne changent pas le sens de la phrase et véhiculent peu d'information, (b) des pronoms de renforcement ou d'emphase (ex. : "On n'en sait rien, nous,....").

On peut s'interroger sur la valeur, la signification de ces virgules qui sont réalisées selon un schéma prosodique différent. Nous avons cherché quel était l'usage de ces virgules par rapport à d'autres.

Nous avons procédé comme suit : dans un premier temps, le texte initial ponctué est dépouillé de ses virgules et présenté aux sujets ; ceux-ci le lisent, ajoutent des virgules là où cela leur semble nécessaire, puis sont enregistrés en lecture à voix haute. Après un délai minimum d'une semaine, les sujets sont à nouveau enregistrés pour la lecture du texte initial non dépouillé de ses virgules, c'est-à-dire présentant les même ponctuations virgules pour tous (les résultats présentés dans les paragraphes suivants concernent ce second enregistrement).

Comparés au texte initial, les textes ponctués par les sujets comportent :
- les mêmes virgules aux frontières de propositions dans 84 % des cas
- les mêmes virgules séparant un pronom d'emphase dans 40 % des cas.

Si la nécessité d'une virgule s'impose largement dans le premier cas, l'usage en est donc facultatif dans le second (le niveau d'études des sujets est au moins la licence). On peut détailler ce résultat en ajoutant que certains sujets mettent peu de virgules dans l'ensemble, d'autres beaucoup plus, et que quelques uns souhaitaient rajouter d'autres virgules lors de la lecture à voix haute. La comparaison des variations prosodiques entre le premier et le second enregistrement ne permet pas de conclure, (1) à une différence de traitement prosodique systématique quand une virgule, absente dans le premier cas, est présente dans le second, (2) à un traitement prosodique rigoureusement identique quand une virgule est présente dans les deux cas. C'est-à-dire que (1) par exemple, si la virgule séparant le pronom emphatique du reste de la phrase n' a pas été inscrite par le sujet, il peut toutefois dans sa lecture à voix haute marquer une pause, (2) deux lectures d'un même texte peuvent donner lieu à des réalisations prosodiques différentes.

Il ne s'agit donc pas en matière de prosodie de donner de règles absolues, nous nous contentons de chercher les tendances généralement adoptées.Nous décrivons ci-après les tendances que nous avons observées : pour les commentaires en fin de phrase, pour les pronoms d'emphase, pour les propositions incises (C. CHOPPY, 1978).

II - COMMENTAIRES EN FIN DE PHRASE

Nous appelons "commentaire en fin de phrase" des propositions telles que : "dit Marie-Laure" dans : "C'est ça, dit Marie-Laure.", ou "si tu sais" dans "dis-le nous ce que tu mettrais, si tu sais". Ce sont des éléments ajoutés en fin de phrase, qui n'en changement pas le sens et véhiculent peu d'information Le corpus comprend huit phrases avec commentaires finaux.

La virgule précédant le commentaire n'a pas été marquée par une pause dans 65 % des cas ; lorsqu'il y a eu pause, sa durée moyenne est de 200 ms, c'est-à-dire qu'il s' agit d'une pause courte. Dans 80 % des cas, le commentaire final est précédé d'une chute Fo (en moyenne 30 Hz). La durée de la dernière voyelle précédant le commentaire est allongée (120 à 130 ms).

Le schéma prosodique qu'on peut associer à la virgule précédant le commenaire final est donc : pas ou peu de pause, chute de Fo, allongement de la voyelle précédente.Dans l'expérience de ponctuation décrite au § I ce type de virgule a été rajoutée dans 82 % des cas.

III - PRONOMS D'EMPHASE OU DE RENFORCEMENT

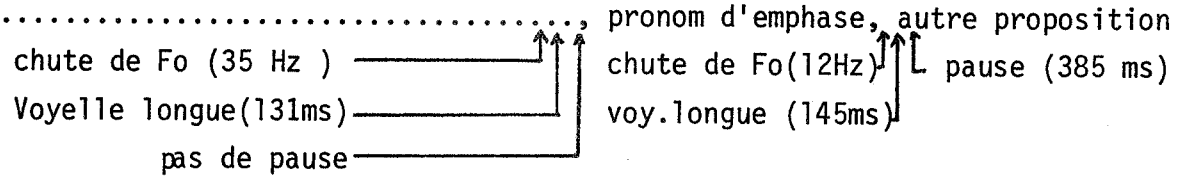
Les pronoms d'emphase (C. BLANCHE-BENVENISTE, 1973) sont une tournure redondante, un "recopiage" d'un nom ou d'un pronom, par exemple :
"Il est à l'heure." devient "Il est à l'heure, lui." (a)
ou "Lui, il est à l'heure." (b)

Nous parlerons d'emphase à DROITE (a) ou à GAUCHE (b) selon que le pronom est "recopié" à droite ou a gauche. Dix phrases du corpus comportent des cas d'emphase à droite ou à gauche, portant sur des pronoms.

Pour l'emphase à gauche, il y a chute de Fo sur le pronom dans 67 % des cas (56 Hz en moyenne), la voyelle finale est longue (158 ms en moyenne), il n'y a pas de pause après le pronom dans 57 % des cas.

Dans des cas d'emphase à droite, il y a aussi chute de Fo sur le pronom. Pour la phrase (a) cela correspond au schéma prosodique de fin de phrase ; mais on observe aussi une chute de Fo pour des phrases telles que :
"On n'en sait rien, nous, et il faut bien qu'on réponde." (c) où le pronom d'emphase est suivi d'une autre proposition (chute de Fo dans 63 % des cas - 12 Hz en moyenne -, durée moyenne de la dernière voyelle : 145 ms, pause après le pronom - ce qui est aussi une frontière de propositions - assez longue : 385 ms en moyenne). La chute de Fo n'est pas très importante (12 Hz), mais il y a également chute de Fo sur le mot précédant le pronom d'emphase ("heure" (a), "rien" (c)) dans 86 % des cas (35 Hz en moyenne). Il n'y a pas de pause avant le pronom d'emphase dans 69 % des cas.

SCHEMA POUR L'EMPHASE A DROITE



IV - PROPOSITIONS INCISES

Il s'agit de propositions indiquant qu'on rapporte les propos de quelqu'un, et qui sont insérées dans le corps de la phrase pour le cas où elles sont rejetées à la fin de la phrase, voir § II, par exemple :
"On pourrait peut-être, proposa quelqu'un, déplacer le canapé dans l'autre pièce."

Il y a dans le corpus cinq phrases comportant des propositions incises ; phrases et incises sont de longueur variée.

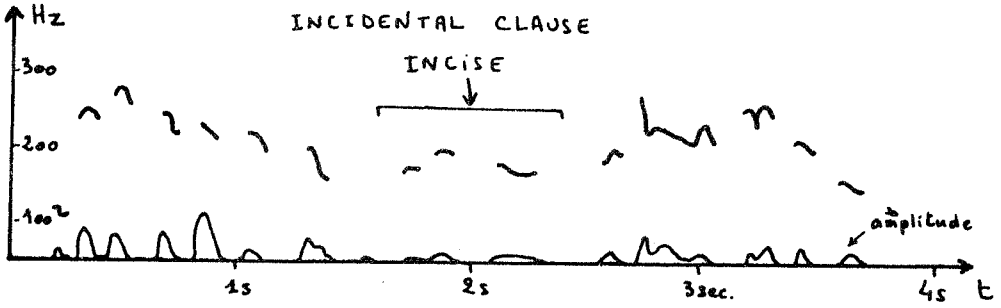
Nous observons sur le tracé de Fo en fonction du temps (fig. 1) que les valeurs de Fo sont inférieures sur l'incise par rapport au reste de la phrase. Pour mesurer ce phénomène nous avons procédé de la façon suivante :
- pour chaque phrase, pour chaque voyelle, nous avons relevé la Fo la plus élevée
- puis nous avons calculé les FO moyennes pour les voyelles des portions de phrase situées avant l'incise, après l'incise, et pour l'incise elle-même.

Dans la majorité des cas (88 %), la Fo moyenne est plus élevée après l'incise que sur l'incise (fig. 2 et 3). La Fo moyenne avant l'incise étant supérieure à la Fo moyenne après l'incise, nous observons donc une Fo décroissant sur l'ensemble de la phrase, mais passant par un palier plus bas sur l'incise.

Les virgules encadrant l'incise sont pour certaines phrases précédées d'une montée de Fo et pour d'autres d'une chute de Fo. En ce qui concerne les pauses éventuelles marquant les virgules qui entourent l'incise : on observe dans 76 % des cas une pause courte (132 ms en moyenne) avant l'incise, et dans 92 % des cas une pause significativement (test t : $p \leq .01$ ou $P < .05$ selon les phrases) plus longue (454 ms en moyenne) après l'incise (fig. 4).

Enfin nous avons testé pour deux locuteurs l'hypothèse selon laquelle l'incise serait dite plus vite (c'est-à-dire à une vitesse d'articulation plus élevée) que le reste de la phrase : nous n'avons pas observé une telle tendance.

Dans l'expérience de ponctuation décrite au § I, les virgules avant incise ont été rajoutées dans 83 % des cas, et les virgules après incise, dans 91 % des cas.



"Je propose qu'on fasse une fête, dit Jacqueline, et qu'on invite tous les copains." - Fo inférieure sur l'incise -

Fig 1.

Fig 2. "I suggest we have a party, Jacqueline said, and invite all our friends." - Lower Fo on the incidental clause -

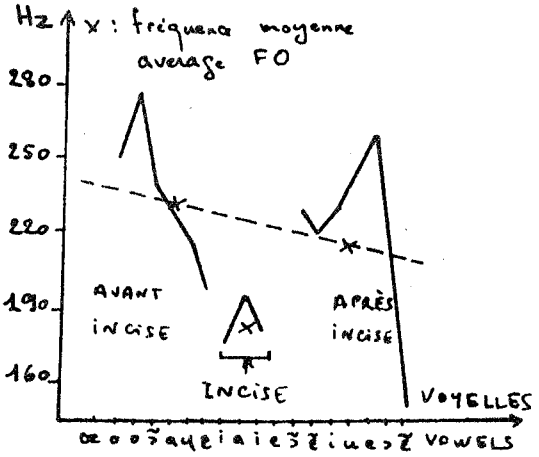
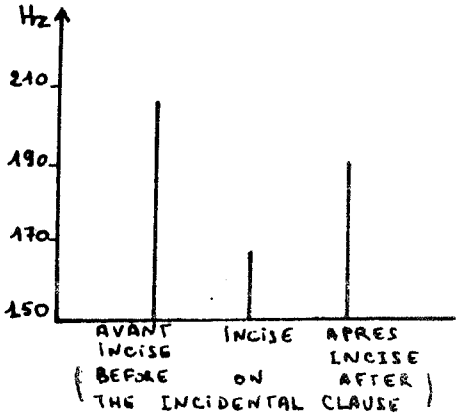


Fig 3. La Fo moyenne est inférieure sur l'incise
The average Fo is lower on the incidental clause



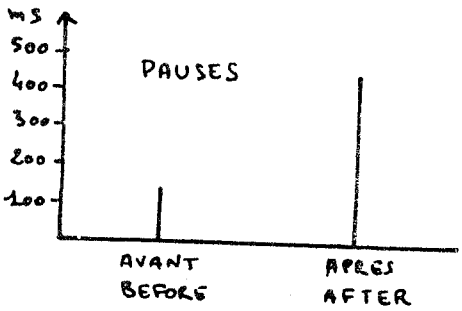


Fig 4. Pauses plus courtes avant incise qu'après.
Pauses are shorter before the incidental clause.

V-CONCLUSION -

Nous avons décrit, dans les cas de commentaires en fin de phrase et de pronoms d'emphase, un schéma prosodique associé à la virgule qui est : chute de Fo, voyelle longue, pas ou peu de pause. L'expérience de ponctuation(§ I) nous montre qu'il s'agit d'une virgule"importante" dans le cas de commentaires finaux, et "facultative" dans le cas des pronoms d'emphase (toutefois, pour l'emphase à droite suivie d'une autre proposition, la seconde virgule est importante et est marquée par une pause).

D'autre part, dans le cas des propositions incises, si l'on n'a pu associer aux virgules qui les délimitent une chute ou une montée de Fo, nous avons mis en évidence une Fo inférieure sur ces propositions ; elles sont précédées d'une pause courte et suivies d'une pause longue ; l'expérience de ponctuation montre que les virgules encadrant l'incise sont importantes.

Ces résultats sont directement utilisables en synthèse de la parole dans le cadre d'applications à large spectre du ,type machine à lire pour les aveugles. Ils peuvent également être un élément du traitement prosodique en reconnaissance de la parole.

REFERENCES

BLANCHE-BENVENISTE, C., 1973; Recherches en vue d'une théorie de la grammaire Essai d'application à la syntaxe des pronoms, Thèse, Paris Librairie Honoré Champion.

CHOPPY, C., 1977, Introduction de la prosodie dans la synthèse vocale automamatique. Thèse de Docteur Ingénieur Paris

CHOPPY, C., 1978, Rapport d'activité I.R.I.A. N° 3

EMERARD, F., 1977, Synthèse par diphones et traitement de la prosadie, Thèse 3ème cycle GRENOBLE

GREVISSE, M., 1969, Le Bon Usage, Ed Duculot

GRUAZ, C., 1977, Fréquence d'emploi des signes de ponctuation dans cinq romans contemporains, in : La Ponctuation, Recherches actuelles et Historiques, CNRS, Publication du G.T.M.

HENKE, W., 1976, Fundamental Période, Note interne du Research Laboratory of Electronics, M.I.T., Cambridge MA, U.S.A.

PASQUES, L., 1977 Ponctuation à l'écrit, arrangement rythmique à l'oral, à propos d'un conte de Marcel Jouhandeau lu par l'auteur, in: La Ponctuation (réf. ci-dessus)

VAISSIERE J., 1971 Contribution à la synthèse par règles du français, Thèse 3ème cycle, Grenoble.

10èmes JOURNÉES D'ÉTUDE SUR LA PAROLE

GRENOBLE - 30 MAI - 1ER JUIN 1979

Effets de la prosodie de phrase sur les variations du F_0 et de la durée syllabique.

Emanuela CRESTI - Franco MARTORANA - Mario VAYRA - Cinzia AVESANI

Scuola Normale Superiore, Piazza dei Cavalieri 7 - 56I00 Pisa
Italia

RESUME

A fin de vérifier l'existence des patterns d'intonation de phrase, tels que Topic-Comment et Comment, on a analysé deux phrases de type déclaratif et présentatif (réalisées deux fois par six locuteurs) par rapport aux paramètres du F_0 et de durée.

La segmentation des phrases a été exécutée en visant à la syllabe phonétique.

On donne aussi des résultats pour ce qui concerne soit les différences de durée qui caractérisent les deux patterns d'intonation, soit les variations de durée syllabique liées à des positions specifiques dans le "bloc tonal". On propose aussi, par égard au F_0, des stylisations possibles des deux patterns proposés.

Effects of the sentence prosody on the variations of the F_0 and of the syllabic length.

Emanuela CRESTI - Franco MARTORANA - Mario VAYRA - Cinzia AVESANI

SUMMARY

According our hypothesis of two intonational patterns of the sentence, like Topic-Comment and Comment, we analysed the F_0 and the syllabic lenght of two sentences, the one declarative and the other presentative, twice performed by six speakers.

The segmentation of the two sentences has been accomplished from the point of view of the phonetic syllable.

We present some issues both on the different length which caracterise the two intonational patterns and the syllabic length variations related to particular positions inside the tonal block.

We propose also some possible stylizations of the F_0 of the two hypothesized patterns.

10èmes JOURNÉES D'ÉTUDE SUR LA PAROLE
GRENOBLE - 30 MAI - 1ER JUIN 1979

Effets de la prosodie de phrase sur les variations du F_0 et de la durée syllabique.

Emanuela CRESTI - Franco MARTORANA - Mario VAYRA - Cinzia AVESANI

La présente recherche s'encadre dans un projet plus général de repérage, description et stylisation des contours de F_0 (patterns) et de structuration rythmique de la phrase italienne, qui sera porté à terme par un système d'analyse-synthèse actuellement en passe de mise au point.

Nous partons d'une hypothèse d'existence des contours généraux de la phrase, dont la forme soit typique et significative, qui interprètent des structures syntaxique superficielles quantifiées et indexées par la forme logique. Nous prémettons d'abord que nous opérons dans le cadre syntaxique de la grammaire générative-transformationnelle (cf. Cresti I979, à paraître).

Nous soutenons, en particulier, que la phrase simple, affirmative, est interprétée par un pattern Topic-Comment (T-C) composé de deux "blocs tonals" ('t Hart-Collier, I975) qui sont en correspondance avec les deux catégories majeures qui composent la phrase: SN et SV, directement dominés par le noyau de la phrase (cf. fig. I et 4).

Une phrase présentative du genre Piove a dirotto (Il pleut à verse) ou E' un quarto alle otto (C'est huit heures moins quart) ou bien encore E' arrivato Luigi (Louis est arrivé), qui soit naturellement prononcée sans aucune emphase ni contraste, s'interprète par un pattern Comment (C) du type de fig. 2 et 5.

Néanmoins, toute phrase dans laquelle on peut supprimer le SN faisant fonction de Topic, ou le transposer, acquiert une valeur présentative et pourra s'interpréter par le pattern Comment de fig. 3 et 5.

Nous avons relevé, au cours des mensurations des deux phrases: I) Lucia ha dormito in cucina (Lucie a dormi dans la cuisine), et 3) Ha dormito in cucina (Il (elle) a dormi dans la cuisine), prononcées deux fois par six locuteurs différents, trois hommes et trois femmes, d'une part que les contours gardent leur regularité, ce qui confirme l'existence des patterns proposés; de l'autre, que la séquence syllabique Ha dormito in cucina, interprétée en "blocs tonals" différents, a les caractéristiques de F_0, une durée et une énergie très differentes. Ceci prouve qu'en ce qui concerne l'intonation les caractéristiques du pattern de la phrase sont dominantes par rapport aux phénomènes de micro-mélodie, liés à la composition phonétique des mots en jeu.

ANALYSE DU PROBLEME

En examinant les différences entre les deux patterns par rapport aux paramètres du F_0 et de durée, mesurée par égard aux syllabes, nous avons pu observer les phénomènes suivants:

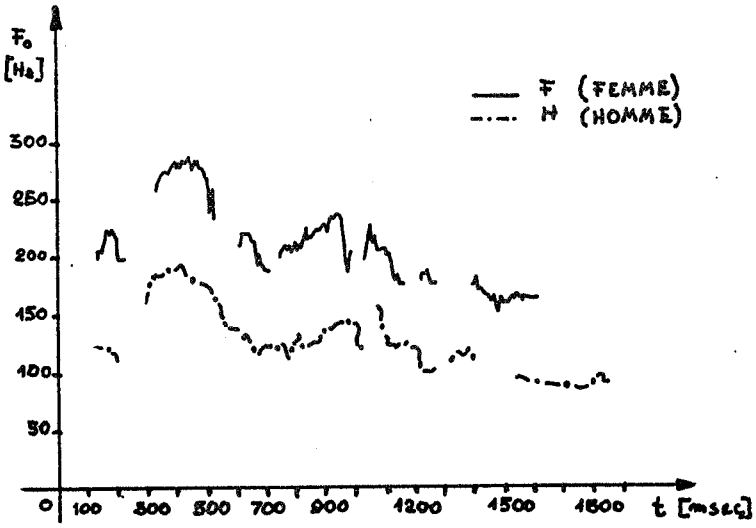


Fig. I - Frase T-C "Lucia ha dormito in cucina"
Profilo di F_0
Phrase T-C - Contour du F_0
T-C sentence - F_0 contour.

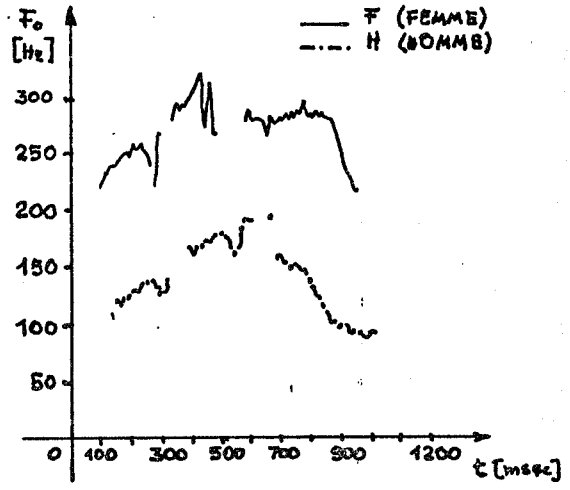


Fig. 2 - Frase C "E' un quarto alle otto"
Profilo di F_0
Phrase C - Contour du F_0
C sentence - F_0 contour.

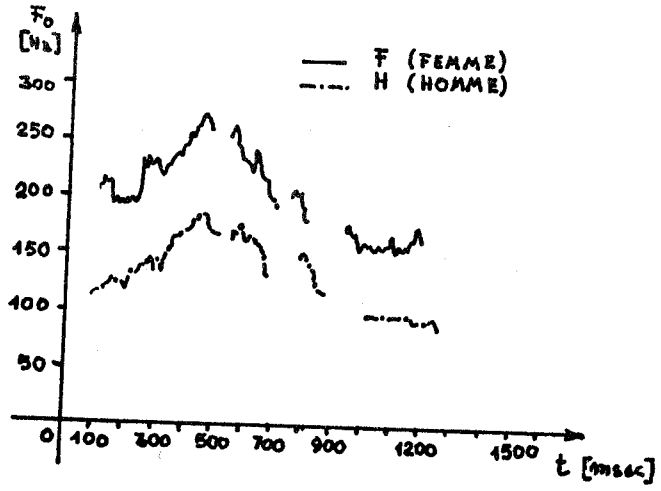


Fig. 3 -Frase C "Ha dormito in cucina"
Profilo di F_o
Phrase C - Contour du F_o
C sentence - F_o contour

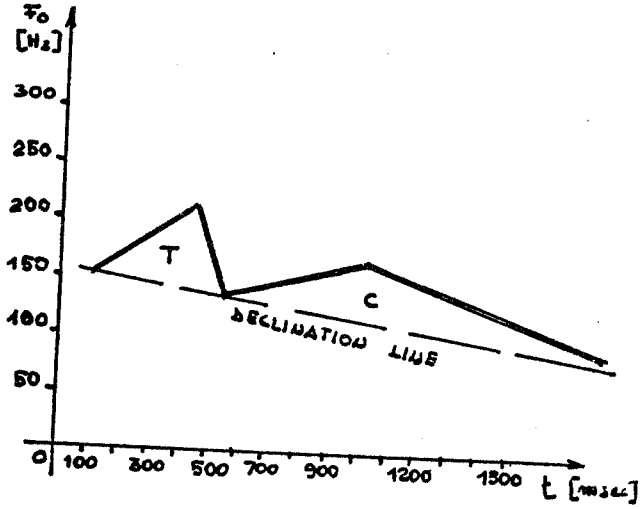


Fig. 4 -Stilizzazione di una frase T-C
Stylisation d'une phrase T-C
T-C sentence stylisation

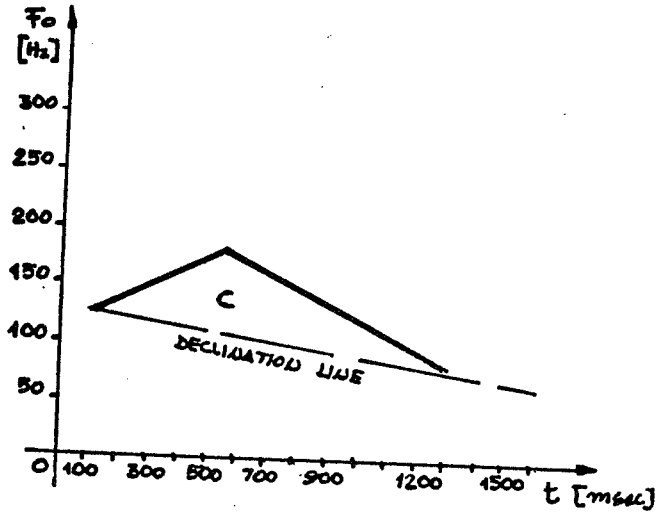


Fig. 5 -Stilizzazione di una frase C
Stylisation d'une phrase C
C sentence stylisation

F_0-Topic-Comment (T-C)

Topic: Nous avons choisi le topic du notre exemple avec un seul mot lexical. La crête d'onde se trouve en correspondence de la syllabe tonique du SN, et l'on constate ensuite une descente rapide.

Comment: La crête se trouve en correspondance de la syllabe tonique du V et puis on constate une descente lente et continue jusqu'à la fin de la phrase.

Comment (C)

Comment: La crête se trouve en correspondance de la syllabe tonique du V, le F_0 tient, ensuite il y a une retombée rapide en correspondence de la syllabe tonique du dernier SN.

Durée-Topic-Comment (T-C)

Topic: On constate un allongement de la dernière syllabe du topic, indépendamment de son accentuation.

Comment: On constate une tendance à transformer en une vocoïde de timbre imprécis (schwa) la dernière voyelle du comment et à former, par conséquent, une nouvelle syllabe qui se compose de l'avant dernière syllabe (en général tonique, en italien) et de la consonne finale jusqu'à former une syllabe longue.

Comment (C)

Comment: On observe une tendance à l'isochronisme syllabique, à l'exception de ce qui touche à la dernière syllabe, qui se soumet au processus précédemment indiqué.

Nous avons fait des essais avec les phrases 4) Il ragazzo ha dormito in cucina (Le garçon a dormi dans la cuisine) et 5) La modista ha dormito in cucina (La modiste a dormi dans la cuisine), dans lesquelles le premier SN, interprété comme topic, a une structure accentuelle semblable à celle de la séquence Ha dormito (Il (elle) a dormi), pour vérifier si la diversité des paramètres enregistrès pour les deux patterns ne dépendait pas de la seule différence d'accentuation; mais la forme du pattern Comment (C) se distingue quand-même de celle du topic d'un pattern Topic-Comment (T-C), même lorsque la structure accentuelle de la séquence syllabique et semblable. Il n'apparaît même pas que la forme du pattern (T-C) soit sensible au nombre de syllabes qui composent l'énoncé; en effet, même dans une très courte séquence comme 6) Mario va a Roma (Mario va à Rome; cf. fig. 6) , prononcée sans aucune pause, l'interprétation des deux blocs du topic-comment (T-C), qui correspondent aux deux catégories majeures syntaxiques, reste clairement reconnaissable. Quand parfois le topic s'avère articulé avec des enchâssements ou des expansions, ceux-ci montrent, à l'enregistrement, des crêtes de remontée (Martin I979, à paraître), mais la forme générale du topic ne change pas.

Les syllabes, qui sont nos unités de référence en ce qui concerne la mensuration des crêtes d'onde, des valeurs en générale du F_0 et de la durée, correspondent à des séquences acoustiques obtenues sur la base des caractéristiques spectrographiques, aussi bien que des valeurs relatives au parcours du F_0 et à l'énergie.

Notre critère a été de choisir des syllabes phonétiques parfois en contraste partiel avec la syllabation phonologique.

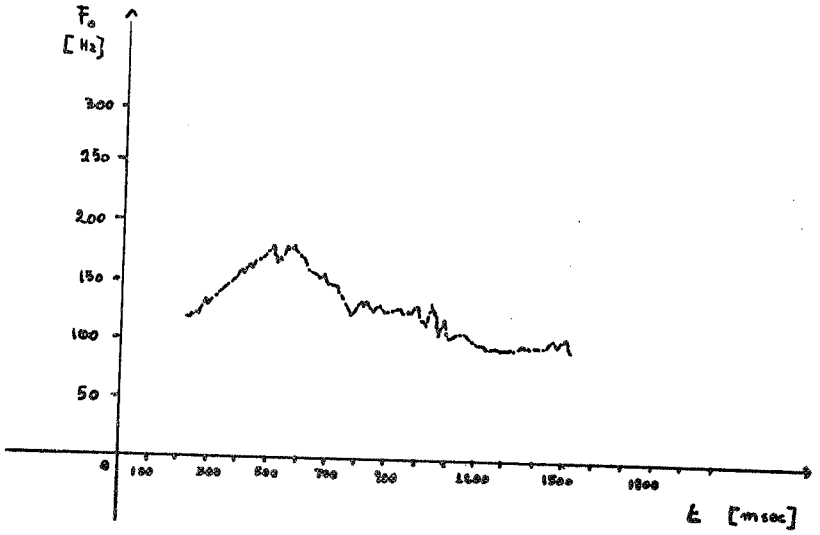


Fig. 6 - Frase T-C "Mario va a Roma"
Profilo di F_0
Phrase T-C Contour du F_0
T-C sentence - F_0 contour

Nous dirons tout de suite que nous nous sommes trouvés face à quelques phénomènes d'exécution assez intéressants. Nous avons constaté, par exemple, des cas de fusion syllabique touchant en particulier la fin du topic, entre la syllabe finale (-cia #) et l'initiale du comment (# a - ddor -), avec la formation de une nouvelle syllabe (- ciaa # ddor -) ou même la formation d'une syllabe (- ciaad # dor -). Dans ce cas, un autre phénomène de fusion auquel aucun de nos locuteurs n'a pu se soustraire, c'est celui de la liaison de la fin du SV avec le début du Spr au sein du comment (- to - in -), qui a donné dans tous les cas une syllabe (- toin -) avec un diphtonque oï ascendant, ou bien (-t@in -) à première voyelle "schwa".

Nous pensons que ces phénomènes regardent des règles d'exécution qui n'infirment pas les règles d'interprétation de l'intonation proposées, mais qui se superposent à elles, comme certaines caractéristiques phonétiques, telles que l'identité phonétique de la dernière voyelle du topic avec la première du comment. Nous avons déjà pu enregistrer dans de nombreuses phrases la fusion du morphème de l'auxiliaire en début de comment avec la dernière syllabe du SN dans le topic.

PROCEDURE EXPERIMENTALE

Notre système de traitement, par ordinateur, de l'analyse acoustique indexe temporellement, de façon uniforme, les paramètres précédents et permet, à l'aide d'un sous-programme, l'écoute et la segmentation auditive des séquences sonores, qui nous mettent en mesure de confronter les séquences obtenues par segmentation auditive aux séquences repérées avec d'autres paramètres.

Notre système d'analyse se divise en deux parties:

I° - un système à microprocessus (Ph. Martin), pour obtenir en temps réel la fréquence fondamentale F_0, représentée d'une part, visiblement, sur écran de monitoring, de l'autre graphiquement sur "plotter" optique Visicorder Honeywell;

2° - un système sur ordinateur (Mancini - Macerata), qui nous donne la possibilité d'obtenir un spectre , la fréquence fondamentale et la valeur

de l'énergie associée à la phrase prise en considération. En outre, nous sommes en mesure d'effectuer des analyses auditives des phrases, fractionnées par intervalles de 800 millisecondes, à l'intérieur desquels nous disposons d'une ultérieure segmentation en 64 sous-intervalles, d'une durée de I2,5 millisecondes. Cette segmentation est obtenue à l'aide d'un appareil (voice-dissector), branché sur le même ordinateur, qui inscrit sur écran visuel les sous-intervalles à l'écoute.

Ces deux systèmes employés parallèlement nous consentent de vérifier les résultats obtenus relativement au F_0, aussi bien que l'enchainement des information quantitatives et les mesures objectives des éléments auditifs.

RESULTATS

Nous donnons ici les valeurs de durée des patterns et des syllabes, avec les moyennes et les pourcentages d'allongement de nos six locuteurs, et les valeurs réelles du F_0 pour un locuteur masculin et un féminin en positions jugées significatives, telles que le début, la crête maximale du bloc, la conclusion.

PHRASE T-C - HOMME		
POSITION	F_0 [Hz]	SYLLABES
T – DÉBUT	125	Lu
T – CRÊTE	195	cIa
T – CONCLUSION	95	A
C – DÉBUT	110	dOr
C – CRÊTE	160	mI
C – CONCLUSION	95	cinə
PHRASE C - HOMME		
DÉBUT	125	Ad
CRÊTE	185	dOr
CONCLUSION	95	nA

PHRASE T-C - FEMME		
POSITION	F_0 [Hz]	SYLLABES
T – DÉBUT	205	Lu
T – CRÊTE	285	cIa
T – CONCLUSION	235	ciA
C – DÉBUT	200	A
C – CRÊTE	240	mI-tən
C – CONCLUSION	165	cinə
PHRASE C - FEMME		
DÉBUT	200	Ad
CRÊTE	275	mI
CONCLUSION	165	cinə

Table I - Valeurs du F_0 pour la phrase T-C et pour la phrase C.
(On a souligné dans la syllabe le segment phonétique auquel le valeur du F_0 se refère).
Fondamental frequency values related to T-C and C sentence.
(We have underlined in the syllable the phonetic segment which carries out the F_0 value).

En ce qui concerne les valeurs de durée, nous observons que tous nos parleurs ont prononcé les phrases proposées à une vitesse considerée normale.

Locuteurs		Durée phrase T-C [msec] Moyenne totale	Moyenne syllabique	Durée phrase C [msec] Moyenne totale	Moyenne syllabique
F	E	1884 ± 50	(8) 235	1188 ± 54	(6) 198
	A	1494 ± 66	" 187	1068 ± 36	" 178
	G	1866 ± 28	" 233	1194 ± 42	" 199
H	L	1668 ± 44	" 208	1116 ± 52	(7) 159
	F	1578 ± 54	(7) 225	1188 ± 86	" 170
	M	1362 ± 42	" 195	978 ± 18	(4) 153

Locuteurs	Durée général phrase T-C [msec] Moyenne	Moyenne syl.	Durée général phrase C [msec] Moyenne	Moyenne syl.
F	1766 ± 180	221	1154 ± 80	192
H	1546 ± 128	221	1123 ± 45	161

Table 2 - Durée moyenne des phrases pour chaque locuteur et durée moyenne générale pour groups de locuteurs masculins et feminins. Length average of the sentences (values for every speaker). General sentence length average (values for groups of male and female speakers).

Phrase T-C . F [msec]		Syllabes	Durée moyen.	
T		Lu	141	544
		cia	403	
C		a	95	1222
		dder	198	
		mi	201	
		tein	202	
		cu	117	
		cinà	409	

Phrase T-C . H [msec]		Syllabes	Durée moyen.		
T		Lu	108	490	
		cia-a	382		
C		dder	196	1058	
		mi	160		
		tein	210		
		cu	118	1152	
		cinà	374		
		ci	270		
		na	162		

Table 3 - Phrase T-C Durée moyenne des syllabes.
T-C sentence - Syllabic length average.

PHRASE C. F [msec]	
SYLLABES	DURÉE MOYEN.
a	96
ddor	222
mi	125
tʒin	198
cu	142
cinə	371

PHRASE C. H [msec]	
SYLLABES	DURÉE MOYEN
a	75
ddor	185
mi	159
tʒin	182
cu	120
cinə	270**
ci	285
na	117

Table 4 - Phrase C Durée moyenne des syllabes.
C sentence - syllabic length average.

*) Syllabation phonétiques Ci-na.
**) Syllabation phonétiques Cinə .

REFERENCES

CRESTI, E., I979, L'intonazione come fenomeno linguistico; -Annali della Classe di Lettere della Scuola Normale Superiore, n°I, a paraître.

't HART, J., COLLIER, R., I975, Integrating different levels of intonation analysis; -Journal of Phonetics, n°3, pp.235-255.

MANCINI, P., MACERATA, A., I977, La strumentazione di analisi fonetica sviluppata alla Scuola Normale Superiore. -Studi di grammatica italiana, VI, pp. 9-2I.

MARTIN, Ph., I978, L'intonation de la phrase en italien; -Studi di grammatica italiana, VIII, a paraître.

VAISSIERE, J., I975, On french Prosody; -Quaterly Progress Report, n°II4, Research Laboratory of Electronics, M.I.T..

VAISSIERE, J., I976, Premiers essais de segmentation automatique de la parole continue en mots à partir des variations du fondamental dans la phrase; -Recherches acoustiques, vol. III, CNET, Lannion.

10èmes JOURNÉES D'ÉTUDE SUR LA PAROLE

GRENOBLE - 30 MAI - 1ER JUIN 1979

LE COMPILATEUR DE REGLES DE REECRITURE TOP ET SON UTILISATION A LA TRANSCRIPTION DU FRANCAIS EN VUE DE LA SYNTHESE.

M. DIVAY & M. GUYOMARD

IRISA RENNES
CNET LANNION
IUT Département "Informatique" BP 150 22302 LANNION

RESUME

Nous décrivons un langage de programmation conçu pour simplifier la prise en compte automatique de règles de réécriture, et son utilisation dans la transcription graphémo-phonétique du français.

Le système de règles est basé sur l'application d'un ensemble partiellement ordonné de règles. La partie gauche de chaque règle indique les graphèmes pris en compte par la règle. La partie droite de chaque règle précise les phonèmes correspondants et éventuellement les contextes requis précédant et suivant les graphèmes pour que la règle s'applique. Les règles context- sensitives sont examinées de telle sorte que, dans un bloc de règles, celles relatives aux prononciations particulières soient examinées en premier. Les dernières règles examinées sont les plus générales. D'autres règles ont été ajoutées pour traiter des problèmes de "e" muet, liaisons, élisions, transcription en lettres de nombres écrits en chiffres...

Le langage est utilisé comme un outil, très souple, de recherche de règles qui peuvent facilement être définies ou modifiées. Après compilation des règles (compilateur TOP), il suffit d'interpréter (interpréteur TOP) le programme objet pour avoir une transcription fidèle aux règles définies.

THE TOP LANGUAGE AND ITS USE IN ORTHOGRAPHIC-TO-PHONETIC TRANSCRIPTION

M. DIVAY & M. GUYOMARD
IRISA RENNES
CNET LANNION
IUT Département "Informatique" BP 150 22302 LANNION

SUMMARY

A programming language designed for the simple description of letter-to-sound rules is described. Its use in an application of grapheme-to-phoneme transformation rules system for French is studied, and some results are given.

The rules system is based on the application of a partially ordered set of phonological rules. The left-hand side of each rule indicates the graphemes involved by the rule. The right-hand side of each rule specifies the corresponding phonemes and possibly the preceding and succeeding graphemic context. The context-sensitive rules are processed in such a way that the rules related to exceptional pronunciations are first examined in the set. The last examined rules are the more general ones. Additional rules have been implemented to deal with the problem of mute "e", liaisons, linkings and numbers written as numerals.

This program has been designed primarily to serve as a flexible research tool. The set of rules can be easily defined, revised and expanded.

10èmes JOURNÉES D'ÉTUDE SUR LA PAROLE
GRENOBLE - 30 MAI - 1ER JUIN 1979

LE COMPILATEUR DE REGLES DE REECRITURE TOP ET SON UTILISATION A LA TRANSCRIPTION DU FRANCAIS EN VUE DE LA SYNTHESE.

M. DIVAY & M. GUYOMARD
IRISA RENNES
CNET LANNION
IUT Département "Informatique" BP 150 22302 LANNION

INTRODUCTION

En français, l'écart entre le graphème et le phonème est assez important, par exemple, dans le mot "oiseau" (/wazo/), aucun des sons ne se retrouvent dans l'orthographe. Un son peut être obtenu à partir de plusieurs lettres : /s/ se retrouve dans les graphèmes : son, passion, ration, grâce, maçon, science, soixante ; /ɛ/ peut s'écrire de 24 façons différentes : mère, Noël, fête, fer, chef-d'oeuvre, peine, sept, tes, congrès, bérêt, intérêt, bey, aspect, legs, est, entremets, paire, chaîne, laid, monnaie, relais, lait, faix, tramway. Au contraire, une seule lettre peut représenter différents sons : "x" représente /ks/ dans "axe", /s/ dans "six", /z/ dans "sixième", /gz/ dans "exact" (BURNEY, P., 1970).

LE PRINCIPE GENERAL DE LA TRANSCRIPTION

Bien que ne connaissant pas certains mots, un locuteur peut, la plupart du temps, les prononcer correctement. Il procède par analogie avec les mots connus, ou applique certaines règles telles que : "s" se prononce /z/ dans un contexte voisé ; cependant, il existe un certain nombre d'exceptions que seule l'expérience permet d'acquérir. Ainsi la règle du "s" donnée précédemment ne s'applique pas pour "soubresauts", "entresol", "cosinus" ... ; "soubresauts" et "entresol" la concaténation de 2 mots qui gardent leur propre prononciation. En français, ce genre d'exceptions est assez rare. En anglais, c'est une des raisons qui fait que la transcription par règles est insuffisante, et doit être précédée d'une décomposition en morphes (ALLEN, J., 1976), (HUNNICUT, S., 1976), (LEE, F., 1972).

La transcription automatique d'une chaîne écrite peut se faire de deux façons : à l'aide d'un dictionnaire contenant l'équivalent phonétique de chaque mot, ou à l'aide de règles de réécriture (COULOM, D., KAISER, D., 1976), (FERVERS, H., LE ROUX, J., MICLET, L., 1976), (LE CORNEC, A., 1972), (RODET, X., POIROT, L., 1976), (SILVA, 1969). L'utilisation d'un dictionnaire est une solution simple si le nombre de mots est limité. Par contre, pour un vocabulaire illimité, incluant les noms propres, l'utilisation d'un dictionnaire n'est plus envisageable. Dans ce dernier cas, il faut développer des règles graphémo-phonétiques afin de passer du texte écrit au son à prononcer. L'élaboration de ces règles s'appuie sur le raisonnement donné dans les exemples (A) et (B) suivants :

(A) (1) "i" se prononce généralement /i/
(2) "oi" se prononce /wa/
(3) "oin" se prononce /wɛ̃/

En accord avec ces 3 règles, les parties "oin", "oi", "i", des mots français suivants :

.../...

lit, ami, petit, (1)

oie, poids, loi, fois, moi, roi, boi, (2)

poing, loin, coin, soin, foin, (3)

sont correctement transcrites seulement si la règle (3) est examinée avant la règle (2) et la règle (2) avant la règle (1).

Ainsi : "coin" → /kwɛ̃/
"moi" → /mwa/
"ami" → /ami/

(2) est un cas particulier de (1) ; (3) est un cas particulier de (2). Ces règles doivent être examinées dans l'ordre (3), (2), (1).

(B) "s" dans un contexte voisé se prononce /z/. Il faut définir des classes (voyelles, consonnes...) auxquelles se réfèrent les règles phonologiques lors de l'examen des contextes droits ou gauches.

Des règles supplémentaires sont à ajouter pour prendre en compte les problèmes de :

(a) "e" muet, le graphème "e" peut être prononcé ou non suivant sa position dans le mot : forme, tellement, petit, ...

(b) Liaisons, la consonne finale non articulée dans le mot isolé est prononcée devant la voyelle initiale du mot suivant : "nous _ attendons"

(c) Enchaînements, la consonne finale articulée dans le mot isolé, est enchaînée, sans pause, avec la voyelle du mot suivant : partir_ en courant, ...

(d) Transcription phonétique des nombres écrits en chiffres : 32 est transcrit /trãtdø/

LES REGLES DE REECRITURE

Le langage TOP (Transcription Orthographique Phonétique) est un langage de description de règles context-sensitives. Un programme écrit en TOP comprend :

Une partie déclaration des codes utilisés

Une partie déclaration de classes (optionnelle)

Une partie règles proprement dites.

Dans la partie code, il s'agit de déclarer les codes externes (s'étendant éventuellement sur plusieurs caractères) choisis pour représenter les graphèmes et les phonèmes. La déclaration de ces codes n'est pas implicite de façon à donner plus de souplesse et de généralité (phonèmes différents pour d'autres langues) au langage TOP.

Les classes sont des ensembles de chaînes de caractères composées de codes externes. On peut déclarer des classes de voyelles, consonnes, fricatives, nasales,... Certaines règles nécessitent un examen des contextes gauche et droit. Ce contexte peut être exprimé comme une chaîne, ou un élément d'une classe, ou les deux à la fois. .../...

Exemple : /voyelle/ +_ ; signifie : si le contexte gauche, de la chaîne examinée est un élément de la classe 'voyelle' et si le contexte droit est un blanc alors la règle s'applique.

Les règles de réécriture s'écrivent comme suit :

étiquette : LHS ⟶ [RHS] [/ [LC] + [RC]] ;

où étiquette est un nombre entier.

LHS représente la chaîne à identifier dans le texte à transcrire.
RHS représente la chaîne qui se substitue à LHS si la règle s'applique. Ce peut être une chaîne vide.

LC et RC représente respectivement le contexte gauche et droit, exprimés comme des expressions régulières:

Certain.|s signifie "certain" ou "certains"

tot+(a, on, ell) signifie contexte gauche : "tot"
contexte droit : "a" ou "on" ou "ell"

11420 : 2 → vingt / + 'ch' . __ ; signifie : 2 se réécrit "vingt" si le contexte droit est un élément de 'ch' (donc un chiffre) suivi d'un espace.

La méthode de transcription consiste à partir du texte à transcrire et à déterminer un sous-ensemble des règles susceptibles de s'appliquer aux caractères courants examinés. Soit T le texte à transcrire, et x le premier caractère à traduire. Si les règles sont :

y → ab ;
xa → pd1/cg1 + cd1 ; (1)
xab → pd2/cg2 + cd2 ; (2)
z → e1/ab + ;
x → pd3/cg3 + cd3 ; (3)
x → pd4/cg4 + cd4 ; (4)
w → zy ;
xa → pd5/cg5 + cd5 ; (5)

Le sous-ensemble des règles qui peuvent s'appliquer comprend :
(1), (2), (3), (4), (5) . Il n'y a pas de notion de séquentialité de règles. Le sous-ensemble est examiné de façon à donner la priorité aux règles particulières, et donc aux plus longues parties gauches (LHS). La méthode consiste à examiner le sous-ensemble par ordre décroissant de longueur de partie gauche et pour une même longueur, dans l'ordre des règles spécifiées par le programmeur. Ainsi, sur l'exemple, les règles sont examinées, jusqu'à ce qu'un bon contexte soit trouvé, dans l'ordre : (2), (1), (5), (3), (4).

De cette façon, la transcription pourrait être faite en un seul examen de la chaîne à transcrire. Ceci n'est pas possible car, certaines transcriptions représentent des marques qu'il faut réexaminer plus tard ; plusieurs passages sur le texte à transcrire sont nécessaires. Il faut donc partitionner l'ensemble des règles en blocs (repérés par "début" et "fin"), chaque bloc de règles

.../...

nécessitant un passage sur le texte à transcrire.

TRANSCRIPTION AU NIVEAU DU MOT ISOLE

Notre but est de transcrire en phonétique n'importe quel texte français, y compris les noms propres. La décomposition en morphes ("soubre" et "sauts" pour "soubresauts") ne se justifie pas en français et peut être traitée parmi les cas particuliers, à l'aide de règles.

La conversion au niveau des mots isolés s'effectue en deux étapes. La première est une phase de normalisation où les marques du pluriel non pertinentes sont supprimées (la lettre "s" en général). Un "s" final qui doit être prononcé ("hélas"→/elas/, "jadis"→/ʒadis/) est conservé afin d'être traduit dans l'étape suivante. Moyennant ce pré-traitement, les mots sont plus faciles à transcrire et ne nécessitent pas de redondance ni de duplication de règles. Cette procédure de normalisation est constituée de 16 règles. Dans la seconde étape, 270 règles (DIVAY, M., GUYOMARD, M., 1977) sont utilisées pour traduire une chaîne graphique en une chaîne phonétique.

Par exemple, trois des règles s'appliquant à la prononciation de la chaîne graphique "ai" sont :

- "ai"→/ɛ/ ; (1) : c'est la prononciation générale de cette chaîne

- "ai"→/ə//f+s.'voyelle' ; (2) (s. 'voyelle' signifie : lettre "s" suivie d'une voyelle). C'est une prononciation particulière de "ai" pour les formes dérivées du verbe faire ("défaisait"→/defəzɛ / , "faisant"→/fəzɑ̃/.

- "ain"→/ɛ̃/ / + 'cgem', - ; (3) : La chaîne "ain" produit le son /ɛ̃/ si le contexte droit est une consonne graphique excepté "m", ou en position finale (représentée par "_"). Par exemple "_pain_" → /_p ɛ̃ _/.

Cependant afin d'assurer une traduction correcte des formes dérivées de "faire", la règle (2) doit être essayée avant la règle (1), alors que la règle (3) peut apparaître à tout endroit car sa partie gauche (LHS) est plus longue que celle des règles (1) et (2). Ainsi les 3 seules séquences acceptables dans l'ordonnancement des règles sont : (3), (2), (1), (2), (3), (1), et (2), (1), (3).

Exemple de conversion de mots isolés

(On examine la transcription de 1 mot et on commente l'interprétation).

Dans les figures 1 et 2, la première colonne représente le ruban d'entrée contenant la chaîne à traduire et la fenêtre rectangulaire encadrée, la partie gauche de la règle courante ; la seconde colonne représente le ruban de sortie contenant, à la fin du traitement, la chaîne phonétique résultante.

.../...

Traduction de "entonnions"

	Graphème	Phonème	
1	_ entonnions _	_ entonnion _	
2	_ entonnion _	_ ã	1ère étape
3	_ entonnion _	_ ãt	
4	_ entonnion _	_ ãt	
5	_ entonnion _	_ ãtɔn	
6	_ entonnion _	_ ãtɔnn	2ème étape
7	_ entonnion _	_ ãtɔnnj	
8	_ entonnion _	_ ãtɔnnjõ	

Fig. 1

La première étape conduit à la suppression du "s" final. Bien qu'en toute rigueur ce caractère ne constitue pas la marque du pluriel, il n'est pas prononcé. Dans la 2e étape, la règle :

"en"→/ã/ dans un contexte droit de consonne orale

est appliquée, puis la lettre "t" est traduite en /t/, ceci étant la règle générale. La 4ième ligne transcrit la lettre "o" en un phonème /o/. La règle "on"→ /õ/ n'a pu être retenue car elle ne s'applique que si la chaîne "on" se trouve en position finale d'un mot ou suivie d'une consonne orale. Les lignes 5 et 6 expriment l'effet de la seule règle débutant par "n". La traduction de la lettre i est plus complexe, elle peut conduire soit à /i/ soit à /j/ selon le contexte droit. Ici la règle :

"i"→/j/ devant une voyelle excepté un "e" muet s'applique.

La règle "on"→/õ/ déjà mentionnée est alors exécutée pour traduire les dernières lettres.

Les difficultés rencontrées au niveau de la transcription des mots isolés proviennent de deux sources :

-a- La transcription des chaînes contenant "ti", qui peut fournir soit /t.../ soit /s.../. (sortie mais inertie).

-b- Les mots s'achevant par "ent" (qui produisent soit /ə/ dans le cas d'un verbe, soit /ã/ dans les autres cas) ou par "er" (qui est prononcé /ɛ/ ou /e/).

La plupart de ces difficultés sont résolues dans notre programme de transcription grâce à l'examen des contextes.

TRANSCRIPTION DE LA PHRASE

La transcription d'une phrase complète n'est pas, en général, la "somme" des transcriptions des mots la constituant. On doit tenir compte des phénomènes particuliers déjà mentionnés. D'un point de vue technique ces caractéristiques ont

.../...

également été mises en oeuvre par le moyen de règles, ainsi, le système est construit à l'aide d'un formalisme unique.

LIAISON ET ENCHAINEMENT

Nous traitons de ces 2 problèmes simultanément car le premier traitement les concernant travaille sur la forme graphique et, à ce niveau, la distinction entre liaison et enchaînement ne peut être faite. Notre solution est basée sur la stratégie suivante :

- Traiter en priorité les liaisons obligatoires
- Traiter quelques liaisons optionnelles faciles à détecter
- Eviter toutes les autres.

Cela est possible en constatant que les liaisons sont nécessaires après des articles, prénoms, certains adjectifs :"les amis"→/lɛzami/; "le petit enfant"→/ləpətitɑ̃pɑ̃/.De tels adjectifs sont fréquemment utilisés mais sont peu nombreux (aux environs de 10). La raison profonde de ceci tient à la structure sous-jacente de la langue française liant un mot à son successeur révélant une forte dépendance du premier par rapport au second. Le traitement résultant comprend 2 phrases : la première (survenant avant la transcription des mots isolés) introduit des marqueurs de liaison - enchaînement ne faisant pas intervenir de considérations syntaxiques. La seconde phase utilise ces marqueurs et le fait que les mots sont sous forme phonétique, pour décider d'une liaison ou d'un enchaînement.

. Elision :

En français, le seul cas d'élision dans la parole continue est celui de la lettre "e" : "dangereux" → /dazrø). Une seule règle traite la majorité des cas : tous les /ə/ sont muisés à l'exception de ceux apparaissant dans la 1ère syllabe ou précédés de 2 consonnes pour être en accord avec la règle des 3 consonnes.

. Ambiguité de prononciation :

La représentation de surface d'une langue parlée est influencée par sa structure profonde. En anglais ce phénomène se révèle principalement par l'intermédiaire de l'accentuation. En français, un problème similaire surgit au niveau phonétique

"il_est_dans_l'est" → / i l ɛ d ɑ̃ l ɛ s t /

"nous_dictons_des_dictions" → / n u d i c t j õ d ɛ d i k s j õ /

Une approche basée sur une analyse syntaxique serait nécessaire pour résoudre convenablement ce problème. L'ensemble de règles ne prend pas en compte de telles considérations et nous effectuons simplement le choix semblant le plus probable.

. Structure fonctionnelle du système traitant les phrases :

Le processus de conversion texte - phonèmes est effectué par l'exécution séquentielle des 10 modules suivants :

- A - Codage : c'est la seule étape à ne pas être implémentée par règle. Elle traduit la représentation des accents (exprimés sur 2 caractères) en un code interne et détecte les erreurs de "syntaxe" relative à l'utilisation du code externe.

.../...

(2) - Normalisation du texte (15 règles) : traite la ponctuation et autres séparateurs afin de normaliser le texte pour les phases suivantes. Les marques de production ne sont pas détruites de façon à transmettre l'information prosodique qu'elle représente jusqu'à la synthèse vocale.

(3) - Transcription des nombres en graphèmes (50 règles) : cette étape intervient avant la 1ère phase de liaison - enchaînement qui est ainsi apte à traiter les nombres sous leurs formes graphiques.

(4) - Première étape de liaison - enchaînement, elle a déjà été explicitée et comprend 22 règles.

(5) (6) - Transcription des mots isolés : c'est une phase divisée en 2 étapes déjà étudiées.

(7) - Suppression des géminées (13 règles). Les géminées sont systématiquement transformées en consonne simple. Cette façon de procéder est simple et parfois insuffisante : "Grammaire" donne alors /gramɛr/ alors que la prononciation correcte est /grammɛr/.

(8) - C'est le second module de liaison-enchaînement (19 règles)

(9) Elisions : elles s'élaborent à l'aide de 3 règles.

(10) - Etape facultative de transcription du résultat en un code imprimable (56 règles).

CONCLUSION

Ce langage a été conçu comme un outil de recherche dans le traitement de texte, de phonétique et de linguistique. L'ensemble de règles peut être facilement défini et mis à jour. Les résultats actuels, dans le domaine de la transcription graphémo-phonétique, fournissant un taux d'erreur de 0.4 % pour les mots isolés et de 2 % pour les liaisons. La chaîne phonétique résultante est transmise à un synthétiseur. Une bonne synthèse doit cependant prendre en compte la variation du fondamental au cours de l'élocution. Actuellement, ce dernier traitement est en cours d'automatisation.

Ce travail a été réalisé au département SST du Centre National d'Etudes des Télécommunications (C.N.E.T.) de LANNION.

REFERENCES

ALLEN, J., 1976, Synthesis of speech from Unrestricted text, proceeding of the IEEE, April 1976

BURNEY, P., 1970, L'orthographe, collection "Que sais-je" no 685, Boulevard St Germain, PARIS

COULOM, D., KAISER, D., 1976, Analyse de réponses rédigées en français courant. Revue de la RAIRO, juin 1976

FERVERS, H., LE ROUX, J., MICLET, L., 1976, ENST-D-76003 ENST PARIS FRANCE

FOUCHER, P., 1969, Traité de prononciation française, PARIS-Editions, Lkincksieck

French Review XXI, 1947, 2 december 47, XXII 6 May 49, XXXI October 56

HUNNICUT, S., 1976, Natural language processing group. Research laboratory of Electronics, M.I.T., Cambrige, Massachusetts

LE CORNEC, A., 1972, La transcription orthographique phonétique du français, Essai de formalisation NT/CEI/CSI/22-CNET LANNION FRANCE

LEE, F., 1972, Machine-to-man communication by speech, Research Laboratory of Electronics, M.I.T., Cambridge, Massachusetts

RODET, X., POIROT, L., 1976, Transcription phonétique automatique du français. SES/INTERNE SERF/76-068 - C.E.A. SACLAY FRANCE

SILVA, 1969, An automatic orthographic-to-phonetic conversion system for French, Institute of Library Research. University of California, LOS ANGELES

TEIL, D., 1975, Conception et réalisation d'un terminal à réponse vocale. Thèse de Docteur Ingénieur, LIMSI, Orsay, FRANCE

DIVAY, M., GUYOMARD, M., 1977, Grapheme to phoneme, transcription for french. IEEE Symposium may 1977. Hartford Connecticut, U.S.A.

DIVAY, M., GUYOMARD, M., 1977, Conception et réalisation sur ordinateur d'un programme de transcription graphémo-phonétique du français", Thèse de 3e cycle, Université de RENNES.

10èmes JOURNÉES D'ÉTUDE SUR LA PAROLE

GRENOBLE - 30 MAI - 1ER JUIN 1979

TROPS : UN SYSTEME DE TRANSCRIPTION ORTHOGRAPHIQUE PHONETIQUE ET DE SYNTHESE DU FRANCAIS.

P. GOYER, D. DEGRYSE, B. GUERIN

Laboratoire de la Communication Parlée
E.N.S.E.R.G.
23 Avenue des Martyrs - 38031 GRENOBLE CEDEX
Equipe de recherche associée au C.N.R.S. - ERA 366

RESUME

"TROPS", système en cours d'étude au Laboratoire de la Communication Parlée est un système de synthèse en temps réel d'un texte français orthographié.

Sa caractèristique essentielle sera un fonctionnement quasi simultané de l'acquisition du texte, de la transcription orthographique phonétique et de la synthèse.

Dans cet article nous decrirons uniquement la procédure de passage d'un texte orthographique français en un texte phonétique.

Cette procédure comporte deux phases essentielles :

- la division du texte en syllabes
- la transcription de ces syllabes en symboles phonétiques.

Cette transcription est effectuée par examen séquentiel de plusieurs listes de références et substitution des codes phonétiques désirés. Cette méthode permet de tenir compte aiséement des particularités du français.

Cette procédure a été programmée sur un miniordinateur LSI 11 et nécessite de l'ordre de 5000 octets de mémoire.

TROPS : UN SYSTEME DE TRANSCRIPTION ORTHOGRAPHIQUE PHONETIQUE ET DE SYNTHESE DU FRANCAIS.

P. GOYER, D. DEGRYSE, B. GUERIN.
Laboratoire de la Communication Parlée
E.N.S.E.R.G.
23 Avenue des Martyrs - 28031 GRENOBLE CEDEX
Equipe de recherche associée au C.N.R.S. - ERA 366

SUMMARY

In this paper, a procedure for trancisbing a french text into phonetic symbols is described. This procedure will be used in a real time system for french speech synthesis. This system is studying at the "Laboratoire de la Communication Parlée".
This procedure can be divided into two major parts :

- syllabification of text, because french pronunciation is governed in particular by the distribution of consonants between syllabes
- phonetic translation of the syllabized text.

The translation scans the syllabized text on successive cycles in groups of four and more, three, two, one characters. Every valid group is compared to a reference list. When a match occurs the correspondant phonetic code is substituted for the group. This procedure also handles the transcription of those letters of which the pronunciation can be context varying.

One version of this procedure has been programmed, in assembly language on a microcomputer LSI 11 and needs only 5 K bytes of memory and its execution speed will authorize its use in a real time system for french speech synthesis.

10èmes JOURNÉES D'ÉTUDE SUR LA PAROLE
GRENOBLE - 30 MAI - 1ER JUIN 1979

TROPS : UN SYSTEME DE TRANSCRIPTION ORTHOGRAPHIQUE PHONETIQUE ET DE SYNTHESE DU FRANCAIS

P. GOYER, D. DEGRYSE, B. GUERIN.
Laboratoire de la Communication Parlée
E.N.S.E.R.G.
23 Avenue des Martyrs - 38031 GRENOBLE CEDEX
Equipe de recherche associée au C.N.R.S. - ERA 366

1.INTRODUCTION

Dans le cadre de la réalisation du système autonome de synthèse du français nous avons développé une procédure de passage d'un texte orthographié français en un texte phonétique.

Cette procédure se compose d'une division du texte en syllabes et d'une transcription de ces syllabes en codes phonétiques (PRATT B. § al., 1967).

La première étape est réalisée par une comparaison des syllabes avec des listes de référence, les syllabes étant successivement découpées en groupe de quatre et plus, trois, deux et une lettre. Cette méthode permet également de prendre en compte des groupes de lettres dont la prononciation dépend du contexte dans lequel elles se trouvent.

Le principal intêrêt d'une telle méthode est de pouvoir être adaptée aiséement à un vocabulaire particulier tout en ayant une occupation mémoire réduite.

2. PROCEDURE DE SYLLABIFICATION

La prononciation française repose sur la syllabification, elle dépend en particulier de la distribution des consonnes entre les syllabes;

La procédure de syllabification divise le texte en syllabes suivant les régles de base suivantes :

- chaque syllabe doit contenir une voyelle
- s'il y a seulement une consonne entre deux voyelles, elle introduit une syllabe (ex: a-mer)
- s'il y a plus d'une consonne, seule la dernière introduit une nouvelle syllabe (ex. : cons-tant)
- certaines paires de consonnes (bl, pr...) introduisent toujours une nouvelle syllabe (ex. : com-prend)

Il y a cependant quelques exeptions à la règle dans les cas particuliers suivants :

- la voyelle finale d'une terminaison polysyllabique en e et es est muette dans la phrase française (ex. :simple, ondes)
- la combinaison des voyelles doubles pose un problème dans le processus de syllabification. Les combinaisons suivantes : aé, éa, éi, ia, io et les voyelles précédées ou suivies de ï sont découpées ; ainsi les autres com-

binaisons telles que ié, ieu, ien ... seront considérées comme formant une seule syllabe.

- le h est toujours considéré comme non aspiré car il est très difficile de lever l'ambiguité
- les deux voyelles de gne et de que sont habituellement muettes.

3. PROCEDURE DE TRANSCRIPTION

Cette procédure tente de résoudre un double problème : d'une part, le fait que certains sons peuvent être représentés en français par plusieurs orthographes, d'autre part le fait q'un certain groupe de lettres peut avoir des sons différents selon le mot dans lequel il se trouve. Dans chaque cas le son doit être défini dans le contexte, de telle sorte qu'il puisse être identifié de façon exacte. Néanmoins, ce problème est grandement facilité par une syllabification correcte.

La procédure de transcription est effectuée en plusieurs phases pendant lesquelles la phrase syllabisée est successivement analysée par groupes de quatre, trois, deux, une lettres.

Dans la première phase, le texte est partagé par groupes de quatres lettres ou plus. Dès qu'un groupe valide est trouvé, il est comparé à une liste de références. Si une correspondance est trouvée, le code phonétique de ce groupe particulier est substitué au texte d'origine. Le texte obtenu à la fin de cette première phase sert de données d'entrée pour la deuxième phase. (ex. : doigt → d wa ; rhum → rom)

Dans la deuxième phase la même analyse est effectuée mais en partageant le texte en groupes de trois lettres et en utilisant une liste de références contenant des groupes de trois lettres. Le texte à analyser, ayant été en partie codé par la première phase, contient aussi bien des lettres que des codes phonétiques, ainsi un groupe valide ne doit pas contenir de codes phonétiques. (ex. : main → m$\tilde{\varepsilon}$; dix → dis)

Dans la troisième phase le texte en partie codé est partagé en groupes de deux lettres et les paires de consonnes inséparables sont codées.

Dans la quatrième phase le texte est à nouveau partagé en groupe de deux lettres mais à la différence des autres phases, les espaces sont pris en compte notamment le groupe /e /.

Dans la cinquième phase le texte est considéré caractère par caractère et le code numérique approprié est substitué à chaque lettre restante.

Cette procédure de comparaison avec des listes de références convient très bien aux lettres ou groupes de lettres dont la prononciation ne varie jamais ; cependant certains ont une prononciation différente selon le contexte dans lequel ils se trouvent.
Ainsi, par exemple, suivant la phase, les cas particuliers suivants sont examinés :

- le groupe "cher" peut avoir deux prononciations suivant qu'il est au début ou en fin de mot (ex. : chercher)
- la terminaison "es" est silencieuse s'il n'y a pas de liaison avec le mot suivant
- si le groupe "gu" est suivi d'une voyelle la lettre "u" n'est pas prononcée
- si les groupes "de", "re" commencent un mot et sont suivis de "ll", alors la lettre "e" est prononcée /e/

Néanmoins, la reconnaissance de la terminaison ent pose un problème qui ne peut être entièrement résolu. En effet, elle peut être silencieuse si c'est une terminaison de verbe, ou bien correspondre à un son /a/ si c'est une fin d'adverbe. Cette ambiguité peut être levée uniquement dans les combinaisons mment, ement, ément, ament qui ne sont pas des terminaisons de verbes.

4. IMPLANTATION ET PERFORMANCES

Cette procédure a été, dans un premier temps, écrite dans le langage FORTRAN IV du calculateur PDP 11 du Laboratoire et aprés mise au point une seconde version optimisée a été développée en langage assembleur.

Cette seconde version necessite 5 K octets de mémoire qui se décomposent en :

- 600 octets pour la procédure principale
- 300 octets pour la syllabification
- 400 octets pour la procédure de transcription
- 500 octets pour les procédures de contrôle des cas particuliers
- 1200 octets pour les données et divers tableaux de travail
- 2000 octets pour les listes de références.

Le temps d'exécution moyen pour la transcription d'une ligne de 60 caractères est de 1,5 s.

Dans le cadre d'une application de synthèse autonome de textes français une troisième version de cette procédure a été développée. Au lieu d'attendre la fin de la phrase, la transcription s'effectue au fur et à mesure de l'arrivée du texte. Ce texte pouvant être fourni par exemple soit par un ordinateur soit par un lecteur optique. Cette procédure permettra d'assurer une quasi-simultanéité entre l'arrivée du texte et la synthèse finale.

5. CONCLUSION

Dans le cadre d'un système autonome de synthèse en temps réel de textes français orthographiés, nous avons développé une procédure de transcription orthographique phonétique.
Cette transcription est basée sur une décomposition du texte en syllabes suivie d'une comparaison avec des listes de références.

Cette procédure permet, tout en ayant une vitesse d'exécution élevée, d'obtenir une transcription qui tient compte d'un grand nombre de particularités du français et peut être aisement adaptée pour des vocabulaires particuliers.
Son principal intêret est de necessiter une taille mémoire réduite et pourra aisément être intégrée dans un système autonome de synthèse.

6. REFERENCES

PRATT B., SILVA G., 1967. PHONTRS, A procedure which uses a computer for tranasbing french text into phonetic symbols. Monash University

10èmes JOURNÉES D'ÉTUDE SUR LA PAROLE

GRENOBLE - 30 MAI - 1ER JUIN 1979

TRANSCRIPTION ORTHOGRAPHIQUE - PHONETIQUE ET SYNTHESE EN TEMPS REEL DE LA PAROLE PAR PREDICTION LINEAIRE

J. LE ROUX et L. MICLET

Département Systèmes et Communications
ENST 46 rue Barrault 75013 PARIS

RESUME

La transcription orthographique phonétique réalisée décide du (ou des) phonème à associer à chaque caractère orthographique en analysant le texte par un arbre de décisions portant sur les caractères voisins, de profondeur de 0 à 20. Quelques règles simples portant sur des mots ou signes à valeur grammaticale permettent de calculer une intonation. Pour la synthèse, les paramètres de prédiction linéaire associés à un phonème donné sont envoyés à un filtre programmable qui engendre le son correspondant pendant la durée du phonème. La transition entre phonèmes successifs se fait par interpolation linéaire sur ces paramètres. Cette méthode nécessite une place mémoire minime, et la commande est réalisée sur microprocesseur. Elle fournit une synthèse intelligible en temps réel.

TEXT TO PHONEMES TRANSCRIPTION AND REAL TIME SYNTHESIS OF SPEECH USING LINEAR PREDICTION CODING

J. LEROUX and L. MICLET

SUMMARY

In the proposed method, the transcription from text to phonemes chooses the phoneme that will correspond to a character, by analysing a decision tree where its neighbours are considered (its depth varies from 0 to 20 tests). The computation of the intonation is deduced from some simple rules on words and signs having a grammatical importance.

In the synthesis of speech, the linear prediction parameters corresponding to a phoneme are sent to a programmable lattice filter which generates the sound of this phoneme during a given period of time. The transition from one phoneme to another is done by linear interpolation on these parameters. The synthesis is controlled by a microprocessor (as memory). This gives an intelligible synthesis in real time.

10èmes JOURNÉES D'ÉTUDE SUR LA PAROLE
GRENOBLE - 30 MAI - 1ER JUIN 1979

TRANSCRIPTION ORTHOGRAPHIQUE - PHONETIQUE ET SYNTHESE EN TEMPS REEL DE LA PAROLE PAR PREDICTION LINEAIRE

J. LEROUX et L. MICLET

INTRODUCTION

Les travaux effectués au Département Systèmes et Communications de l'E.N.S.T. en synthèse de parole ont pour origine le mise au point d'un synthétiseur hybride (GUEGUEN & AL, 1975) aisément commandable. Nous avons développé autour de cet appareil un certain nombre d'outils qui avaient pour but non pas de créer un système d'excellente qualité acoustique, mais de réduire autant que possible la complexité du système (algorithmes simples et rapides, emplacement mémoire minime) de façon à faire une synthèse en temps réel sur un matériel peu encombrant (actuellement un microprocesseur TMS 9900 et le synthétiseur hybride). A partir de ces résultats nous pouvons mettre en évidence les points où cette technique est suffisante et trouver ceux où une amélioration est nécessaire. Les deux parties importantes de ce travail concernent d'une part la préparation de la synthèse à partir d'un texte orthographié (transcription orthographique - phonétique, génération automatique d'une intonation), et d'autre part la synthèse de sons à partir de la suite des codes phonémiques et de l'intonation trouvés à l'analyse du texte, en utilisant les méthodes liées à la prédiction linéaire (MARKEL, 1976).

TRANSCRIPTION ORTHOGRAPHIQUE PHONEMIQUE

Chaque lettre est analysée en fonction de son contexte sous la forme d'un arbre de décisions où chaque test revient à regarder si une lettre est une consonne ou une voyelle ou à la comparer à une lettre donnée de l'alphabet (FERVERS & AL, 1976). Le nombre de prononciations possibles étant déjà limité pour une lettre, on considère la lettre suivante : suivant son état on réduit éventuellement ce nombre et on recommence de la même façon en analysant au maximum les six lettres suivantes et les trois lettres précédentes, jusqu'à ce qu'il n'y ait plus qu'une seule prononciation possible. Cet arbre de décisions est relativement court : 400 tests au total, de 0 à 10 tests successifs au maximum pour toutes les lettres sauf E (20 tests successifs au maximum) ce qui correspond à un temps de calcul inférieur à 600 µ s par lettre. La décision prise pour une lettre permet éventuellement de sauter l'analyse de la lettre suivante (celle de N et de G dans ETANG par exemple). Ce programme tient compte (imparfaitement) des liaisons. La plupart des règles habituelles de prononciation du français s'y retrouvent sauf celles qui dépendent de la signification d'un mot en fonction de son contexte (problème que nous n'avons pas étudié) et les exceptions pour lesquelles a été introduit un petit dictionnaire (Fig. 1).

GENERATION DE L'INTONATION

La phrase est découpée en parties (proposition) séparées par des silences aux ponctuations et avant les mots ayant une signification grammaticale (conjonctions, pronoms relatifs). Dans chaque subdivision ainsi obtenue, les séparateurs entre mots sont éventuellement transformés en silences brefs pour subdiviser les parties de phrase trop longues. On fixe alors le mouvement intonatif pour chaque groupe séparé par des silences selon leur position dans la phrase (DELATTRE, 1966). A chacun des cinq mouvements intonatifs correspond une courbe qui permet de calculer la période du fondamental pour chaque phonème du groupe séparé par deux silences (FERVERS & AL, 1976).

A ce moment, on a déduit du texte orthographié une suite de phonèmes à prononcer ainsi que la période du fondamental correspondant à chacun d'entre eux. La transcription est très rapide mais le calcul de l'intonation nécessite un programme plus complexe et se fait relativement lentement (recherche dans un dictionnaire de mots grammaticaux, critères de segmentation, calculs assez lourds de l'intonation pour des résultats plutôt banals) (Figure 2).

SYNTHESE

Le synthétiseur est un filtre en treillis du dixième ordre, construit en technologie hybride et programmable (GUEGUEN & AL, 1975). Il permet, en affichant sur ses multiplieurs digitaux-analogiques les valeurs numériques des coefficients de corrélation partielle, d'obtenir la réponse d'un système linéaire (représentant à un instant une forme du canal vocal) à une entrée programmable (suite d'impulsions pour représenter l'excitation due aux cordes vocales et bruit pour les frictions) (Fig 3).

Cet ensemble de 13 paramètres (10 coefficients du filtre K_i, amplitude du bruit B, amplitude V et période N des impulsions) peut être modifié toutes les 10 msec. C'est en les faisant évoluer au cours du temps en fonction des données issues du programme d'analyse du texte orthographié, qu'on obtient la parole synthétique.

Ce synthétiseur étant actuellement géré par un microprocesseur TMS 9900, nous avons choisi d'effectuer la synthèse par une méthode aussi simple et rapide que possible.

Pour chaque phonème de base un jeu de dix coefficients, de corrélation partielle a été calculé et mémorisé (les plosives étant décomposées en deux parties). Chaque phonème dure approximativement 80 ms et la transition d'un phonème à l'autre 40 ms. Pour chaque phonème, le microprocesseur envoie au synthétiseur les 12 paramètres (K, B, V). La période du fondamental N est recalculée toutes les 10 ms par interpolation linéaire entre les deux valeurs données pour le début de chaque phonème par le programme de génération d'intonation, et envoyée au synthétiseur.

Pour éviter une transition trop brusque d'un phonème au suivant, le microprocesseur calcule toujours par interpolation linéaire l'évolution des paramètres nécessaires à la synthèse (K, B, V, N) lors de la transition

d'un phonème au suivant. Le domaine de stabilité du filtre linéaire correspond à des coefficients de corrélation partielle K compris entre -1 et +1 (domaine convexe), l'interpolation linéaire assure sa stabilité lors des transitions et domme pour cette période une évolution raisonnable du spectre lors du passage d'un phonème voisé à un autre (Figure 4).

Cette synthèse par phonème est intelligible et les transitions d'un phonème au suivant semblent tout à fait acceptables sauf pour les plosives : même si la partie voisée et l'explosion sont bien modélisées, le passage de la plosive au phonème suivant est souvent mal reproduit, sauf dans certaines configurations plosive - voyelle. Il est alors difficile de distinguer les plosives les unes des autres dans les logatomes. Nous cherchons une solution à ce problème en introduisant un phonème très bref juste après l'explosion (éventuellement fonction de la plosive et du phonème suivant) et en faisant la transition à partir de ce phonème intermédiaire. Si cette méthode n'améliore pas les résultats nous aurons secours à une forme simplifiée de transition par diphones (LARREUR, EMERARD, 1976).

Dans l'état actuel le microprocesseur synthétise en temps réel la phrase à partir du texte codé sous forme phonétique (il n'y a pas de mémorisations intermédiaires) grâce à un programme de moins de 300 instructions et une mémoire de 600 mots de 16 e.b. où sont conservés les paramètres nécessaires à la synthèse (16 paramètres pour chacun des 37 sons élémentaires).

CONCLUSION

Nous avons présenté une méthode complète de synthèse parmi les plus simples. Elle peut être implantée grâce à des moyens informatiques réduits et une programmation facile. Bien que la qualité de la voix obtenue semble acceptable, elle devrait pouvoir être améliorée sur deux points : une bonne synthèse des plosives et un meilleur "naturel" dans l'intonation obtenue ; domaines où nous ne pouvons guère espérer d'aboutir sans la collaboration de phonéticiens.

Nous menons parallèlement des études analogues pour la transmission de la parole à faible débit : l'analyse de la corrélation du signal vocal pour une durée de 20 ms permet de synthétiser 1 parmi 1000 sons élémentaires dont les paramètres ont été mémorisés. La transmission ne nécessite que l'envoi du numéro correspond à ce son (10 e.b.) au lieu des coefficients de corrélation partielle déduits de l'analyse par prédiction linéaire (49 e.b.) pour un résultat de qualité comparable.

REFERENCES

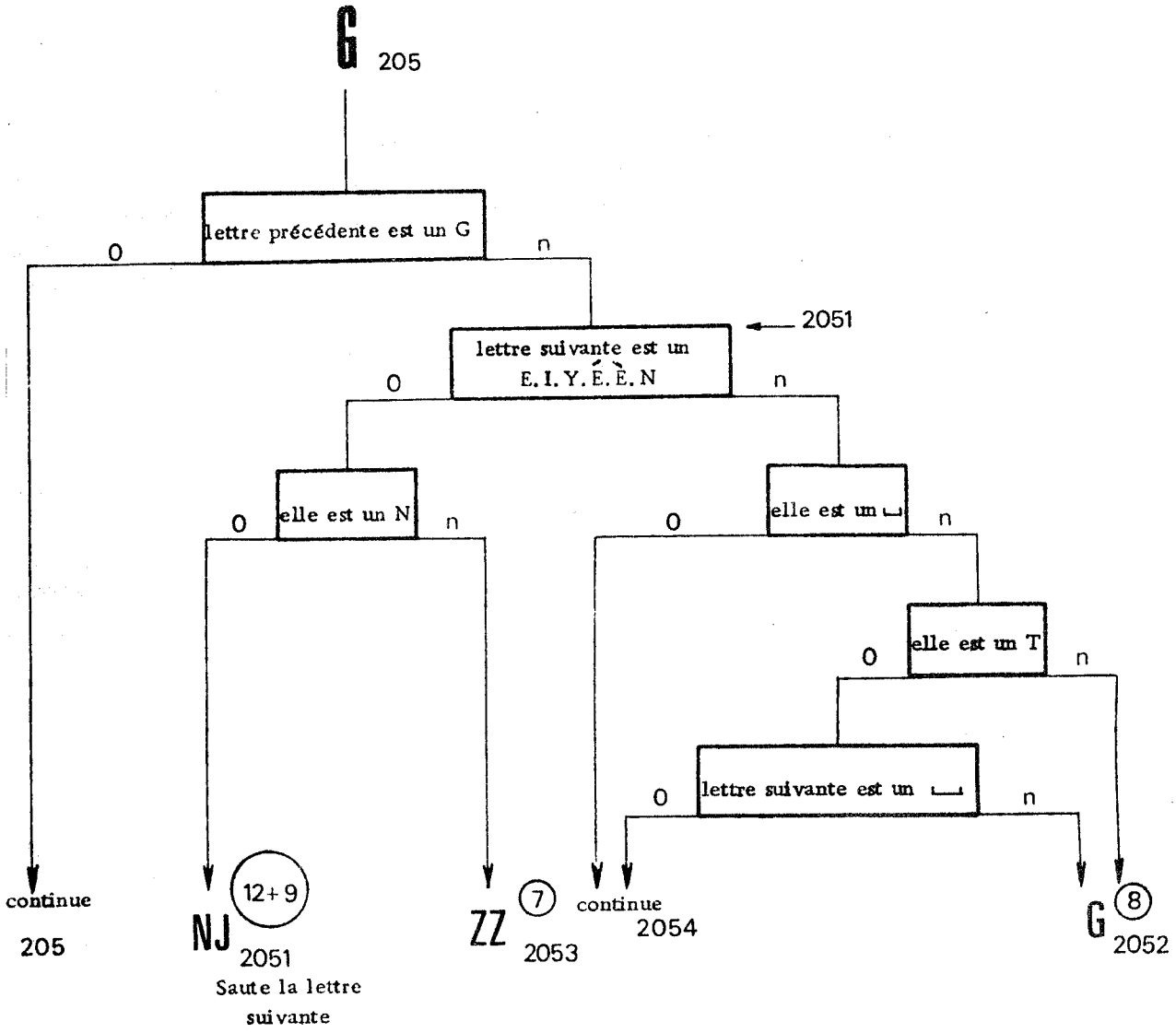
GUEGUEN, C. , LEROUX, J. & AL, 1975, Un synthétiseur à structure programmable ; 6° J.E.P., Toulouse.

MARKEL, J.D , 1976, Linear prediction of speech ; Springer Verlag.

FERVERS, H., & AL, 1976, Programme de transcription orthographique phonémique ; Rapport E.N.S.T.

DELATTRE, P., 1966, Les 10 intonations de base du français ; French review Vol. 40, Baltimore, Octobre et Décembre.

LARREUR, D., EMERARD, F., 1976, Speech synthesis by dyads and automatic intonation processing ; 1976 IEEE international conference on ASSP ; Philadelphie, Avril.



Continue	pas de prononciation
NJ	comme dans Agnès
ZZ	comme dans Gérard
G	comme dans Gaston

Figure 1 : TRANSCRIPTION PHONEMIQUE DE LA LETTRE G
TRANSCRIPTION OF THE CHARACTER G TO A PHONEME

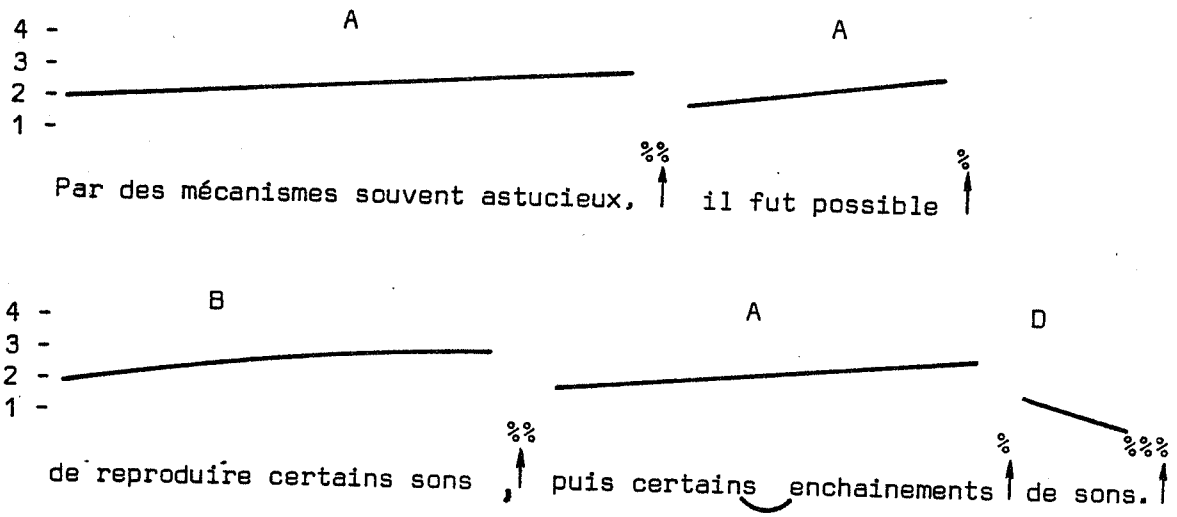


Figure 2 : EXEMPLE DE SCHEMA INTONATIF OBTENU
EXAMPLE OF INTONATION GIVEN BY THE PROGRAM

(il y a cinq schémas possibles A B C D E)

Le signe % indique l'insertion d'un silence
Le signe ◡ indique que la liaison a été prise en compte.

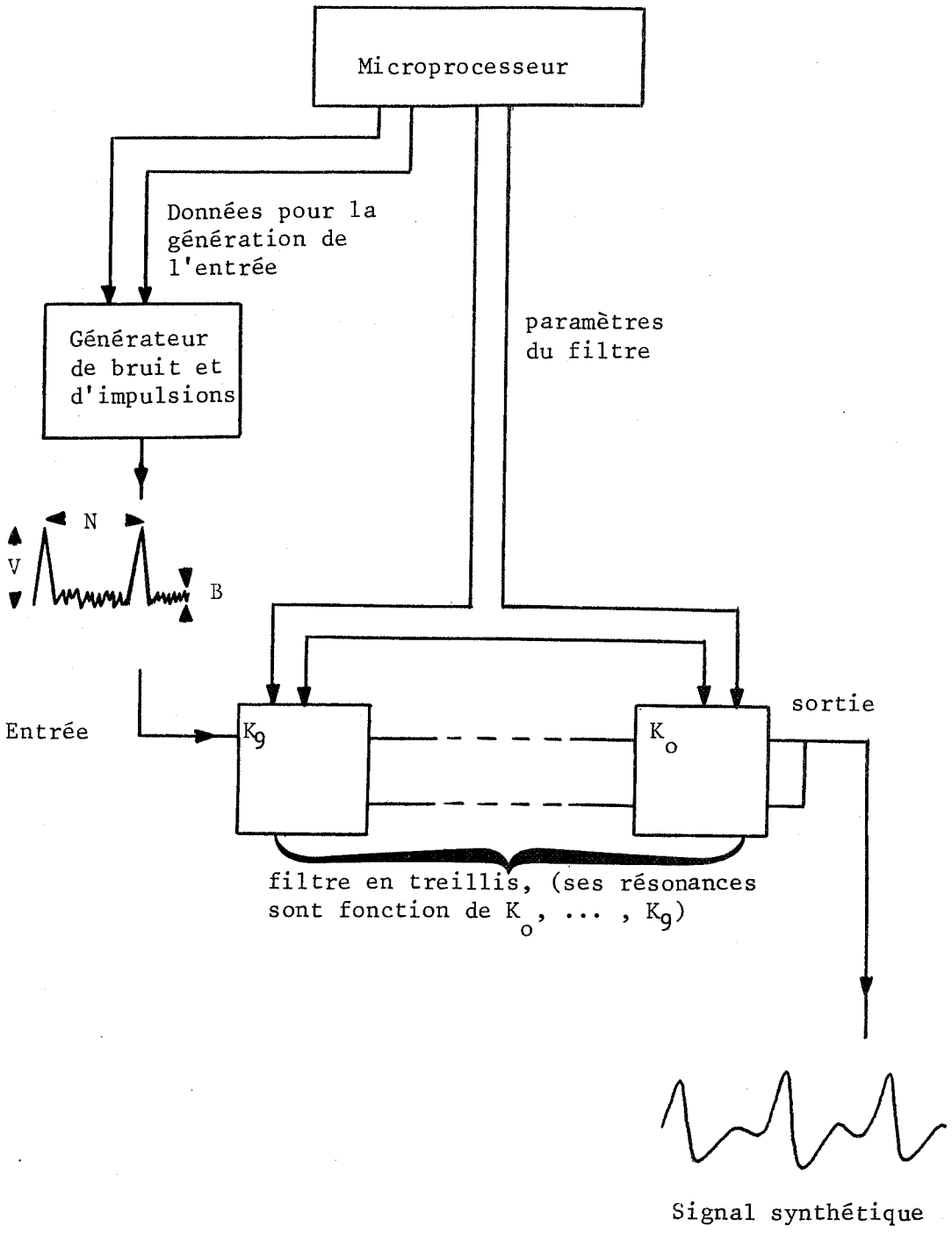


Figure 3 : SYNTHESE DE SON UTILISANT LE CODAGE PREDICTIF
SPEECH SYNTHESIS USING LINEAR PREDICTION

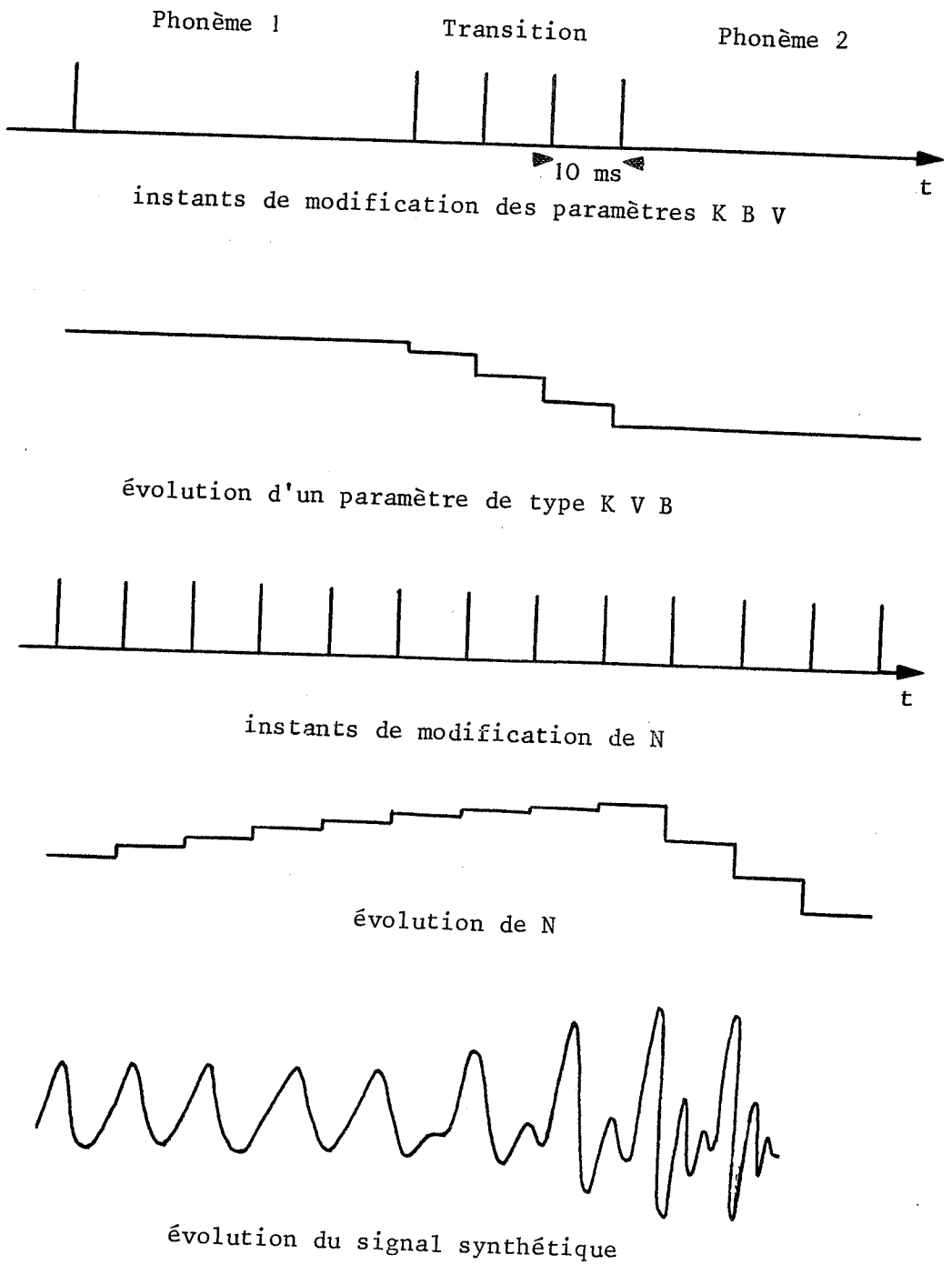


Figure 4 : COMMANDE DU SYNTHETISEUR ET EVOLUTION DU SIGNAL SYNTHETIQUE LORS D'UNE TRANSITION
CONTROL OF THE SYNTHETIZER AND EVOLUTION OF THE SYNTHETIC SIGNAL DURING A TRANSITION

10èmes JOURNÉES D'ÉTUDE SUR LA PAROLE

GRENOBLE - 30 MAI - 1ER JUIN 1979

UN ANALYSEUR SYNTAXIQUE POUR LA SYNTHESE DU TEXTE

Philippe MARTIN

Institut de Phonétique
Université de Provence
Experimental Phonetics Laboratory
University of Toronto

RESUME

Les séquences de paramètres prosodiques nécessaires à la synthèse peuvent être dérivées de la structure prosodique qui elle-même peut être déterminée, dans certaines conditions, à partir de la structure syntaxique de l'énoncé. La génération automatique de ces paramètres pour la synthèse de textes nécessite donc une analyse syntaxique préalable. Les analyseurs actuellement utilisés, basés généralement sur des mécanismes génératifs ou distributionnels, peuvent se révéler trop puissants et trop complexes pour cette application particulière.

On propose ici, pour le calcul de la hiérarchie syntaxique, un procédé plus simple basé sur les caractéristiques de dépendance des catégories syntaxiques. Ce procéeé permet de déterminer directement la structure de dépendance de la phrase et, partant, la ou les structures prosodiques correspondantes. On montre que les propriétés des systèmes de dépendances envisagés sont différentes de celles relatives aux grammaires de dépendance "classiques".

SYNTACTIC ANALYSIS FOR SPEECH SYNTHESIS

Philippe MARTIN

SUMMARY

The prosodic parameters needed to synthesize sentences can be obtained from a prosodic structure which is strongly correlated with a corresponding syntactic hierarchy (MARTIN, 1975). Automatic reading of a text requires thus a syntactic analysis of each sentence. Syntactic parsers, based for example on phrase-structure grammars, are available, but are generally too complex to be used in this particular application.

It is felt that a simpler approach is needed to determine syntactic hierarchies. The method presented here is based on specific dependency relations existing between syntactic categories. These relations are used to determine a dependency structure as well as a prosodic structure of a sentence. Some important properties of dependency systems are described, and are shown to be different from those related to "classical" dependency grammars (cf. HAYS, 1964; GAIFMAN, 1965).

10èmes JOURNÉES D'ÉTUDE SUR LA PAROLE
GRENOBLE - 30 MAI - 1ER JUIN 1979

UN ANALYSEUR SYNTAXIQUE POUR LA SYNTHESE DU TEXTE

Philippe MARTIN

La synthèse de la parole à partir d'un texte, qu'elle soit réalisée à partir de phonèmes, de diphonèmes, ou de mots entiers, exige la génération de paramètres prosodiques relatifs à chaque énoncé synthétisé. Ces paramètres peuvent être assignés manuellement (ex. EMERARD et LARREUR, 1976) ou automatiquement (CHOPPY, LIENARD et TEIL, 1975) selon une procédure liée ou non à l'organisation syntaxique de la phrase. En fait, c'est le rapport de l'intonation à la syntaxe qui semble encore faire problème: cette relation peut en effet tantôt être caractérisée par de fortes similitudes tantôt, par des différences apparemment irréconciliables (VAISSIERE, 1975; DI CRISTO, 1975).

Toutefois, si on a intérêt, dans une perspective théorique globalisante, à considérer a priori la prosodie comme indépendante de la syntaxe pour en déterminer ensuite les contraintes syntaxiques de variation, il reste que, pour des applications telles que la synthèse de la parole, on est libre d'imposer une correspondance étroite entre les hiérarchies syntaxique et prosodique, provoquant par cet homomorphisme une redondance propre à améliorer l'intelligibilité du message synthétisé. La similitude ne devra alors être rompue que dans des cas d'ambiguïté syntaxique ou de structure non-coplanaires, pour lesquels la hiérarchie prosodique et, partant, la séquence des paramètres prosodiques nécessaires à la synthèse, ne pourront être directement dérivées de l'organisation syntaxique de l'énoncé (cf. MARTIN, 1979).

Si on suppose connu le mécanisme de génération des paramètres prosodiques à partir d'une certaine hiérarchie prosodique (EMERARD, 1977; MARTIN, 1975), la synthèse de texte pourra donc se faire, en général, à partir d'une analyse syntaxique de chaque phrase. Or, les analyseurs disponibles réalisant cette opération sont en général très complexes, et, donc, coûteux à utiliser, en ce qu'ils fournissent des solutions trop "riches" pour l'application envisagée, pour laquelle seule la connaissance de la hiérarchie syntaxique est nécessaire en dehors, par exemple, de toute spécification de catégorie syntaxique.

On s'attachera ici à circonvenir les caractéristiques d'une classe d'analyseurs spécifiquement adaptés à la recherche de la hiérarchie prosodique et des paramètres acoustiques qui en assurent l'indication. Répondant à des besoins différents de ceux relatifs à la reconnaissance automatique, ils pourront se révéler plus simples et donc plus faciles à mettre en oeuvre.

GRAMMAIRE DE DEPENDANCE

La ou les structures prosodiques relatives à un énoncé donné peuvent, en général, être obtenues à partir de la hiérarchie syntaxique, c'est-à-dire du classement hiérarchique des unités syntaxiques minimales que sont les mots. Pour déterminer cette hiérarchie, on peut se baser sur les relations de dépendance nécessairement contractées par les unités et qui définissent implicitement une structure de dépendance.

Les grammaires de dépendance étudiées par divers auteurs (ex. HAYS, 1964; GAIFMAN, 1965) sont en général de type génératif, en ce qu'ils rendent compte, par un ensemble de règles de réécriture, de mécanismes de production d'énoncés bien formés. Les axiomes qui les définissent sont, par exemple (ROBINSON, 1967):

Dans une séquence de n éléments d'une phrase:

a) un élément et un seul est indépendant;
b) chacun des autres éléments dépend directement d'un autre;
c) aucun élément ne dépend directement de plus d'un autre;
d) si un élément A dépend directement d'un élément B, et qu'un élément C se trouve placé entre A et B dans une séquence, alors C dépend directement de A ou de B ou d'un autre élément intermédiaire qui se trouverait entre A et B.

Ces axiomes portent donc sur une relation binaire de dépendance r , contractée entre les couples d'éléments ou de groupes d'éléments de la séquence. Plus précisément, une telle relation peut être:

symétrique et positive: A dépend de B et B dépend de A (solidarité). Notation: $A \longleftrightarrow B$
symétrique et négative: A ne dépend pas de B et B ne dépend pas de A (combinaison). Notation: A B
asymétrique : A dépend de B mais B ne dépend pas de A (sélection). Notation: $A \longrightarrow B$

Les relations de dépendance sont transitives:

si A r B et B r C, alors A r C, pour r : $\longleftrightarrow$, $\longrightarrow$ et $\longleftarrow$

On peut admettre de plus qu'elles sont réflexives:
Pour toute unité syntaxique X, on a X r X, c'est-à-dire qu'un élément dépend et est solidaire de lui-même.

Au lieu d'une perspective "globale" relative à une grammaire dans laquelle les unités minimales sont dérivées à partir de divisions successives de la phrase de départ, on adoptera ici un point de vue "local", privilégiant les propriétés de dépendance de chaque élément de la séquence. Ceci va conduire à une révision des contraintes déterminées par les axiomes des grammaires de dépendance, et à l'obtention d'un système décrivant les propriétés de dépendance de classes d'éléments syntaxiques. Ce système permettra alors de déterminer la structure de dépendance d'un énoncé donné à partir des propriétés de chaque classe présente.

REVISION DES AXIOMES

Dans une séquence telle que

Marie et Alexandre

l'examen des propriétés de dépendance de chaque élément peut conduire à remettre en question les 3 premiers axiomes cités plus haut. Dans cet énoncé, en effet, la conjonction (et) présente une double relation de sélection, à droite envers l'élément (Alexandre), et à gauche par rapport à l'élément (Marie). Le réseau de dépendances peut être considéré globalement, selon les groupes éventuellement créés:

Marie ⟵ (et ⟶ Alexandre)

ou localement, d'après les relations contractées par chaque élément indépendamment de la formation d'unités plus grandes:

Marie ⟵ et ⟶ Alexandre

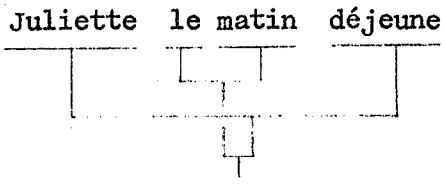
Dans les 2 cas, un seul élément, (Marie), est effectivement indépendant dans la séquence, mais la dépendance de l'unité (Alexandre) est indirecte et réalisée par l'intermédiaire de la conjonction. Ceci contredit donc les axiomes b) et c):

- si un élément et un seul est indépendant dans la séquence, chacun des autres éléments dépend directement ou indirectement d'un autre;
- un élément peut dépendre directement de plus d'un autre.

Soit d'autre part l'énoncé

Juliette le matin déjeune

dans lequel le verbe sélectionne le nom propre, et où le groupe d'éléments solidaires (le matin) ne dépend pas des autres éléments. La structure de dépendance obtenue est donc non-connexe, puisque le réseau de dépendance positive ne lie pas toutes les unités entre-elles. De plus, la hiérarchie est non-coplanaire, en ce que l'arbre qui la représente ne peut avoir ses branches disposées dans un même plan



Ce type de configuration montre que les axiomes a) et d) ne peuvent être conservés:

- plus d'un élément peut être indépendant dans une séquence; s'il y en a plus d'un, la hiérarchie de la séquence résulte de la juxtaposition de ces éléments et des éléments qui en dépendent;

- un élément C situé entre 2 éléments contractant un rapport de dépendance ne dépend pas nécessairement d'un autre élément.

Une relation de dépendance indirecte est une relation qui peut être déduite de la transitivité, par opposition à une relation directe qui ne résulte pas de cette propriété. Cette distinction permet de décrire des relations asymétriques (sélections) pouvant exister entre des groupes d'éléments en fonction des liens existant entre les éléments composants.

Dans l'énoncé

Il mange le gigot froid

la dépendance de l'adjectif (froid) peut se faire par rapport au groupe (il mange le gigot). Le réseau de dépendance est donc:

((il ⟷ mange) ⟵ (le ⟷ gigot)) ⟵ froid

Sélectionnant le groupe, l'adjectif dépend de chacun des éléments (il mange) et (le gigot)

(il ⟷ mange) ⟵ (le gigot) (froid)

La relation de l'adjectif au verbe est alors une dépendance directe, rendant compte de relation vis à vis du groupe, par opposition au réseau

(il ⟷ mange) ⟵ (le ⟷ gigot) ⟵ (froid)

où la relation adjectif-verbe est une relation indirecte résultant de la transitivité.

Un cas de même nature, mais dans lequel un élément est cette fois doublement sélectionné, se présente dans

(après) ⟵ [(vinrent) ⟶ (les ⟷ ennemis)]

où l'ordre verbe-sujet dépend de la présence de l'élément (après) avant le groupe. Ce réseau peut alors être représenté par des relations directes

après vinrent ⟶ les ⟷ ennemis

relations contractées par des éléments minimaux où des groupes d'éléments solidaires.

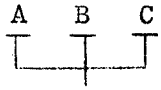
CONTRAINTES DE PROXIMITE A DROITE

Une hiérarchie syntaxique relative à une séquence de n éléments est donc déterminée par un réseau de $\frac{n}{2}(n-1)$ relations de dépendance binaire contractées par chaque couple d'éléments. Ces relations peuvent être asymétriques ou symétriques positives ou négatives (correspondant à une absence de dépendance).

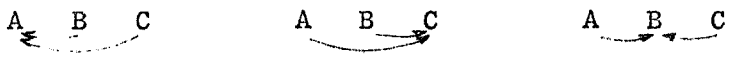
Les relations symétriques déterminent des hiérarchies où plus de 2 éléments peuvent être regroupés en un seul niveau. Ainsi, les réseaux

A ⟷ B ⟷ C et A B C

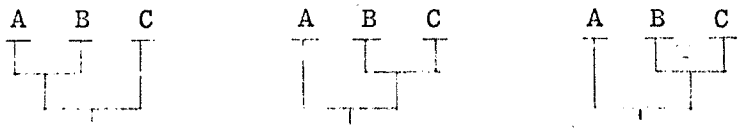
correspondent à la même hiérarchie



Par contre, on peut contraindre les relations asymétriques à n'indiquer que des hiérarchies dans lesquelles chaque groupe est nécessairement formé de 2 éléments plus petits. Il en résulte que des réseaux tels que

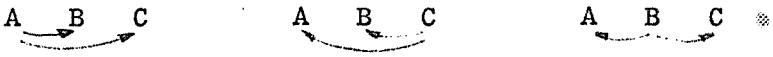


correspondent à des hiérarchies organisées en 2 niveaux, soit respectivement



L'élément le plus proche, ou le plus proche à droite, est donc utilisé pour former une unité plus grande en un premier niveau de regroupement. Ce mécanisme constitue une première contrainte imposée au système de dépendance.

De même, pour des réseaux où un élément dépend directement de plus d'un autre élément, comme



on peut imposer au mécanisme d'indication des hiérarchies de regrouper en priorité les éléments les plus proches, ou les plus proches à droite.

On peut remarquer que cette contrainte de proximité à droite ne pourra pas être appliquée à un système de dépendance de type prosodique.

DENOMBREMENT

Le nombre de réseaux de dépendance distincts que peut présenter une séquence de n unités est égal à $2^{n\ (n-1)}$ (n/2 (n-1) relations de 4 types distincts.) Ces réseaux correspondent à des hiérarchies dont le nombre n'est apparemment pas encore connu pour n quelconque (BARBUT et MONJARDET, 1970).

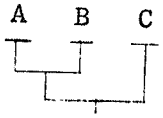
Toutefois, le nombre C_n de hiérarchies coplanaires est donné par la formule de récurrence:

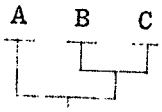
$$(n+1)\ C_{n+1} = 3\ (2n-1)\ C_n - (n-2)\ C_{n-1} \text{ pour } n > 1$$

Les différentes hiérarchies (coplanaires et non-coplanaires) indiquées par (n-1) relations antisymétriques de sélection correspondent aux différentes chaînes que l'on peut dénombrer dans un treillis de partitions. Ces chaînes ne prennent donc pas en compte les hiérarchies partiellement ou totalement indiquées par des relations symétriques, où des regroupements peuvent concerner plus de 2 unités à la fois.

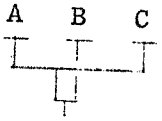
Le nombre de hiérarchies syntaxiques (coplanaires et non-coplanaires) et de hiérarchies prosodiques (toujours coplanaires) sera donc, en général, différent.

Ainsi, la séquence de 3 mots A, B et C peut être organisée selon 4 hiérarchies différentes:

(1) 

(2) 

(3) 

(4) 

Les hiérarchies (1) (2) et (3) sont coplanaires, alors que (4) est non coplanaire; (1), (2) et (4) correspondent à des réseaux de dépendances asymétriques, alors que (3) est indiqué par des relations symétriques (ou par 1 réseau de sélection équivalent, tel que A $\longrightarrow$ B $\longrightarrow$ C, équivalent à A $\longleftrightarrow$ B $\longleftrightarrow$ C).

Une séquence de 3 unités peut présenter $2^{3(3-1)} = 64$ réseaux de dépendance différents, qui ne peuvent donc indiquer que 4 hiérarchies possibles.

APPLICATIONS

Les considérations qui précèdent constituent le cadre général dans lequel doivent s'étudier les systèmes spécifiques pour une langue donnée. Chaque élément syntaxique est susceptible de contracter une ou plusieurs relations de dépendance, et cette propriété est locale, c'est-à-dire qu'elle ne concerne que l'élément lui-même et non le contexte.

Le calcul du réseau relatif à une séquence donnée suppose donc qu'on puisse assigner à chaque élément, ou à chaque classe d'éléments, une propriété de dépendance envers une ou plusieurs autres classes, la réalisation effective d'une relation (symétrique ou antisymétrique) étant liée à la présence de cette ou de ces autres classes dans la séquence.

L'établissement des classes et de leurs propriétés pourra se faire

à partir d'un corpus non fermé, constitué d'énoncés dont on aura établi les structures de dépendance. Les classes choisies au départ correspondront par exemple aux catégories syntaxiques connues: Article, Nom, Adjectif, Nom propre, Verbe, Auxiliaire, Adverbe, Conjonction, etc.

Les relations de dépendances sont a priori directionnelles. Ainsi, la catégorie verbe sélectionne à gauche un nom, un nom propre ou un pronom, alors qu'un article sélectionne un nom ou un adjectif à droite. La dépendance entre éléments peut être obligatoire (entre un nom et un verbe par exemple) ou facultative (entre un verbe transitif et un nom).

Un énoncé simple tel que

Le frère de Juliette a acheté un mauvais livre

fait apparaître les différentes classes d'éléments Article, Nom, Préposition, Nom propre, Auxiliaire, Verbe, Adjectif, pourvues des propriétés suivantes:

Art $\longrightarrow$ (Nom)	: l'article sélectionne obligatoirement un nom placé à sa droite;
(Art) $\longleftarrow$ Nom	: un nom sélectionne obligatoirement un article situé à sa gauche;
( ) $\longleftarrow$ Prép $\rightarrow$ ( )	: une préposition sélectionne les éléments ou groupes d'éléments placés à sa gauche et à sa droite;
Aux $\longleftrightarrow$ (V)	: un auxiliaire est solidaire du verbe à sa droite;
(N) $\longleftarrow$ V	: un verbe sélectionne un nom à gauche
Adj $\longrightarrow$ (N)	: un adjectif sélectionne un nom placé à sa droite
(N) $\longleftarrow$ Adj	ou à sa gauche
(V_{tr}) $\longleftarrow$ N	: un nom sélectionne facultativement un verbe transitif situé à gauche

L'examen d'un deuxième énoncé comme

Juliette est gentille

conduit à modifier les propriétés citées:

$\left\{\begin{matrix}(N_p)\\(N)\end{matrix}\right\} \longleftarrow$ V	: un verbe sélectionne à gauche un nom ou un nom propre
$V_{être} \longleftrightarrow$ (Adj)	: un verbe (être) est solidaire d'un adjectif qui serait placé à sa droite.

L'analyse d'un corpus doit alors faire apparaître une stabilisation dans la liste des propriétés, considérées comme étant suffisantes pour le calcul du réseau de dépendance, et donc de la structure, d'un énoncé donné. Ce calcul se fait alors en établissant les relations satisfaisant aux propriétés de chaque classe des éléments de la séquence. L'énoncé étant a priori bien formé, les dépendances effectivement réalisées pourront être établies, et la hiérarchie correspondante déterminée.

REFERENCES

BARBUT, M. et MONJARDET, B. (1970) Ordre et classification, Hachette, Paris

CHOPPY, C. LIENARD, J-S. et TEIL, D. (1975) Un algorithme de reconnaissance automatique sans analyse syntaxique, Actes des 6èmes JEP, Toulouse, 387-395

DI CRISTO, A. (1975) Recherches sur la structuration prosodique de la phrase française, Actes des 6èmes JEP, Toulouse, 94-116

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

EMERARD, F et LARREUR, O. (1976) Synthèse par diphones, Recherches/ Acoustique, CNET, Lannion, III, 293-314

GAIFMAN, H. (1965) Dependency Systems and Phrase-Structure Systems, Information and Control, (8) 304-337

HAYS, D.G. (1964) Dependency Theory: a Formalism and Some observations, Language, (40) 511-525

MARTIN, Ph. (1975) Analyse phonologique de la phrase française, Linguistics, (146), 37-67

MARTIN, Ph. (1979) L'intonation de la phrase italienne, Studi di Grammatica Italiana, à paraître

ROBINSON, J.J. (1967) Dependency Structures and Transformational Rules, Language (46) n°2, 259-287

VAISSIERE, J. (1975) Caractérisation des variations de la fréquence du fondamental dans les phrases françaises, Actes des 6èmes JEP, Toulouse, 39-50

10èmes JOURNÉES D'ÉTUDE SUR LA PAROLE

GRENOBLE - 30 MAI - 1ER JUIN 1979

GENERATION AUTOMATIQUE DE PHRASES EN PHONETIQUE A PARTIR DE FORMULES SEMANTIQUES.

D.MEMMI - J.S.LIENARD

LIMSI B.P. 30 91406 ORSAY

RESUME

A partir d'énoncés sémantiques donnés dans une formulation de type logique, on engendre par programme les phrases françaises parlées correspondantes, en notation phonétique pour commander un synthétiseur de parole.

On effectue ainsi une traduction directe du sens à la forme sonore, pour constituer un maillon de la chaine de réponse d'un système de dialogue homme-machine par la parole.

Dans une telle approche où on ne passe pas par l'écrit, le découpage en groupes prosodiques est immédiat. Mais il a fallu réaliser une étude grammaticale du français parlé, qui diffère sensiblement de la forme écrite que décrivent d'habitude les grammaires.

COMPUTER GENERATION OF PHONETIC SENTENCES FROM SEMANTIC FORMULAS.

D.MEMMI - J.S.LIENARD

SUMMARY

In order to hold a dialog with a computer system, the computer should be able to generate new messages expressing any meaning without having to record them all beforehand.

The program described here reads semantic formulas as input and can then generate the corresponding spoken sentences in French. These output sentences are produced directly in phonetic code, prosody included, so as to drive a speech synthesizer.

This direct translation from meaning to spoken form is part of a global research project to achieve man-machine communication, together with a speech recognition system.

There is no reference to the written language in this program, and the prosodic groups are then evident. On the other hand it was necessary to work out an original linguistic analysis for the spoken language, while only the written form is usually described.

10èmes JOURNÉES D'ÉTUDE SUR LA PAROLE
GRENOBLE - 30 MAI - 1ER JUIN 1979

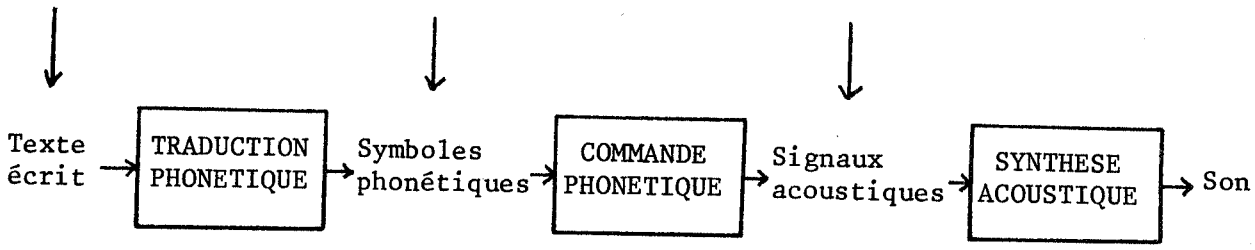
GENERATION AUTOMATIQUE DE PHRASES EN PHONETIQUE A PARTIR DE FORMULES SEMANTIQUES

D. MEMMI - J.S. LIENARD

INTRODUCTION

La synthèse de la parole peut être considérée avec trois entrées différentes :

- signaux acoustiques (spectre et fondamental),
- symboles phonétiques et marqueurs prosodiques,
- texte écrit.

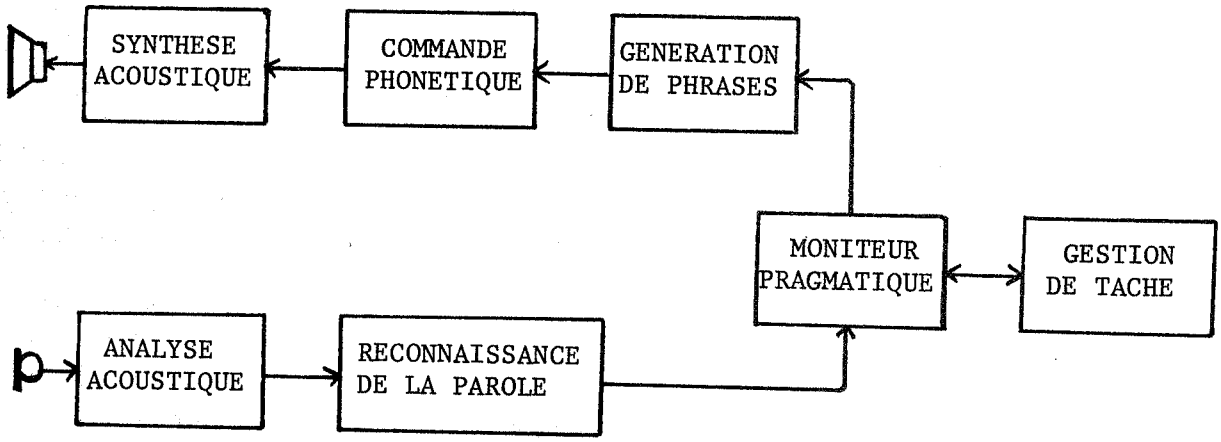


Il reste de nombreux problèmes à résoudre dans chacun de ces trois maillons. L'un des plus difficiles est l'élaboration d'une prosodie "naturelle" à partir d'un texte écrit. La raison en est que la prosodie fait appel à des connaissances de heut niveau : syntaxe (car le texte écrit possède généralement une forte structuration syntaxique), sémantique (on parle indifféremment de "groupes de sens" ou de "groupes prosodiques" ; de plus toute émission de parole possède une certaine expression ou intention de nature sémantique), pragmatique (la conduite de la prosodie dépend de la situation actuelle, du contexte connu simultanément par le locuteur et l'auditeur).

Nous avons déjà émis l'idée que la syntaxe n'était qu'un facteur parmi d'autres dans la conduite de la prosodie (CHOPPY, LIENARD et TEIL, 1975), dans la mesure où la syntaxe reflète des structures sémantiques et pragmatiques sous-jacentes. Nous sommes encore incapables de réaliser une analyse sémantique complète des textes écrits ; de plus nous souhaitons séparer nettement la langue écrite de la langue parlée, qui n'a pas le même domaine d'application ni la

même fonction (LIENARD, 1977).

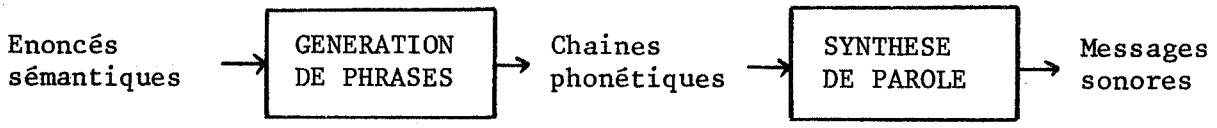
C'est pourquoi nous proposons de partir de messages sémantiques dans lesquels l'information de haut niveau est une donnée, et de transformer ces messages en parole synthétique, moyennant la connaissance d'un lexique et d'une syntaxe choisis dans un sous-ensemble de la langue naturelle. Ces messages sémantiques, pour l'instant élaborés par un opérateur, sont destinés à être ultérieurement construits par un processus d'intelligence artificielle utilisant sans doute un réseau sémantique et des possibilités d'inférence (voir par exemple COULON, KAYSER et al., 1977) ; ceci dans le cadre du projet global de dialogue oral homme-machine du L.I.M.S.I. :



Le programme proposé effectue une transformation directe du message sémantique en une chaîne phonétique accompagnée de marqueurs prosodiques, sans passer par la forme écrite. Nous aurions pu faire une génération de messages écrits, comme par exemple (WINOGRAD, 1972) et ensuite utiliser nos programmes de traduction phonétique et prosodique (voir TEIL, 1975 et PROUTS, 1979). Mais nous avons choisi d'effectuer tout le travail en phonétique pour plusieurs raisons : d'une part la forme orale du message ne doit pas être subordonnée à la forme écrite ; d'autre part on peut imaginer que certaines réalisations de communication homme-machine ne fassent pas du tout usage de la forme écrite, et dans ce cas la présence d'un module de traduction phonétique ne se justifie pas.

Ensuite il était intéressant, sur le plan purement linguistique, d'examiner et de formaliser la grammaire de la langue orale.

Donc à partir d'énoncés sémantiques donnés dans une formulation de type logique, on engendre les phrases françaises correspondantes, en notation phonétique afin de commander un synthétiseur de parole. Ce programme est écrit en PL/1 et s'intègre au système de dialogue parlé en cours d'élaboration au LIMSI, dont il constitue une partie de la branche de réponse :



DONNEES DU PROGRAMME

Il faut d'abord lire une première fois des entrées lexicales (en ordre quelconque) pour construire un dictionnaire donnant, pour chaque notion sémantique, le mot qui l'exprime en français avec sa prononciation et toutes ses caractéristiques morphologiques et syntaxiques nécessaires à la construction de la phrase où on l'utilise :

Ex : SOUPE Nom Régulier Féminin SOUP
AIMER Verbe 1er groupe Transitif AIM
(sens, catégorie, morphologie, syntaxe, prononciation)

Puis en entrée on lit des énoncés sémantiques formulés en notation fonctionnelle, c'est-à-dire préfixée mais avec des parenthèses pour ne pas avoir à connaître à l'avance le nombre d'arguments d'un prédicat : $F(x_1, x_2, \ldots, x_n)$.

Cette notation logique bien connue permet de représenter toute relation lexicale ou grammaticale pour exprimer le sens d'un message :

Ex : AIMER(DEF(JEUNE(ENFANT)), DEF(SOUPE))
("Le jeune enfant aime la soupe").

Noter qu'à ce stade les éléments de la formule représentent des notions sémantiques et non des mots du français. Mais les éléments utilisés sont proches des mots de la langue naturelle et on n'a pas essayé de définir des primitives plus simples que les mots ordinaires, sauf pour les notions grammaticales

(négation, pluriel...). D'ailleurs on aurait pu aussi bien choisir une autre représentation plus ou moins équivalente comme une notation par cas donnant explicitement la fonction (agent, objet....) des arguments. Voir (E. CHARNIAK & Y. WILKS, 1976) par exemple pour les représentations possibles.

RESULTATS DU PROGRAMME

On obtient en sortie la phrase en français parlé exprimant le sens de la formule sémantique entrée. Le résultat est donc une suite de symboles phonétiques et de marqueurs prosodiques donnant toute l'information nécessaire à la prononciation de cette phrase par un synthétiseur de parole. Le code utilisé est choisi en fonction de l'équipement.

Ex : 'LEJEN*F*-(M$LASWP'
("Le jeune enfant aime la soupe").

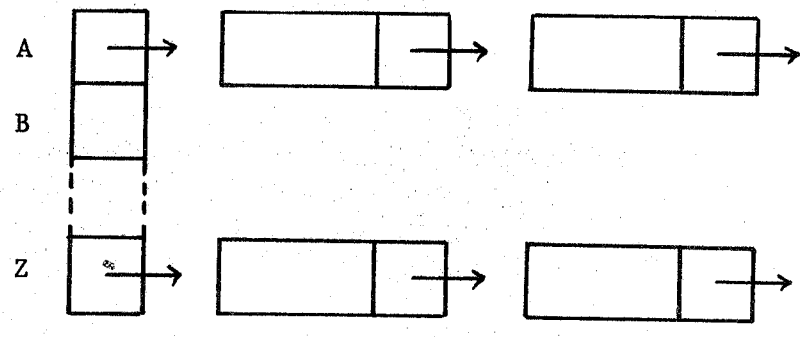
Le découpage en groupes prosodiques par des marqueurs permet de calculer aisément la prosodie de la phrase sans avoir à faire d'analyse supplémentaire, puisque les marqueurs sont insérés dès la construction de la phrase phonétique.

ANALYSE DU PROGRAMME

On peut distinguer trois parties successives : la construction d'un dictionnaire, la lecture des énoncés sémantiques, puis la génération proprement dite à l'aide d'une grammaire orale.

1. Dictionnaire

Les entrées lexicales lues dans un ordre quelconque sont rangées comme dans un dictionnaire ordinaire : par ordre alphabétique à chacune des lettres de l'alphabet, donc en 26 sections. Les entrées sont en effet chaînées en liste afin de pouvoir modifier le dictionnaire à tout moment :



Il faut noter que les entrées lexicales comportent en plus un pointeur vers d'éventuelles formes irrégulières, ce qui fait gagner de la place sur les formes régulières.

On n'a pas essayé de réaliser un dictionnaire important, mais on a veillé à ce que la structure soit extensible pour pouvoir ajouter au besoin n'importe quel mot, même irrégulier. Il est ainsi facile de s'adapter à un domaine nouveau.

2. Arborescence sémantique

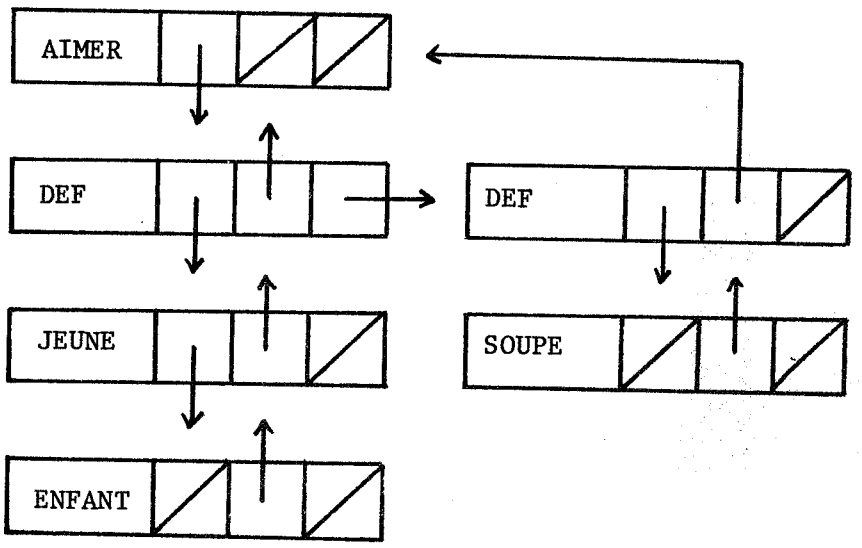
On lit la formule sémantique d'entrée, et on construit en structure de liste le graphe correspondant, qui est une arborescence. La représentation en machine de cette arborescence suit la convention "fils aîné - frère cadet", qui permet un nombre quelconque de successeurs à chaque noeud (c'est-à-dire d'arguments par prédicat).

Puis on parcourt l'arborescence ainsi formée pour l'imprimer et vérifier qu'elle est bien formée.

Ex : soit la formule sémantique :

AIMER(DEF(JEUNE(ENFANT)), DEF(SOUPE))

L'arborescence correspondante sera donc :



3. Morphologie et syntaxe

Cette dernière partie, qui va travailler à partir du dictionnaire et de l'arborescence, est la plus complexe car elle représente théoriquement une grammaire du français oral.

Or l'orthographe française, très conservatrice, est nettement éloignée de la prononciation réelle, et les grammaires ne traitent traditionnellement que du français écrit, malgré quelques tentatives plus récentes (comme RIGAULT, 1971). Il a donc fallu faire un travail original pour mettre au point les règles orales, surtout pour la morphologie. Par exemple des formes verbales qui diffèrent à l'écrit ("aime, aimes, aiment") ont une prononciation identique, et le féminin oral de nombreux adjectifs se forme en ajoutant une consonne ("peti(t), petit(e)") et non en ajoutant un "e" !

On fait d'abord une première passe sur l'arborescence pour consulter le dictionnaire et déterminer les genres, nombres et personnes nécessaires aux accords. En effet on ne peut pas accorder un adjectif ou un verbe par exemple avant d'avoir trouvé le nom ou pronom auquel ils se rapportent ; on empile donc les indicateurs nécessaires.

Ceci fait, on parcourt à nouveau l'arborescence. On effectue alors les accords morphologiques en dépilant les indicateurs précédemment déterminés, et on place les mots accordés en une nouvelle liste selon les règles de la syntaxe. Tout ceci est bien sûr accompli grâce aux indications du dictionnaire sur les particularités de chaque mot.

Par exemple DEF (BON(SOUPE)) donnera "La bonne soupe", mais

DEF (BON(VIN)) deviendra "Le bon vin" et on placera le verbe conjugué à la bonne personne après son sujet :

AIMER (NOUS, DEF(VIN)) donne "Nous aimons le vin", ou plutôt l'équivalent oral.

Enfin on assure le traitement des liaisons, élisions et contractions, qui ne dépendent que de l'ordre final des mots dans la phrase. Ainsi on aura "les garçons" mais "les_hommes", "la soupe" mais "l'eau", et "aux" au lieu de "à les", toujours sous forme orale.

Il ne reste plus qu'à imprimer le résultat phonétique et à le transmettre au synthétiseur qui assemblera les formes acoustiques et calculera la prosodie pour prononcer la phrase.

CONCLUSIONS

Dans son état actuel, le programme peut générer des phrases simples (du type sujet-verbe-complément) affirmatives, négatives, interrogatives et interro négatives au présent. Sujets et compléments comprennent article, adjectif et nom, ou pronom. Le verbe peut être transitif ou intransitif, ou bien avoir un objet indirect.

Ce n'est bien sûr qu'une petite partie de la grammaire du français parlé, qui contient un très grand nombre de règles. Mais c'est un noyau permettant de générer des messages simples mais parfaitement normaux, car la syntaxe est naturelle et la morphologie couvre les formes irrégulières.

Et il s'agit d'un travail en cours qui sera poursuivi plus avant. Des règles plus nombreuses devront être introduites dans la grammaire du programme. En particulier on envisage de traiter d'autres temps des verbes, et une syntaxe plus complexe, ce qui demandera aussi une modification des règles déjà existantes.

Ceci dit, on voit que la parole pose des problèmes particuliers, généralement méconnus, et qui sont loin d'être tous résolus. En particulier il faudrait une synthèse de très bonne qualité pour expérimenter avec les profils prosodiques, car on ne sait pas encore très bien comment les déterminer. Et cela serait d'autant plus vrai si on voulait donner à la parole produite une certaine expressivité, comme le doute ou l'emphase.

Enfin il faudra s'attaquer au problème de la génération des énoncés sémantiques eux-mêmes en fonction du dialogue, sans doute par déduction et inférence logico-sémantiques. En complétant ainsi la chaîne de réponse, on pourra alors envisager de dialoguer par la parole avec l'ordinateur, au moins dans un domaine donné.

REFERENCES

CHARNIAK, E. & WILKS, Y. (eds), 1976, Computational Semantics, North-Holland.

CHOPPY, C., LIENARD, J.S. & TEIL, D., 1975, Un algorithme de prosodie automatique sans analyse syntaxique, 6e J.E.P., Toulouse.

COULON, D., KAYSER, D. et al., 1977, Description générale d'un système de réponse aux questions, C.R.I.N. 77-R-047.

LIENARD, J.S., 1977, Les processus de la communication parlée, Masson, Paris.

PROUTS, B., 1979, Traduction phonétique de texte écrit en français, 10e J.E.P., Grenoble.

RIGAULT, A. (ed), 1971, La grammaire du français parlé, Hachette.

TEIL, D., 1975, Conception et réalisation d'un terminal à réponse vocale, Thèse de Docteur-Ingénieur, Paris VI.

WINOGRAD, T., 1972, Understanding natural language, Academic Press.

10èmes JOURNÉES D'ÉTUDE SUR LA PAROLE

GRENOBLE - 30 MAI - 1ER JUIN 1979

TRADUCTION PHONETIQUE DE TEXTES ECRITS EN FRANCAIS

PROUTS Bernard

L.I.M.S.I. du C.N.R.S. B.P. 30 91406 ORSAY CEDEX

RESUME

Cette communication présente un programme effectuant simultanément la traduction graphème-phonème et la mise en place automatique de marqueurs prosodiques pour des textes écrits en français.

Ce programme repose sur l'exploitation d'un dictionnaire d'environ 1000 éléments. A chaque élément (règle ou exception) sont associées une ou plusieurs des informations suivantes: traduction phonétique, possibilités, obligations ou interdictions de liaison aval ou amont, rôle prosodique.

L'insertion, la suppression ou la modification d'éléments est rendue très simple par l'utilisation de programmes développés à ces fins.

Ce programme actuellement implanté sur un système microprocesseur INTEL 8080 occupe moins de 16 K octets et permet la transcription en temps réel (0,3 s pour une phrase de 100 caractères).

PHONETIC TRANSLATION OF TEXTS WRITTEN IN FRENCH

PROUTS Bernard

SUMMARY

This paper presents a program that performs both graphemic-phonetic translation and automatic prosody-mark setting of texts written in French.

The principle of the program is based on the use of a dictionary comprising about 1000 elements. Each of them (a rule or an exception) is associated with one or several of the following items : its phonetic translation, its prosodic characteristic, the possible, obligatory or forbidden initial or final liaisons.

Routines that have been specially developed, easily allow to insert, delete or modify any element.

This program implemented on an INTEL 8080 microprocessor system occupies less than 16 K bytes and performs real-time transcription (0.3 s for a 100-character sentence).

10èmes JOURNÉES D'ÉTUDE SUR LA PAROLE
GRENOBLE - 30 MAI - 1er JUIN 1979

TRADUCTION PHONETIQUE DE TEXTES ECRITS EN FRANCAIS

PROUTS Bernard

INTRODUCTION

Le L.I.M.S.I. mène actuellement une étude sur la conception et la réalisation d'un système autonome de synthèse du français à base de microprocesseur. Ce système accepte en entrée une chaîne orthographique; la première étape du traitement consiste en une transcription graphème-phonème, la chaîne phonétique ainsi créee constituant l'entrée du module synthétiseur.

Le problème de la transcription graphème-phonème a déja été étudié au L.I.M.S.I. (TEIL, D., 1969); cette étude donna lieu à la réalisation d'un programme implanté sur le synthétiseur Icophone 5 en 1971.

Ce programme possédait quelques lacunes que sa structure arborescente rendait difficile à corriger. Depuis, plusieurs programmes de transcription ont vu le jour en FRANCE; le dernier en date, à notre connaissance, étant celui de DIVAY et GUYOMARD (1977). La description des règles est faite dans un langage spécial développé pour cet usage; le taux d'erreur est faible; la transcription est effectuée à raison d'une moyenne de 20 mots/s sur IRIS 80.

Nous avons cherché une méthode permettant d'une part une modification aisée des règles et d'autre part une implantation sur petite machine. La structure que nous avons retenue présente en outre l'avantage de s'adapter parfaitement à une nouvelle fonction : la mise en place automatique des marqueurs prosodiques par une méthode lexicale.

PRINCIPE DE LA TRANSCRIPTION

Le principe, très simple, repose sur l'exploitation d'un dictionnaire.

Tout d'abord, l'insertion dans le texte à traduire de marqueurs de liaison et de marqueurs prosodiques est effectuée en tenant compte, pour un certain nombre de mots, d'informations contenues dans le dictionnaire.

La phase de traduction proprement dite consiste en un examen de gauche à droite de la chaîne à traduire et une recherche dans le dictionnaire de l'occurrenceidentique à celle de la chaîne d'entrée. On obtient ainsi tout ou partie de sa traduction et l'incrémentation correspondante du pointeur de la chaîne d'entrée. Cette méthode permet la prise en compte du seul contexte droit (cas le plus général). La traduction de certaines occurrences, par exemple : " RI + voyelle ", nécessitant la prise en compte du contexte gauche est assurée par une procédure particulière dont l'adresse est fournie par le dictionnaire.

Exemple de la traduction du mot "COMPTER"

Chaîne d'entrée	Règle	Chaîne de sortie
ŧCOMPTERŧ	ŧ11ŧ	ŧ
COMPTERŧ	C11K	ŧK
OMPTERŧ	OMPT31/	ŧK/
TERŧ	T11T	ŧK/T
ERŧ	ER§21)	ŧK/T)
ŧ	ŧ11ŧ	ŧK/T)ŧ

STRUCTURE DU DICTIONNAIRE

Le dictionnaire contient un millier d'éléments de longueurs différentes, rangés séquentiellement en respectant la contrainte suivante : toute règle sous-chaîne d'une autre règle ne doit pas précéder cette dernière. Chaque élément contient au plus 5 zones:

1- la longueur de la chaîne graphémique (limitée à l'heure actuelle à 15 caractères)
2- un indicateur pouvant prendre 5 valeurs
   - cas général
   - appel à une procédure spéciale
   - liaison amont interdite
   - mot "non prosodique"
   - simultanéité des 2 cas précédents
3- chaîne graphémique
4- pointeurs de la chaîne d'entrée (graphémique) et de la chaîne de sortie (phonétique) ou éventuellement pointeur de procédure
5- chaîne phonétique

Le dictionnaire est divisé en 34 parties, les 31 premières correspondent aux mots ou règles ayant pour initiale l'un des 31 graphèmes (alphabet + é, è, ô, ï, ç).

La 32ème partie du dictionnaire contient les chaînes graphémiques dont la traduction est particulière du fait de leur position initiale dans le mot (précédées d'un blanc).

Exemple : il vient d'entamer un fruit amer.

La 33ème partie contient les mots à liaison aval obligatoire ou interdite. La 34ème contient les mots jouant un rôle lors de l'insertion des marqueurs prosodiques. Pour éviter toute duplication les 2 parties précédentes contiennent quelques mots de cette catégorie, l'identification en étant faite par l'indicateur cité plus haut.

Pour l'écriture des règles, nous avons utilisé, outre les 31 graphèmes et le blanc, quelques symboles permettant de réduire le nombre de règles (symbole voyelle, symbole consonne, symbole représentant un blanc en un "s" suivi d'un blanc, symbole représentant un caractère quelconque.....).

L'adressage de ce dictionnaire est fait sur les 2 premières lettres, ce qui réduit considérablement le temps d'accès à une règle au prix d'une occupation mémoire de 2 K octets supplémentaires correspondant à la table d'adresse.

ALGORITHME

On distingue 4 étapes successives :

1 - les caractères de la chaîne d'entrée sont transcrits suivant un code regroupant les lettres présentant certaines analogies de traitement.

Simultanément, compte tenu d'une taille de buffer donné, le programme détermine la longueur maximum de texte qu'il traitera en 1 passage pour avoir une incidence minimum sur la traduction des liaisons.

2 - les nombres écrits en chiffres (maximum 9) sont convertis en mots en respectant l'orthographe afin d'unifier les traitements.

3 - Un automate insère les marqueurs prosodiques et les marqueurs de liaisons.

Le positionnement des marqueurs prosodiques est déterminé par l'algorithme décrit dans la communication de D. TEIL (1979).

Les marqueurs de liaison sont de 3 types : "a", "b", "c". Un marqueur de type "a" est placé devant les mots susceptibles d'engendrer une liaison aval; un marqueur de type "b" est placé devant les mots susceptibles d'engendrer une liaison amont et n'appartenant pas à la liste des mots à liaison amont interdite.

Les marqueurs de type "c" concernent le traitement des liaisons à l'intérieur des mots composés contenant un trait d'union.

4 - La quatrième phase est la phase de traduction proprement dite, effectuée selon le principe exposé plus haut. Les liaisons sont traitées simultanément :

- un marqueur de type "a" entraîne la recherche du mot qui le suit dans le dictionnaire des liaisons aval obligatoires ou interdites (33ème partie).

- un marqueur de type "b" appelle la procédure de traitement des liaisons régulières : seules sont faites les liaisons en Z (mot terminé par S ou X) et ce uniquement dans le cas ou le mot suivant se termine par S ou X (pluriel).

- Un marqueur de type "c" permet de faire les liaisons en Z et T (trait d'union précédé par S, X, Z, T)

REALISATION

La réalisation a été, dès le début de notre étude, envisagée sur microprocesseur. Néanmoins, afin de bénéficier des facilités offertes par les langages évolués et l'utilisation de grands fichiers, le développement et la mise au point de l'algorithme et du dictionnaire ont été faits sur les ordinateurs d'un centre de calcul (IBM 370/168). Pour cette phase de simulation, nous avons préféré le FORTRAN à tout autre langage en raison desfacilités de conversion en assembleur qu'il présente.

L'élaboration du dictionnaire a été faite selon la démarche suivante.

A partir des règles établies par D. TEIL (1969), nous avons écrit un premier noyau (environ 200 règles) qui nous a servi à phonétiser un lexique de 17 000 mots.

Pour chaque erreur de traduction (1) nous avons formulé une règle de la manière la plus générale possible compatible naturellement avec l'algorithme.

A l'issue de ce premier balayage du lexique des 17 000 mots, nous avons obtenu un dictionnaire d'environ 800 règles. Certaines des règles écrites pendant cette première passe ont eu des conséquences inattendues sur des mots traités antérieurement. Nous avons opéré une deuxième passe en cours de laquelle chaque nouvelle règle élaborée à partir d'une erreur de traduction était analysée par un programme chargé d'extraire du lexique les mots concernés par cette règle. Nous pouvions ainsi déterminer d'une part la validité de la règle et d'autre part la liste des exceptions devant faire l'objet d'une nouvelle règle de chaîne graphémique plus longue.

Les règles sont intégrées au dictionnaire par un programme effectuant le classement, le compactage et l'édition de la table d'adresse. L'adressage est fait sur les 2 premières lettres de chaque règle et, à l'intérieur de chacune de ces subdivisions, les règles sont classées suivant l'ordre des longueurs croissantes de leur chaîne graphémique, excepté le cas où une chaîne est sous chaîne d'une autre. Les règles d'utilisation la plus fréquente étant celles de chaîne graphémique les plus courtes, ce classement permet d'accroître la vitesse de l'algorithme. Les performances obtenues en simulation, tant du point de vue de la vitesse d'exécution (500 mots /s sur IBM 370/168), que de l'occupation mémoire nous ont permis d'opter pour un microprocesseur standard : INTEL 8080. Le dictionnaire est transféré par une ligne conversationnelle (TSO) d'un fichier du centre de calcul à une disquette du système microprocesseur. Par cette opération très simple, le programme implanté sur microprocesseur peut bénéficier à chaque instant de la dernière version de dictionnaire mise au point.

RESULTATS

Nous n'avons traité à l'heure actuelle, ni les sigles ou abréviations, ni les homographes hétérophones. Nous pensons que le nombre de sigles ou abréviations utilisés au sein d'une application donnée est relativement faible.

(1) Nous avons pris pour référence le dictionnaire de prononciation de MARTINET (1973).

Il est par conséquent préférable d'introduire leur traduction dans le dictionnaire plutôt que d'alourdir le traitement par un procédure particulière. On distingue 2 sortes d'homographes hétérophones (CATACH, N., 1977) ceux pour lesquels l'ambiguité peut-être levée par la syntaxe (ex: portions, président ...) et ceux qui nécessitent une analyse sémantique (ex: fils). Le problème pourra parfois être résolu dans le premier cas par une recherche de mots grammaticaux, la structure de notre programme se prête facilement à un tel traitement. Dans le deuxième cas, on ne peut que limiter le nombre d'erreurs grâce à une liste de fréquence d'utilisation dans le contexte considéré.

En collaboration avec l'équipe HESO du CNRS (CATACH, MEISSONNIER), nous avons réalisé un programme de comptage de l'utilisation des différentes règles du dictionnaire. Mis en oeuvre sur un jeu de textes d'environ 300 cartes perforées, il laisse apparaître que 3% des règles traitent environ 70% des cas et sur l'exemple considéré 39% des règles traitent l'ensemble des textes. Cette remarque revêt une certaine importance si l'on veut utiliser le programme sur un système de faible capacité mémoire.

Le taux d'erreur obtenu à l'heure actuelle avec un dictionnaire de 1000 règles est très faible. Nous pensons qu'il est illusoire de le chiffrer car il peut varier suivant le type de texte traité. Les erreurs portent essentiellement sur des noms propres à consonnance étrangère, des mots peu usités, des liaisons non faites ou des cas correspondant aux limites exposées ci-dessus.

Dans le cadre d'une application déterminée on peut ramener le taux d'erreur à des valeurs encore plus faibles, en insérant des règles propres au contexte d'utilisation (juridique, médical, bancaire ...). Toute modification de dictionnaire est en effet très simple en utilisant les programmes mentionnés plus haut.

Pour la version implantée sur le système microprocesseur INTEL 8080, le dictionnaire muni de sa table d'adresses occupe actuellement environ 12 K octets, le programme moins de 4 K octets. Une version prototype nécessiterait environ 14 K octets de mémoire morte (ROM) et 2 K octets de mémoire vive (RAM). Le temps de traduction d'une phrase de 20 mots est d'environ 0,3 seconde.

CONCLUSION

Malgré l'absence d'analyse syntaxique, les résultats obtenus pour le traitement des liaisons et la traduction des terminaisons "ent" sont satisfaisants. Les erreurs de liaisons subsistantes sont pour la grande majorité des liaisons obligatoires omises, ce qui semble préférable à des liaisons interdites effectuées.

Pour un taux d'erreur comparable ou moindre selon la taille du dictionnaire choisie, l'algorithme qui fait l'objet de cette communication présente deux caractéristiques essentielles par rapport aux réalisations déjà existantes : il effectue d'une part la mise en place automatique des marqueurs prosodiques, il est d'autre part le seul, à notre connaissance, opérationnel sur un système microprocesseur, ce qui constitue une première étape importante pour la réalisation d'un synthétiseur autonome.

REFERENCES

CATACH, N., 1977, Projet de recherche : "Ambiguité, information et redondance graphique dans la chaîne écrite du français"

DIVAY, M. , GUYOMARD, M. , 1977, Conception et réalisation sur ordinateur d'un programme de transcription graphémo-phonétique du français (thèse de 3ème cycle, Université de RENNES)

MARTINET, A. , WALTER, H. , 1973, Dictionnaire de la prononciation française dans son usage réel.

TEIL, D. , 1969, Etude de génération synthétique de parole à l'aide d'un ordinateur (Mémoire d'ingénieur CNAM, PARIS)

TEIL,D. , 1975, Conception et realisation d'un terminal à réponse vocale (thèse de docteur ingénieur, Université de PARIS VI)

TEIL,D. , 1979, Comparaison de plusieurs algorithmes de marquage prosodique (10ème JEP du GALF, GRENOBLE)

10èmes JOURNÉES D'ÉTUDE SUR LA PAROLE

GRENOBLE - 30 MAI - 1ER JUIN 1979

GENERATION AUTOMATIQUE DE MARQUEURS PROSODIQUES EN VUE DE LA SYNTHESE D'UN TEXTE QUELCONQUE

P. QUINTON (*), F. EMERARD, P. GRAILLOT, D. LARREUR

CNET LANNION

RESUME

On présente ici un projet de synthèse à partir d'un texte quelconque en français, utilisant une analyse syntaxique pour la génération de la prosodie.

Le texte pris en compte par l'analyseur est au préalable traité par un programme d'analyse morphologique dont les données sont d'une part un dictionnaire contenant les mots les plus courants de la langue française (mots grammaticaux, verbes courants et mots à forte occurence), et d'autre part des règles de classification morphologique à partir des désinences. Le résultat de ce traitement est donc une suite de classes morphologiques dont certaines sont multiples (lorsque la désinence ne permet pas de décider).

Une analyse syntaxique est alors effectuée sur cette suite ; elle permet d'une part de trouver la classification la plus probable de chaque mot par contexte, d'autre part, de construire l'arbre syntaxique et de générer alors les marqueurs par application de transformations d'arbres.

Le but du projet est d'explorer les possibilités offertes par cette méthode. La principale difficulté est liée à l'incertitude de la classification morphologique des mots inconnus.

Si elle est surmontée, elle permettrait de s'affranchir de la connaissance a priori des mots du texte à synthétiser.

(*) Adresse actuelle : IRISA, Université de Rennes, Avenue du Général Leclerc 35042 RENNES

Automatic prosodic marking for a Text-to-Speech system.

P. QUINTON*, F. EMERARD, P. GRAILLOT, D. LARREUR
CNET - LANNION

SUMMARY

In a system designed to generate synthetic speech by diphones from an input text, an important component is the prosodic treatment.

An algorithm to generate Fo, duration and intensity has been previously developed (EMERARD, 1977). However the present scheme requires additional marks in order to indicate syntactic clause and phrase boundaries.

This paper describes the structure of an algorithm using a syntactic analysis for an automatic prosodic processing.

Figure 1 shows the general organization which is divided into four modules.

The first one analyses the text with a typographic test procedure of the string of orthographic symbols.

The second one consists in a statistical morphological analysis : its data give

- a basic dictionary which contains "function" words (prepositions, articles, pronouns...) and the more frequent radicals of verbs,
- morphological rules based on the termination of words.

The output of the module supplies statistical distribution of each word among possible morphological classes.

In the third module, the syntactic analizer builds up all the possible syntactic structures according to a context free grammar and gives the more likely syntactic tree.

In the last one, the input text is synthesized using a grapheme-phoneme transcriber (DIVAY & GUYOMARD, 1977) and a diphone synthesizer.

This work is related to a development project of a speech - output machine with a LPC real time synthesizer on a microprocessor system.

*Present address: IRISA, Université de Rennes, Avenue du Général Leclerc
35042 RENNES

INTRODUCTION

Le problème de la génération automatique de la prosodie qui est abordé ici s'inscrit dans le cadre d'un projet de synthèse vocale à partir d'un texte quelconque en français.

Le projet s'intègre dans la réalisation actuellement en cours au CNET d'une "boîte à parole" constituée d'un circuit intégré, synthétiseur par prédiction linéaire, et d'un microprocesseur. De ce fait, le logiciel qui réalisera le passage du texte écrit au signal de parole sera soumis à de sévères contraintes qu'il faut prendre en compte dans les algorithmes retenus.

On présente ici non pas les solutions du problème mais plutôt une approche du fonctionnement du système envisagé, ainsi qu'un recensement des difficultés à surmonter et des outils dont on peut disposer.

Cet article comporte cinq parties. Dans la partie I, on justifie l'approche syntaxique utilisée en faisant référence à d'autres travaux sur la synthèse ; la partie II décrit le fonctionnement général du système ; dans la partie III, on s'intéresse plus particulièrement au bloc de traitement qui concerne l'analyse morphologique ; dans la partie IV, on décrit le mécanisme de l'analyse syntaxique et de la génération des marqueurs ; enfin, la partie V est consacrée aux objectifs du système et en particulier, aux problèmes d'évaluation de ses performances.

I - SYNTAXE ET SYNTHESE DE LA PAROLE

Si les équipes travaillant dans le domaine de la synthèse de la parole se sont attachées dans un premier temps à produire d'abord une parole intelligible, elles se préoccupent depuis quelques années -et du fait des applications envisagées- du naturel et de l'agrément de la voix (Allen, 1976 ; Uméda et al., 1975 ; Choppy et al., 1976 ; Le Roux et al., 1976 ; Emerard, 1977 ; Rodet, 1977...).

Aujourd'hui, la plupart des systèmes de synthèse utilisent dans le message à synthétiser des repères syntaxiques qui servent à introduire certaines des évolutions prosodiques observées à l'analyse.

Certaines équipes, qui réalisent un traitement prosodique simplifié, se contentent de la détermination des frontières des unités de sens grâce au repérage automatique des mots grammaticaux (prépositions, conjonctions etc...). L'identification de ces mots ne pose pas de problème particulier puisqu'ils n'appartiennent en général qu'à une seule classe grammaticale et sont en nombre limité.

En ce qui concerne le système de synthèse du CNET, le problème est plus complexe dans la mesure où le traitement prosodique est plus élaboré. Pour l'instant, l'introduction de la prosodie nécessite de positionner manuellement des marqueurs en certains points du message considérés comme pertinents ; ceux-ci permettent d'introduire par programme, pour chaque mot les précédant et sur chaque syllabe, des variations spécifiques (liées à la longueur du mot et à sa position dans la phrase) de Fo, de durée et d'intensité.

Pour les trois types de phrases étudiées jusqu'ici (énonciatif, impératif, interrogatif), 14 marqueurs sont prévus qui renvoient à 10 patrons prosodiques de mot (ces patrons sont liés à la longueur du mot).

Pour des phrases assez courtes et assez simples du type :

Syntagme préverbal + Syntagme verbal + Syntagme post verbal.

il semble que l'on puisse "se contenter" de l'identification

- de la ponctuation de fin de phrase (?!.),
- des frontières du groupe verbal,
- des mots grammaticaux pour permettre une séparation des unités de sens à l'intérieur des syntagmes.

Cependant il ne fait pas de doute que toute amélioration du naturel de la voix passe par un traitement encore plus précis de la prosodie et donc par une décomposition encore plus fine de la phrase en ses éléments syntaxiques. Ainsi l'étude de la prosodie ne saurait être menée indépendamment de la syntaxe (voire de la sémantique ?).

II - ALLURE GENERALE DU SYSTEME ENVISAGE

Le schéma de la figure I représente la structure globale du système qui comprend quatre blocs.

II-1- Saisie du texte :

Les fonctions de ce bloc peuvent être regroupées en deux phases : une analyse typographique et un contrôle de la chaîne d'entrée.

Dans la version finale, l'entrée pourra être constituée d'une suite quelconque des caractères disponibles sur une machine à écrire : lettres, signes de ponctuation, chiffres, symboles mathématiques et caractères spéciaux. Il faudra donc savoir effectuer une analyse de la chaîne d'entrée.

- pour repérer les majuscules qui permettent de connaître les noms propres et les sigles,
- pour tenir compte de la typographie, en particulier des titres, en vue du traitement de la prosodie,
- pour avoir accès à la ponctuation qui fournit des informations pour l'analyse morphologique et surtout pour la prosodie, par exemple les virgules ou le point d'interrogation qui permet de connaître les phrases à structure interrogative,
- parce que l'orthographe joue un rôle important : l'analyse morphologique se fonde principalement, comme on le verra, sur les désinences des mots.

Le contrôle de l'exactitude de la chaîne d'entrée -que ce soit de la ponctuation ou de l'orthographe- reste pour le moment problématique dans la mesure où il n'est pas envisagé de disposer d'un dictionnaire exhaustif. Il est d'ailleurs probable que ce contrôle devra être fait en relation avec l'analyse morphologique et syntaxique. De toute façon, quel que soit ce contrôle, se pose le problème de déterminer le niveau d'erreur critique quant aux performances du système.

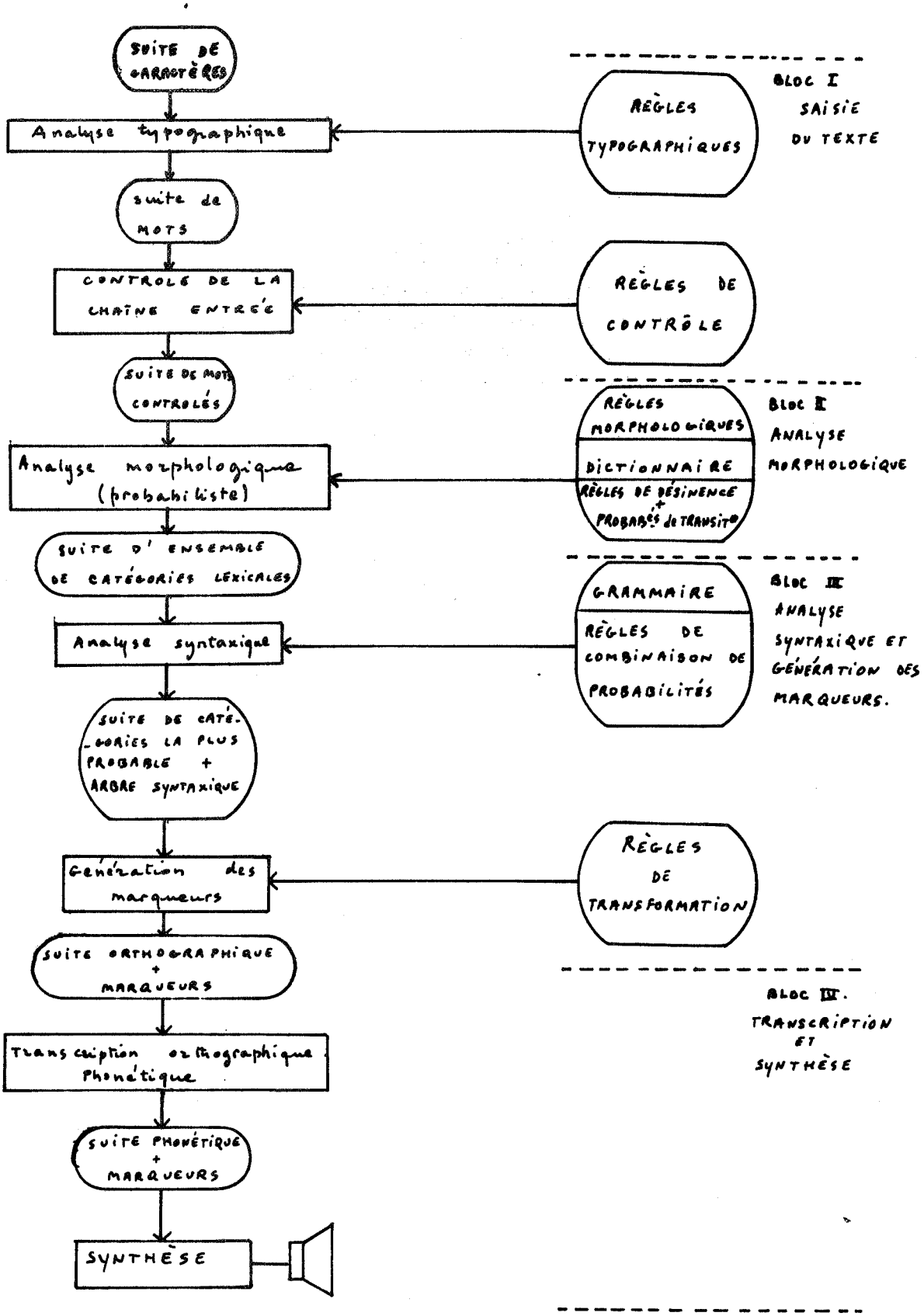


Figure 1 : Structure générale du système

II-2- Analyse morphologique :

Grâce à la consultation d'un dictionnaire de base et en s'aidant de règles grammaticales, elle permet d'établir pour chaque mot la liste des catégories grammaticales (substantif, verbe, adverbe... etc...) auxquelles il est susceptible d'appartenir. A chaque classe est associée une probabilité qui sert d'entrée à l'analyse syntaxique.

II-3- Analyse syntaxique et génération des marqueurs :

A partir de règles de grammaire, elle détermine à la fois la suite la plus probable de classes lexicales ainsi que son arbre syntaxique à partir duquel sont ensuite générés les marqueurs prosodiques.

II-4- Transcription et synthèse :

Cette partie du traitement est déjà opérationnelle : elle comprend d'une part la transformation des formes orthographiques en formes phonétiques (DIVAY et GUYOMARD, 1977), d'autre part la synthèse par diphones (EMERARD, 1977)

On peut noter que la transcription orthographique-phonétique pourra utiliser l'information issue des traitements antérieurs pour lever certaines ambiguïtés (exemple type : les poules du couvent couvent...). Toutefois l'interaction est assez limitée car le nombre d'ambiguïtés phonétiques que l'on peut lever à partir de la connaissance des classes lexicales est généralement considérée comme relativement faible. Le schéma de la figure I laisse supposer que le traitement est entièrement séquentiel. En réalité, les modules seront réalisés de façon à pouvoir s'exécuter de façon concurrente, afin que chacun puisse bénéficier au maximum des informations fournies par les autres.

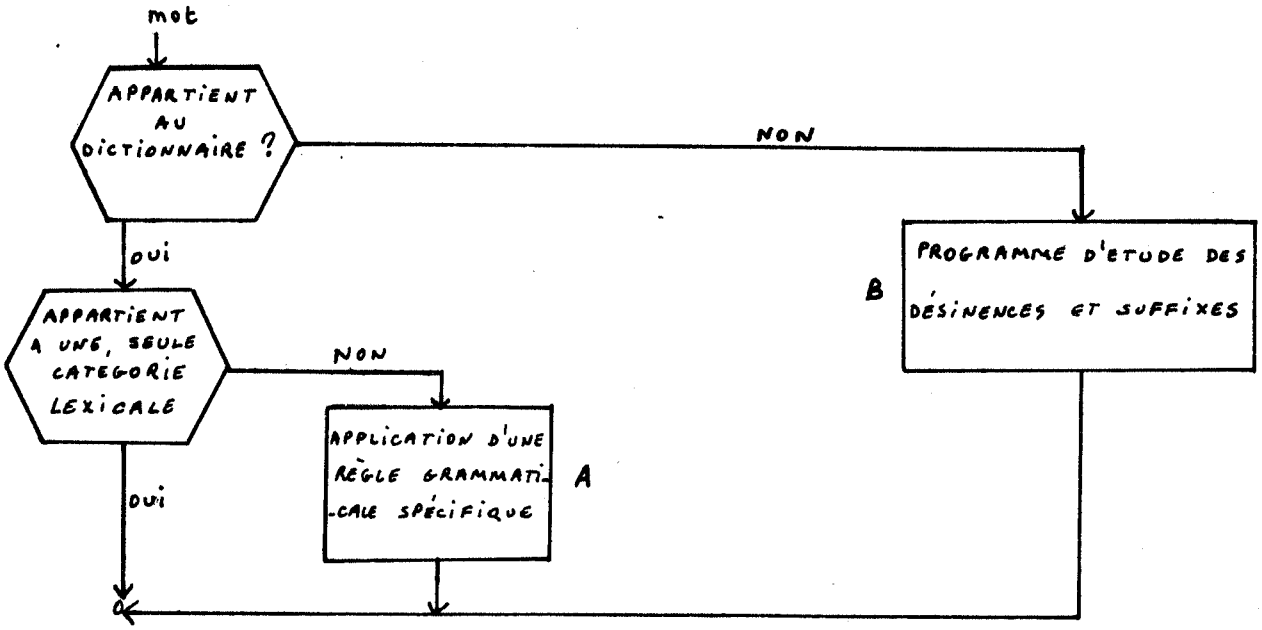
III - ANALYSE MORPHOLOGIQUE PROBABILISTE

Comme nous l'avons déjà vu, les contraintes informatiques liées au projet nous interdisent un dictionnaire de taille trop importante.

Les informations contenues dans le dictionnaire dépendront de la précision demandée à l'analyse syntaxique,ce qui est lié à la définition des nouveaux marqueurs. Mais quelle que soit la précision retenue, il faut prévoir de stocker les radicaux des verbes les plus courants, les éléments des catégories fermées (prépositions, articles, adverbes, conjonctions, ...) et les mots constituants des exceptions aux règles grammaticales générales. Par contre, très peu de mots d'autres catégories lexicales -mis à part ceux qui sont connus pour leur fréquence élevée- y figureront. Une justification de cette option provient du fait que si pour deux mots successifs on détermine que l'un et l'autre peuvent être à la fois adjectif ou substantif, le problème le plus difficile pour une analyse syntaxique est de déterminer dans quel ordre de classe les deux mots se suivent (plusieurs structures "substantif + adjectif", "adjectif + substantif"... sont possibles en français). Or nous n'avons pas à lever cette ambiguïté, car l'ordre de ces classes ne semble pas avoir d'influence sur la prosodie (ou plus exactement, il n'a pas été mis en évidence jusqu'ici).

Nous comptons organiser le dictionnaire selon le modèle proposé par l'IMAG * (GRANDJEAN, E., 1975), c'est-à-dire répertorier les différents couples catégories lexicales -règles de désinence. Un mot est alors défini dans le dictionnaire par son radical et un représentant de son couple. Les règles de désinence sont représentées par des automates.

L'allure générale du traitement est représentée sur le schéma suivant :



Les résultats des procédures A et B peuvent être exprimées en une suite de probabilités d'appartenance du mot aux différentes catégories lexicales. Cette option est compatible avec l'analyseur syntaxique ci-après. En outre elle permet d'inclure,comme filtre supplémentaire entre l'analyse morphologique et l'analyse syntaxique, un module de traitement des probabilités de transition. En effet différents travaux (JELINEK et al., 1975) ont montré la richesse de l'information apportée à la connaissance de la catégorie lexicale du mot n par celles des mots n-2, n-1, n+1, n+2. Le module,intéressant par sa simplicité, permet d'affiner, à partir des mots pour lesquels le dictionnaire a déterminé de manière univoque la catégorie lexicale, la chaîne sortie de l'analyse morphosyntaxique.

IV - ANALYSE SYNTAXIQUE ET GENERATION DES MARQUEURS

La génération comporte deux étapes : l'analyse syntaxique puis la génération proprement dite. L'analyse consiste simultanément à déterminer quelle est la séquence de classes lexicales qui est la plus probable, et à construire l'arbre syntaxique de cette séquence en appliquant des règles décrites dans une grammaire hors contexte. La génération des marqueurs consiste à appliquer à l'arbre syntaxique des transformations associées aux règles de la grammaire.

---

* Institut de Mathématiques appliquées de Grenoble.

IV-1- Analyse :

Notons $m_1 \ldots m_n$ la suite des mots à analyser ; à chaque mot m_i, l'analyse lexicographique fait correspondre un ensemble de couples (x_i^j, w_i^j), $j = 1, \ldots, K_i$ où x_i^j est une catégorie lexicale de la grammaire et w_i^j une probabilité.

Le problème est de trouver parmi les suites $\alpha = x_1^{j(1)} x_1^{j(2)} \ldots x_n^{j(n)}$ appartenant au langage engendré par la grammaire celle qui maximise la probabilité $p(\alpha) = \prod_{i=1}^{n} w_i^{j(i)}$.

Ce problème peut être résolu par la méthode qui est utilisée pour la reconnaissance de phrase dans le système de reconnaissance de la parole KEAL (QUINTON, 1977) et qui consiste à effectuer simultanément l'analyse syntaxique de toutes les chaînes α en déterminant en même temps la probabilité $p(\alpha)$; si plusieurs solutions α existent, on prend celle dont la probabilité $p(\alpha)$ est la plus élevée. Cette façon de procéder permet d'effectuer le choix de la classe lexicale des mots inconnus en utilisant toute l'information syntaxique ce qui ne serait pas le cas par exemple si on faisait intervenir uniquement le contexte immédiat du mot. Ainsi, dans l'exemple de la figure II, l'arbre syntaxique permet de déterminer que le mot "journaliste" est un substantif alors que le contexte immédiat ne permet pas de le savoir (la séquence pronom objet (le) + verbe + article partitif (du) étant possible comme le montre la séquence soulignée de la phrase suivante : "il le jette du haut du toit"). Or le fait de savoir que "journaliste" n'est pas un verbe est essentiel au positionnement des marqueurs dans le système du CNET.

IV-2- Génération :

La génération des marqueurs se fait en effectuant une interprétation de l'arbre syntaxique au moyen de transformations élémentaires associées aux productions de la grammaire.

Dans l'exemple choisi, la phrase munie de ses marqueurs est :

le journaliste @ du magazine # travaille @ dans son bureau * avec le directeur.(†) Le placement des marqueurs se fait au niveau des règles 2 et 3 (figure II). Par exemple, la règle 2 :

<Phrase> ⟶ <sujet> <groupe verbal> <complément> <complément>

induit sur la phrase le parenthésage suivant :

(†)

" @ " = Fo descendant ; durée de la dernière voyelle précédant ce marqueur allongée de 40 ms.

" # " = Fo montant ; la dernière voyelle dure environ 200 ms ; introduction d'une pause supérieure à 150 ms.

" * " = Fo montant ; la dernière voyelle dure environ 200 ms ; une pause de 65 ms est introduite.

(le journaliste du magazine) (travaille) (dans son bureau) (avec le directeur)

sujet verbe complément complément

En associant à cette règle l'action qui consiste à réécrire la séquence

⟨sujet⟩ # ⟨groupe verbal⟩ @ ⟨complément⟩ * ⟨complément⟩

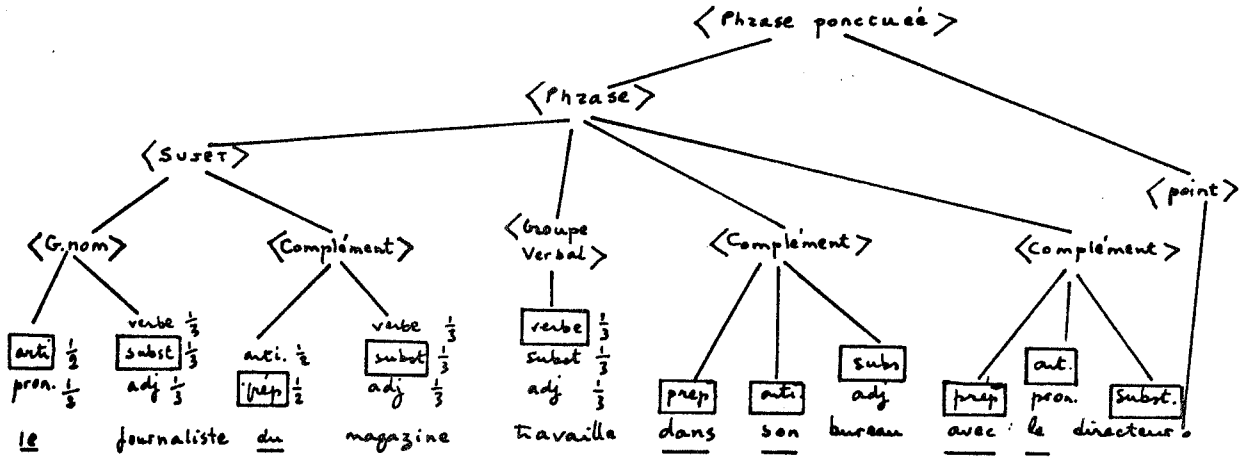
on place la plus grande partie des marqueurs prosodiques. Il suffit d'associer à la règle 3 qui est :

⟨sujet⟩ ⟶ ⟨groupe nominal⟩ ⟨complément⟩

l'action de réécriture :

⟨groupe nominal⟩ @ ⟨complément⟩

et enfin de remplacer tous les non terminaux par les mots sous jacents pour obtenir la phrase munie de ses marqueurs.



1 ⟨phrase ponctuée⟩ ⟶ ⟨phrase⟩ "point"
2 ⟨phrase⟩ ⟶ ⟨sujet⟩ ⟨groupe verbal⟩ ⟨complément⟩⟨complément⟩
3 ⟨sujet⟩ ⟶ ⟨groupe nominal⟩ ⟨complément⟩
4 ⟨groupe nominal⟩ ⟶ "article" "substantif"
5 ⟨complément⟩ ⟶ "article" "substantif"
6 ⟨complément⟩ ⟶ "préposition" "article" "substantif"

Figure 2 : Exemple de traitement syntaxique pour la phrase "Le journaliste du magazine travaille dans son bureau avec le directeur". Pour chaque mot, on donne les catégories lexicales auxquelles il peut appartenir avec une probabilité.

VI - OBJECTIFS A ATTEINDRE

Le critère de réussite du positionnement automatique des marqueurs ne devra pas être évalué en termes de score mathématique mais plutôt en fonction d'une perte d'intelligibilité à des tests auditifs. En effet certaines fautes sur les marqueurs auront des répercussions faibles sur l'intelligibilité de la phrase (voire sur l'agrément), mais d'autres la rendront soit totalement incompréhensible soit compréhensible au prix d'un effort de reconstruction du sens. Ce sont donc celles-ci qu'il conviendra de pénaliser.

Le critère de réussite du système montre bien l'interaction qui existe entre la complexité du traitement et la qualité désirée de la parole synthétique. En effet, chaque bloc de traitement peut être plus ou moins détaillé et faire intervenir un ensemble plus au moins exhaustif de règles. Le choix des catégories lexicales retenues pour l'ensemble du système est un exemple typique car déterminé en fonction du traitement prosodique envisagé.

Ceci explique pourquoi, aux contraintes informatiques citées dans l'introduction, s'ajoute la nécessité de pouvoir faire évoluer de façon très souple le système. En conséquence toutes les étapes du traitement seront paramétrées.

CONCLUSION

Notre but est de réaliser un système, constituant un compromis entre la qualité de la parole synthétisée et la complexité du traitement, qui soit implantable d'ici trois ans sur un système microprocesseur.

La division de ce système en modules,dont chacun ne nécessite qu'un temps de traitement et une quantité de données à stocker assez faibles, permet d'espérer que soit réalisée une synthèse vocale à partir d'un texte quelconque en français.

BIBLIOGRAPHIE

ALLEN, J., O'SHAUGHNESSY, D., "A comprehensive model for fundamental frequency generation". I.E.E.E Int. Conf. on Acoustics..., Philadelphie, 701-704, 1976.

CHOPPY, C., LIENARD, J.S., "Prosodie automatique pour la synthèse par diphonèmes" 8èmes J.E.P., Aix, 211-217, 1977.

DIVAY, M., GUYOMARD, M., "Conception et réalisation sur ordinateur d'un programme de transcription graphémo-phonétique du français". Thèse, Université Rennes, Avril 1977.

EMERARD, F., "Synthèse par diphones et traitement de la prosodie". Thèse, Université Grenoble, Mars 1977.

GRANDJEAN, E., "Conception et réalisation d'un dictionnaire pour un analyseur intéractif de langues naturelles". Thèse, CNAM Grenoble, Février 1975.

JELINEK, F., BAHL, L.R., MERCER, R.L., "Design of a linguistic statistical decoder for the recognition of continuous speech". I.E.E.E Trans. on Information Th. Vol. IT-21. Mai 1975.

LEROUX, J., FERVERS, H., MICLET, L., "Programme de transcription orthographique-phonétique en langue française". Rapport E.N.S.T 1976.

QUINTON, P., "Utilisation d'un analyseur syntaxique pour la reconnaissance de la parole continue". Annales des Télécommunications. Septembre 1977.

RODET, X., "Analyse du signal vocal dans sa représentation amplitude-temps ; synthèse de la parole par règles", Thèse, Paris VI, 1977.

UMEDA, N., TERANISHI, R., "The parsing program for automatic text-to-speech synthesis", "I.E.E.E. ASSP n° 23. Avril 1975.

10èmes JOURNÉES D'ÉTUDE SUR LA PAROLE

GRENOBLE - 30 MAI - 1ER JUIN 1979

COMPARAISON DE PLUSIEURS ALGORITHMES DE MARQUAGE PROSODIQUE

D. TEIL

L.I.M.S.I. du C.N.R.S. - 91406 ORSAY

RESUME

L'élaboration automatique de la prosodie en vue de la synthèse de parole à partir du texte écrit comporte deux étapes:

a/ calcul des marqueurs délimitant les groupes prosodiques

b/ calcul de l'évolution des paramètres physique de la prosodie en fonction du découpage précédent.

Nous nous interessons surtout ici à la première étape pour laquelle nous avons écrit quatre algorithmes différents, qui ne font intervenir la syntaxe que de manière très rudimentaire. Un ensemble de phrases synthétisées selon ces hypothèses ont été soumises à un test de préférence auprès de plusieurs auditeurs. L'algorithme retenu est celui qui fournit le plus grand nombre de marqueurs.

COMPARAISON DE PLUSIEURS ALGORITHMES DE MARQUAGE PROSODIQUE

D. TEIL L.I.M.S.I. du C.N.R.S. - ORSAY

SUMMARY

Automatic prosody elaboration from written text in speech synthesis includes two steps:

- calculation of prosodic group mark-setting
- calculation of prosody physical parameter evolution with respect to the preceding segmentation.

Our investigations mostly concern the first step. Four different algorithms using very rudimentary syntax were worked out. A set of sentences synthetized according to these hypotheses were presented to several listeners for a preference test.

The algorithm that was retained yields the greatest number of marks.

10èmes JOURNÉES D'ÉTUDE SUR LA PAROLE
GRENOBLE - 30 MAI - 1ER JUIN 1979

COMPARAISON DE PLUSIEURS ALGORITHMES DE MARQUAGE PROSODIQUE

D. TEIL L.I.M.S.I. du C.N.R.S. - ORSAY

INTRODUCTION

Le problème de la synthèse de la parole par la méthode des diphonèmes présente deux sortes de difficultés pour obtenir une voix dite " naturelle " : c'est d'une part l'intelligibilité et d'autre part ce que l'on pourrait appeler l'esthétique de la voix.

Grossièrement l'intelligibilité est liée en grande partie à la constitution du dictionnaire des formes acoustiques alors que l'aspect esthétique est surtout rendu par la prosodie.

L'élaboration automatique de la prosodie comporte deux étapes:
a/ le calcul des marqueurs délimitant les groupes prosodiques
b/ le calcul de l'évolution des paramètres physiques de la prosodie en fonction du découpage précédent.

L'expérience qui est décrite ici concerne uniquement l'étude de plusieurs algorithmes de placement automatique des marques à partir de considérations de type lexical. L'utilisation de la syntaxe est réduite ici à la consultation d'une liste de mots grammaticaux appelés mots outils.

A la suite d'études et d'analyse de textes par différentes personnes nous avons défini quatre algorithmes susceptibles de fournir une prosodie correcte.

Pour faire le choix parmi ces algorithmes nous avons synthétisé un corpus de 15 phrases en appliquant les quatre lois prosodiques et nous les avons soumises à plusieurs auditeurs pour établir une classification de préférence.

LES LOIS DE MARQUAGE

Les deux premières lois de positionnement automatique des marqueurs de coupe prosodique prennent en considération la longueur des mots et leur appartenance à la liste des mots grammaticaux. Ces lois ont été définies à partir des travaux de C. CHOPPY (1975).

- Première loi : coupe prosodique après tout mot qui n'est pas dans la liste des mots outils.
- Deuxième loi : coupe prosodique après tout mot long (plus d'une syllabe) s'il n'est pas dans la liste des mots outils.
- Troisième loi : coupe prosodique à la transition mot lexical - mot outil (proposition faite par D. MEMMI à la suite de l'analyse de quelques textes).
- Quatrième loi : coupe prosodique obtenue en appliquant en même temps les algorithmes des lois 2 et 3.

CONDITIONS DE L'EXPERIENCE

Nous avons volontairement simplifié le problème du marquage en ne considérant qu'une seule catégorie de marqueurs prosodiques à l'intérieur des phrases et pour ce test nous avons choisi 15 phrases ennonciatives sans ponctuation interne.

Le choix de ces phrases a été guidé par des considérations de structure de phrase et de longueur de mots (certaines phrases n'ont que des mots courts). La liste des mots outils pris en compte au moment de cette expérience n'était pas complète en particulier le verbe FAIRE utilisé dans la phrase 11 et l'adjectif PETIT des phrases 9 14 n'y figurent pas. Ces mots ont été volontairement considérés comme des mots lexicaux dans le but de connaître l'influence de coupes prosodiques parasites et de déterminer si des coupes supplémentaires influent sur l'aspect prosodique général de la phrase.

DEROULEMENT DU TEST

Les quinze phrases du test ont été découpées manuellement en appliquant les lois définies précédemment (annexe 1).

Pour certaines phrases les lois de découpage conduisent au même regroupement prosodique.

Nous n'avons synthétisé et enregistré que les phrases ayant des découpages différents. Par exemple, les lois 1, 3 et 4 donnent le même découpage de la phrase 1; nous avons mis dans le test cette version et la version fournie par la loi n° 2.

Les quatre lois appliquées à la phrase 13 donnent le même résultat de découpage; pour cette phrase nous avons enregistré 2 fois la même séquence.

La synthèse a été réalisée sur le système Icophone 6 (synthétiseur à formants parallèles); les schémas mélodiques appliqués sont des schémas en dent de scie. (LIENARD J.S., TEIL D., 1977)

La loi de rythme allonge la durée des voyelles accentuées pour les groupes de début de phrase et la durée de la dernière voyelle des groupes de fin de phrase.

Chaque phrase est d'abord énoncée oralement de façon à limiter le problème de l'intelligibilité, puis elle est éditée deux fois par version synthétisée.

L'auditeur doit donner sa préférence en notant la prosodie de la façon suivante :

1 - si la prosodie semble mauvaise
2 - si elle est moyenne
3 - si elle est assez bonne
4 - si elle est bonne
et si possible il devra indiquer les défauts constatés.

DEPOUILLEMENT DES RESULTATS

Nous avons soumis ce test à sept auditeurs. Les préférences pour une loi plutôt qu'une autre ne sont pas très marquées : des versions différentes d'une même phrase ont été entendues comme identiques pour certains, et les deux versions identiques de la phrase 13 n'ont pas été toujours notées de la même façon.

Le premier but du test étant de choisir une loi de découpage prosodique, le dépouillement a consisté en première approximation à faire la somme des notes obtenues pour chacune des lois.

A partir de ce comptage simplê il apparaît que la loi 1 (celle qui donne le plus grand nombre de coupes) est celle qui globalement donne les meilleurs résultats et la loi n° 3 celle qui globalement a les plus mauvaises notes.

Une autre manière de compter donnant des résultats comparables a également été utilisée : nous avons pondéré par le coefficient - 2 les notes 1
- 1 les notes 2
+ 1 les notes 3
et + 2 les notes 4.

Les résultats positifs correspondent aux appréciations les plus favorables. Les résultats sont portés en annexe 2.

CONCLUSION

Ce test d'appréciation de la prosodie n'a pas la prétention d'être parfait mais il donne une méthode qui sera reprise pour d'autres tests du même type.

Le nombre de résultats obtenus est nettement insuffisant et surtout les auditeurs ont beaucoup de mal à isoler la prosodie dans les phrases synthétisées; l'aspect intelligibilité et qualité de parole influent très fortement sur l'appréciation des résultats.

Pour la même raison, il semble que l'ordre dans lequel les phrases synthétisées sont présentées a de l'importance : pour certains auditeurs la première version est systèmatiquement moins bien notée que la suivante.

Les règles rudimentaires de variations prosodiques demandent à être affinées en particulier les contours mélodiques et les variations temporelles du message.

En conclusion, il est très important pour établir des tests sur la prosodie de disposer d'une synthèse de parole de très bonne qualité. D'après ces résultats, il semble que la loi qui donne les meilleurs résultats soit celle qui donne le plus grand nombre de coupes.

REFERENCES

CHOPPY C., LIENARD J.S., TEIL D., 1975, Un algorithme de prosodie automatique sans analyse syntaxique; 6ème J.E.P. TOULOUSE.

LIENARD J.S., TEIL D., & al., 1977, Diphone synthesis of French: vocal response unit and automatic prosody from the texte; IEEE International Conference on Acoustics, Speech and Signal Processing. HARTFORD.

Annexe 1

1 coupe donnée par la loi 1
2 coupe donnée par la loi 2
3 coupe donnée par la loi 3
4 coupe donnée par la loi 4

1 - Elles s'en allaient $\frac{1}{2}$ dans la montagne $\frac{1}{2}$ et la haut $\frac{1}{3}$ le loup $\frac{1}{3}$ les mangeait.

2 - La table $\frac{1}{3}$ a été dressée $\frac{1}{2}$ ce matin.

3 - La table 1 dressée $\frac{1}{2}$ ce matin $\frac{1}{2}$ a brûlé.

4 - Le pot 1 jaune 1 rit $\frac{1}{3}$ de se grand rat 1 blanc.

5 - Le tout gros 1 chat 1 gris $\frac{1}{3}$ a croqué $\frac{1}{2}$ une toute petite $\frac{2}{4}$ souris $\frac{1}{2}$ blanche.

6 - Une toute petite $\frac{1}{2}$ souris $\frac{1}{2}$ blanche.

7 - Il a vu $\frac{1}{3}$ la balle 1 bleue 1 rouler $\frac{1}{2}$ à terre.

8 - Nous avons vu $\frac{1}{3}$ le plat $\frac{1}{3}$ de riz 1 rouler $\frac{1}{2}$ à terre.

9 - Le petit $\frac{1}{2}$ chat $\frac{1}{3}$ de la maison $\frac{1}{2}$ mange $\frac{1}{3}$ du poisson.

10- Le journaliste $\frac{1}{2}$ travaille $\frac{1}{2}$ dans son bureau $\frac{1}{2}$ avec le directeur.

11- Le facteur $\frac{1}{2}$ fait $\frac{1}{3}$ sa tournée $\frac{1}{2}$ à bicyclette.

12- Veuillez $\frac{1}{2}$ indiquer $\frac{1}{2}$ votre numéro $\frac{1}{2}$ de sécurité $\frac{1}{2}$ sociale.

13- Votre réponse $\frac{1}{2}$ n'est pas correcte.

14- Le gentil $\frac{1}{2}$ petit $\frac{1}{2}$ chat $\frac{1}{3}$ de la maison $\frac{1}{2}$ boit $\frac{1}{2}$ du lait.

15- Un tremblement $\frac{1}{2}$ de terre $\frac{1}{3}$ a dévasté $\frac{1}{2}$ les cultures $\frac{1}{2}$ de coton.

Nombre de coupes: loi 1 = 52
loi 2 = 31
loi 3 = 31
loi 4 = 44

Annexe 2

RESULTATS DU TEST:
pour toutes les phrases

	Loi 1	Loi 2	Loi 3	Loi 4
notes 1	1	3	7	3
notes 2	23	17	31	26
notes 3	20	37	18	24
notes 4	21	8	9	12
total	191	180	159	175

Résultats pondérés en affectant la valeur -2 à la note 1
-1 à la note 2
+1 à la note 3
+2 à la note 4

	Loi 1	Loi 2	Loi 3	Loi 4
notes 1	-2	-6	-14	-6
notes 2	-23	-17	-31	-26
notes 3	+20	+37	+18	+24
notes 4	+42	+16	+18	+24
total	+37	+30	-9	+16

10èmes JOURNÉES D'ÉTUDE SUR LA PAROLE

GRENOBLE - 30 MAI - 1ER JUIN 1979

SYSTEME DE GENERATION DES PHRASES PHONETIQUES.

G. TEP

Laboratoire CERFIA
Université Paul Sabatier
118, route de Narbonne
31077 TOULOUSE CEDEX

RESUME

Le système proposé génère, à partir d'une phrase graphique écrite en français, une phrase phonétique couramment parlée dans laquelle sont traités l'enchaînement, la liaison, l'élision, la dénasalisation et la pause.
Il utilise principalement un lexique de conversion où chaque enregistrement est composé d'un mot graphique, sa correspondance phonétique, sa catégorie syntaxique et les autres caractéristiques nécessaires à la conversion. Grâce à ce lexique, les homographes ont pu être traités convenablement.

La composante phonologique utilisée dans cette application est différente de celle faisant l'objet de l'article G. PERENNOU et G. TEP de ces mêmes journées. Mais il faut noter que les mécanismes généraux décrits ici peuvent, à des modifications mineures près, être utilisés avec d'autres composantes phonologiques.

Semblablement, il est possible d'accorder plus "d'activité" aux algorithmes de conversion graphème-phonème (cf [4] Divay-Guyomar 1977).

Dans l'exposé, nous avons plutôt montré comment le lexique peut contribuer à la conversion d'une phrase graphique en une chaîne de phonèmes.

10èmes JOURNÉES D'ÉTUDE SUR LA PAROLE
GRENOBLE - 30 MAI - 1er JUIN 1979

SYSTEME DE GENERATION DES PHRASES PHONETIQUES.

G. TEP

1. INTRODUCTION

Ce système a été créé en vue de la synthèse de la parole. Il génère automatiquement, à partir d'une phrase graphique, une phrase phonétique prononcée couramment. On tient compte donc, dans cette phrase phonétique obtenue, du traitement de la liaison, de l'enchaînement, de l'élision de la dénasalisation et de la pause.

Le schéma de principe de ce système est donné par la figure 1.

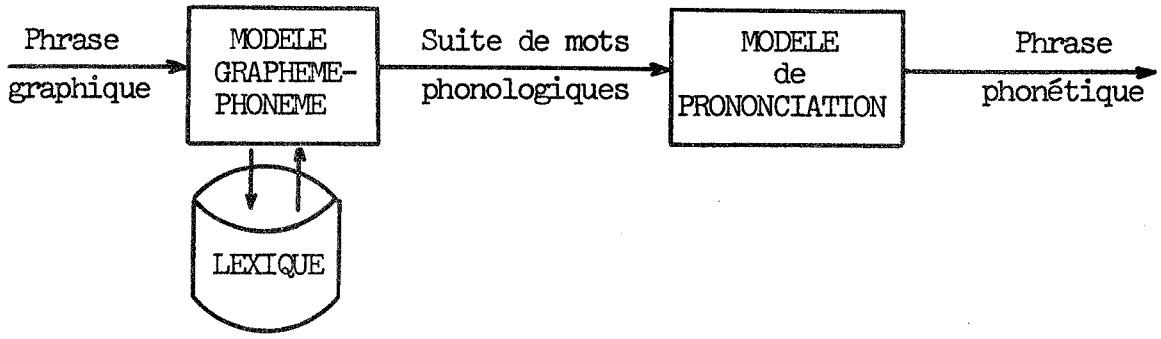


Figure 1 : Système de génération des phrases phonétiques.

A partir d'une phrase, ce système commence par la découper en mots graphiques. Pour chacun des mots obtenus, il consulte le lexique pour obtenir le mot phonétique et les différentes caractéristiques correspondant au mot à convertir. Par cette consultation, on connaît la catégorie syntaxique du mot, si le mot en question peut avoir ou non une liaison antérieure, une liaison postérieure ; s'il peut ou non s'élider en fin de mot etc... On fait ensuite le traitement sur le mot phonétique obtenu.

Ainsi, la phrase graphique est traitée mot par mot, pour obtenir à la sortie du modèle graphème-phonème, une suite de mots phonologiques avec les différents indicateurs de liaison, d'élision etc... qui serviront au modèle de prononciation. Celui-ci doit faire le traitement de liaison, de l'enchaînement, de la dénasalisation, de l'élision et de la pause pour arriver à générer la phrase phonétique couramment parlée.

2. ORGANISATION DU LEXIQUE

Pour pouvoir faire la conversion graphique-phonétique, notre système utilise un lexique [5] composé de deux milles mots courants du français. L'organisation de chaque enregistrement du lexique est représentée par la figure 2.

3. MODELE GRAPHEME-PHONEME

Ce modèle permet de générer, à partir d'une phrase graphique, une suite de mots phonologiques obtenus en consultant le lexique précédent. Il est composé de différents modules suivants :

3.1. Module de découpage de la phrase graphique en mots :

Dans la phrase graphique d'entrée, les mots sont séparés soit par des blancs, soit par des virgules. On peut aussi donners plusieurs phrases en entrée, chacune d'elles est reconnue par le point (.) ou par le point d'interrogation (?) qui la sépare des autres.

Le module de découpage reconnaît les mots par des blancs qui les séparent. Quand le mot contient une apostrophe ('), le module le scinde en deux mots sauf le cas de locution. Dans le cas où les mots sont séparés par une virgule (,) dans une phrase ou par un point (.) ou par un point d'interrogation (?) dans deux phrases successives, ce module découpe les mots en deux dont le premier contient le "," ou le "." ou le "?" en dernière position. Ceci servira dans le traitement de liaison et de la pause.

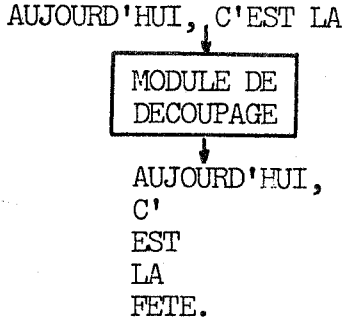


Figure 3. Exemple d'entrée-sortie du module de découpage.

3.2. Module de conversion graphème-phonème :

Ce module génère, à partir de chacun des mots graphiques fournis par le module de découpage, le mot phonologique correspondant grâce au lexique de conversion. Il est composé de trois modules dont le fonctionnement est représenté par la figure 4 suivante :

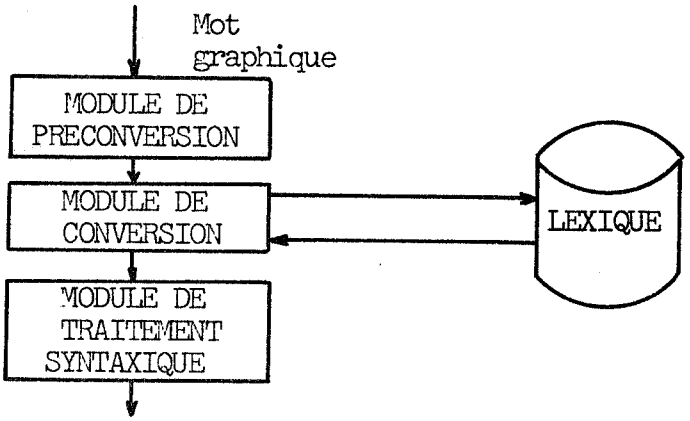


Figure 4 : Module de conversion graphème-phonème.

3.2.1. Module de préconversion :

Il teste si le mot graphique est suivi par une virgule, un point ou un point d'interrogation. Si ce cas est vrai, il positionne un indicateur qui

Racine graphique	Racine phonétique	Pointeur aux Désinences graphiques	Pointeur aux Désinences phonétiques	Liaison antérieure	Liaison postérieure	Elision	Catégorie syntaxique	Genre	Nombre	Personne	Critère sémantique
ancien	ɑ̃sj	A2	B2	*			J				
blanc	blɑ̃	A1	B1				J				U
est	ɛ			*	t		V		S	3	
étude	etydə	N1		*		ə	N	F			
modifi	mɔdif	V1	W1				V				
oiseau	wazo	N2		*			N	M			N
trop	tro				p		D				
⋮	⋮										

Figure 2 : Lexique de conversion

Les mots contenus dans la partie racine graphique (phonétique) ne sont pas tous complets. La plupart sont des radicaux du mot qui devront concaténer aux désinences pour obtenir le mot entier. Quelques unes de ces listes de désinences utilisées, concernant les noms (N), adjectifs (J) et verbes (V), sont :

Désinences graphiques	Désinences phonétiques
N1 (/ s)	
N2 (/ x)	
A1 (/ he s hes)	B1 (/ ʃə / ʃə)
A2 (/ ne s nes)	B2 (ɛ̃ ɛnə ɛ̃ ɛnə)
V1 (e es e ons ez ent)	W1 (i i i jɔ̃ je i)

Le caractère "/" dans les différentes listes signifie que le mot dans la partie "racine graphique" (phonétique) est complet il n'y a pas d'autres désinences à ajouter.

Le caractère "*" dans la partie "liaison antérieure" veut dire que le mot dans la ligne correspondante peut avoir une liaison antérieure.

Les caractères dans la partie "liaison postérieure" sont des caractères susceptibles de se lier avec le mot suivant.

Si le mot peut s'élider en fin de mot, on l'indique par la présence de "ə" dans la partie "élision".

Genre : Masculin (M), Féminin (F) ou Indéterminé (I)
Nombre : Singulier (S), Pluriel (P)
Personne : 1ère, 2ème, 3ème personne ; Cri.sém. : U(couleur),N(animé)...

servira pour la pause. Par conséquent, le mot en question ne peut pas avoir de liaison avec le mot qui le suit immédiatement. Ensuite, ce module supprime la virgule, le point ou le point d'interrogation et fait appel au module de conversion.

3.2.2. Module de conversion :

Ce module consulte le lexique de conversion pour sortir le mot phonétique correspondant. Il détermine si le mot graphique est un nom au singulier ou pluriel ; un adjectif masculin ou féminin, singulier ou pluriel ; un verbe à telle personne du singulier ou pluriel ; ou autres catégories syntaxiques. A partir de ces informations, il ira chercher les désinences dans la liste appropriée pour concaténer à la racine du mot. Nous donnons à la figure 5, un exemple de fonctionnement de ce module.

MOT GRAPHIQUE	CATEGORIE SYNTA-XIQUE	LIAISON ANTE-RIEURE	LIAISON POSTE-RIEURE	ELISION	RADICAL PHONETIQUE DANS LEXI-QUE	DESINEN-CE	MOT PHONO-LOGIQUE
ami	nom	oui	-	-	ami	-	xami(1)
culture	nom	-	-	ə	kyltyrə	-	kyltyr(2)
étude	nom	oui	-	ə	etydə	-	xetyd
table	nom	-	-	-	tablə	-	tablə
trop	adverbe	-	p	-	tro	-	trop(3)
blanche	adjectif	-	-	-	blɑ̃	fém.sing : ʃə	blɑ̃ʃə
moyen	adjectif	-	-	-	mwaj	mas.sing : ɛ̃	mwajɛ̃ə(4)
modifions	verbe	-	-	-	modif	1ère pers pluriel: jɔ̃	modifjɔ̃

Figure 5 : Fonctionnement du module de conversion

(1) le caractère "x" qui ne représente aucun phonème permet de savoir que le mot peut avoir une liaison antérieure.

(2) il y a élision de ə final du mot.

(3) on ajoute le caractère de liaison postérieure au mot.

(4) on ajoute le caractère "ə", qui ne représente aucun phonème, aux adjectifs terminés par les sons /ɛ̃/, /œ̃/, /ɔ̃/, /ɑ̃/ pouvant se dénasaliser.

3.2.3. Module de traitement syntaxique :

Grâce aux catégories syntaxiques du mot, on peut savoir si on fait ou non de liaison, de dénasalisation. Nous donnons ci-après quelques exemples de ce traitement syntaxique :

Suite de mots graphiques	Liaison	Pas de liaison	Sortie du module de traitement syntaxique
ils ont un livre	ilz‿ɔ̃ t‿œ̃ livr	-	ilz x ɔ̃ t xœ̃ n livrə (1)
ont ils un livre	-	ɔ̃til œ̃ livrə	ɔ̃t xil xœ̃ n livrə (2)
les oiseaux	lɛz‿wazo	-	lɛz xwazo
donnez les aux enfants	-	dɔnele ozãfã	donez le xoz xãfã
hommes illustres	-	ɔm ilystrə	ɔm xilystrə (3)
petit oiseau	pətit‿wazo	-	pətit xwazo (4)
sot et méchant	-	so e meʃã	so e meʃã (5)
je vais essayer	ʒə vɛz‿eseje		ʒə vɛz xeseye

Figure 6 : Traitement syntaxique de liaison

(1) on ajoute le phonème de liaison postérieure /z/ au mot [il] si celui-ci est suivi d'un verbe.

(2) /z/ n'est pas ajouté à [il] car celui-ci est suivi d'un article.

(3) il y a élision finale du mot[ɔ mə]. Ici la liaison est facultative ; nous avons choisi la prononciation courante sans faire la liaison.

(4) le phonème de liaison /t/ est ajouté à l'adjectif [peti] car il est suivi d'un nom.

(5) le /t/ est supprimé car l'adjectif [so] n'est pas suivi d'un nom.

Suite de mots graphiques	Dénasalisation	Pas de dénasalisation	Sortie du module de traitement syntaxique
au moyen âge	omwajɛnaʒ	-	o mwajɛ̃ʒ xaz (6)
le moyen terme	-	ləmwajɛ̃tɛrmə	lə mwajɛ̃ʒ tɛrmə
un bon ami	œ̃ bɔnami	-	œ̃n bɔ̃ʒ xami

Figure 7 : traitement syntaxique : dénasalisation

(6) on fait la dénasalisation sur les adjectifs terminés par les sons /ɛ̃/, /ɔ̃/, /œ̃ /, /ã/ et suivis par un nom commencé par une voyelle.

Dans le cas des homographes, on ne connaît pas, à priori, la prononciation ; nous générons, dans ce cas, autant de prononciations possibles pour chacun de ces homographes (grâce au lexique) suivi de leurs catégories syntaxiques. Ensuite, c'est le filtre syntaxique qui élimine les suites de mots syntaxiquement incorrects. Nous donnons à la figure 8 le filtrage syntaxique des homographes.

Suite de mots graphiques	Génération autant de prononciations possibles	Sortie du module de traitement syntaxique
Les poules du couvent couvent.	1.le pul dy kuvɑ̃ kuvɑ̃ (art.nom art nom nom) 2.le pul dy kuvɑ̃ kuvə (art nom art nom verbe) 3. le pul dy kuvə kuvɑ̃ (art nom art verbe nom) 4.le pul dy kuvə kuvə (art nom art verbe verbe)	le pul dy kuvɑ̃ kuvə

Figure 8 : filtrage syntaxique des homographes.

Dans cet exemple (figure 8), il y a ambiguïté de prononciation entre les deux homographes "couvent". Le filtre syntaxique analyse les catégories syntaxiques des mots contenues dans chacune des listes de la façon suivante :

liste 1 : Art nom art nom nom incorrect
liste 2 : Art nom art nom verbe correct
liste 3 : Art nom art verbe nom incorrect
liste 4 : Art nom art verbe verbe incorrect

Donc après filtrage, seule la liste 2 "le pul dy kuvɑ̃ kuvə" est correcte.

4. MODELE DE PRONONCIATION

Ce modèle génère la phrase phonétique de prononciation courante grâce aux divers indicateurs de liaison, de dénasalisation contenus dans la suite de mots phonologiques, sortie du modèle graphème-phonème. Il traite aussi les élisions et les pauses. Il est composé de trois modules représentés par la figure 9 :

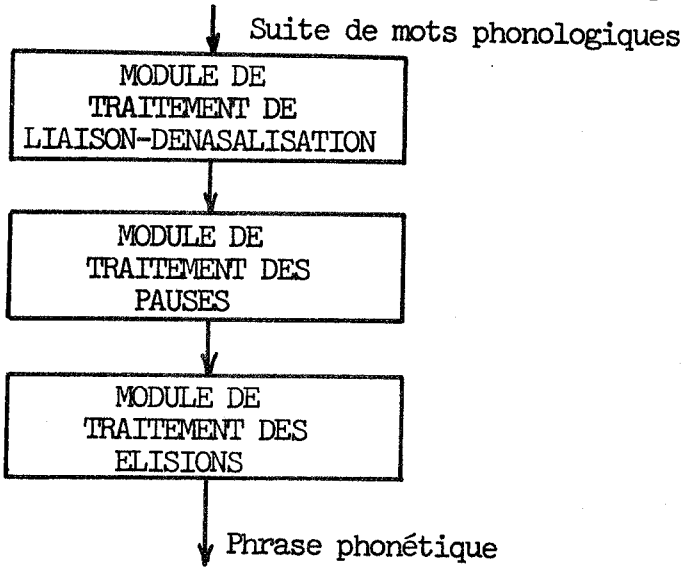


Figure 9 : Modèle de prononciation.

4.1. Module de traitement de liaison-dénasalisation :

Ce module s'exécute de la façon suivante :

4.1.1. Cas de liaison-dénasalisation :

Si le mot commence par le caractère x, on peut savoir que ce mot peut lier au mot précédent. Cette liaison ou enchaînement, est possible si le dernier caractère du mot précédent est l'un des phonèmes suivants : /z/,/t/,/n/, /r/,/p/,/l/,/s/,/k/. Si ce dernier caractère est le chiffre 9, alors il y a dénasalisation du mot précédent.

	Phrases graphiques	Sortie du modèle graphème-phonème	Sortie du module liaison-dénasalisation
LIAISON	Les amis trop aimable	lez xami trop xɛmablə	lezami tropɛmablə
DENASALISATION	Au moyen âge un bon ami	o mwajɛ̃9 xaz œ̃n bɔ̃9 xami	omwajɛnaz œ̃bɔnami

Figure 10 : Liaison-dénasalisation.

4.1.2. Cas sans liaison ni dénasalisation :

Les mots, ne commençant pas par le caractère x, ne peuvent pas avoir de liaison avec le mot précédent. Il faut donc supprimer le dernier caractère du mot précédent si ce caractère est l'un des phonèmes de liaison /z/,/t/,/n/, /p/ et 9 (pour la dénasalisation).

	Phrases graphiques	Sortie du modèle graphème-phonème	Sortie du module liaison-dénasalisation
PAS DE LIAISON	les garçons trop jolis	lez garsɔ̃ trop ʒoli	le garsɔ̃ tro ʒoli
PAS DE DENASALISATION	un bon livre	œ̃n bɔ̃9 livrə	œ̃ bɔ̃ livrə

Figure 11 : Pas de liaison ni de dénasalisation.

4.2. Module de traitement des pauses :

Les pauses (ou silences) sont des arrêts dans la production de la parole. Pour ce traitement automatique, on insère une pause représentée par un caractère blanc (␣) entre les mots ou les phrases à chaque fois que l'on rencontre les ponctuations suivantes : virgule (,), point (.), ou point d'interrogation (?). Les pauses produites par d'autres causes (catégorie syntaxique des mots, nature syllabique) ne sont pas prises en compte, pour le moment, par notre système de génération.

4.3. Module de traitement des élisions :

Le premier traitement de l'élision finale du mot a été fait lors de l'étape précédente. Ce traitement n'est pas applicable dans le cas où les mots sont enregistrés par leurs radicaux (partie élision du lexique est vide).

Le deuxième traitement, qui est l'objet de ce paragraphe, a donc pour but de traiter toutes les élisions possibles restant dans la phrase phonétique. Ces élisions se feront suivant que le phonème /ə/ est en :

- syllabe initiale du mot,
- syllabe intérieure du mot,
- syllabe finale du mot,
- monosyllabe.

Le fait de traiter ces élisions après le traitement des pauses permet de simplifier l'algorithme de fonctionnement. En effet, à la sortie du module des pauses, la plupart des mots sont concaténés entre eux, donc l'élision de /ə/ en monosyllabe se fait rare, de même la présence de /ə/ en syllabe initiale et finale est moins fréquente. L'essentiel du traitement se ramène donc au traitement de /ə/ en syllabe intérieure du mot. D'autre part, après le traitement des pauses, nous avons décidé de ne pas traiter le /ə/ en syllabe finale du mot car très souvent, on entend prononcer le /ə/ en cette position. Il ne nous reste donc à traiter que le cas où le /ə/ est en syllabe initiale ou intérieure du mot.

4.3.1. Elision en syllabe initiale du mot :

Notre système supprime le /ə/ en cette position.

PHRASE GRAPHIQUE	PHRASE PHONETIQUE	TRAITEMENT DES ELISIONS
ce que	səkə	sk ə
je m'en vais	ʒəmɑ̃vɛ	ʒmɑ̃vɛ
ce n'est pas ça	sənɛpasa	snɛpasa

Figure 12 : Elision en syllabe initiale du mot.

4.3.2. Elision en syllabe intérieure du mot :

L'algorithme commence par repérer la position de /ə/. Si le /ə/ est précédé par une seule consonne, il y a élision sauf s'il est suivi de deux phonèmes suivants [lj] et [rj].

PHRASE GRAPHIQUE	PHRASE PHONETIQUE	TRAITEMENT DES ELISIONS
samedi	samədi	samdi
grande gloire	grɑ̃dəglwarə	grɑ̃dglwarə
on me le refuse	ɔ̃mələrəfyzə	ɔ̃mlərfyzə
vous les aimeriez	vulezɛmərje	vulezɛmərje
batelier	batəlje	batəlje

Si le /ə/ est précédé par deux consonnes consécutives, le /ə/ reste, il n'y a pas d'élision.

Exemple : autrement ɔtrəmã

fortement fɔrtəmã

5. RESULTATS

Nous donnons ci-dessous le résultat du traitement automatique du système de génération des phrases phonétiques prononcées couramment à partir d'une phrase graphique en entrée.

```
ENTREZ VOTRE PHRASE LCRITE ?
?AU MOYEN AGE,LES HOMMES PREFERENT LES OISEAUX.  €

## SUITE DES MOTS PHONETIQUES-SORTIE DU MODELE GRAPHEMES.PHONEMES :

0 MWA*59 XAJE L6Z X&M FR6F7RET L6Z XWAZO

## PHRASE PHONETIQUE-SORTIE DU MODULE DE LIAISON.DENASALISATION :

0 MWA*7NAJE L6Z&M FR6F7RE L6ZWAZO

## PHRASE PHONETIQUE-SORTIE DU MODULE DES PAUSES :

0MWA*7NAJE L6Z&MFR6F7REL6ZWAZO

## PHRASE PHONETIQUE FINALE-SORTIE DU MODULE DES ELISIONS :

0MWA*7NAJE L6Z&MPR6F7RL6ZWAZO        omwajenaʒə lezɔ̃mpreferlezwazo
```

6. CONCLUSION

Le système de génération des phrases phonétiques que nous avons présenté dans ce travail, est très général au point de vue de traitement de la liaison, de la dénasalisation et de l'élision. Il est indépendant des mots existant dans le lexique. Par conséquent, la décision d'ajouter des mots nouveaux dans le lexique n'entraîne pas de perturbation dans notre système.

Cependant, le module de traitement des pauses n'est pas complet, puisqu'il ne traite que le cas où l'on rencontre les signes de ponctuation. On pourra aussi décider d'insérer une pause dans la phrase en s'appuyant sur la catégorie syntaxique des mots, mais ce n'est pas du tout systématique. Il faut tenir compte aussi du nombre de syllabes ainsi que du rythme de prononciation de la phrase car, plus le locuteur prononce chaque son lentement, plus il doit reprendre souvent son souffle.

Enfin, le temps de consultation du lexique n'est pas négligeable dans ce système de génération. Cette perte de temps est compensée par la possibilité de traiter des homographes, grâce au lexique où on enregistre, entre autres, la catégorie syntaxique du mot. Sans le filtrage syntaxique, il est impossible de donner la prononciation correcte de ces homographes.

BIBLIOGRAPHIE

(1) L. BARRET
Méthode de prononciation du français standard.
Didier. 1968

(2) F. DELL
Les règles et les sons. Introduction à la phonologie générative.
Collection savoir. Herman. 1973

(3) P. DELATTRE
Studies in French and Comparative phonetics.
Mouton & Cie. 1966

(4) M. DIVAY. M. GUYOMARD
Conception et réalisation sur ordinateur d'un programme de transcription graphémo-phonétique du français.
Thèse de 3è cycle. Rennes. Avril 1977

(5) G. GOUARDERES
Organisation du lexique en vue de l'analyse de la parole en continu.
Rapport diplôme CNAM. Toulouse. Janvier 1977

(6) R. JAKOBSON, G. FANT, M. HALLE
Préliminaires to speech Analysis.
1951

(7) A. MARTINET, H. WALTER
Dictionnaire de la prononciation française dans son usage réel.
France expansion. 1973

(8) G. PERENNOU, J.P.HATON
Reconnaissance automatique de la parole.
8ème école d'été d'Informatique de l'AFCET. NAMUR (Belgique) 1978

(9) G. TEP
Contribution à l'étude phonologique d'un système d'analyse et de synthèse de la parole continue.
Thèse de 3ème cycle (informatique). Toulouse. Septembre 1978

THEME II

LA FORMALISATION DU LEXIQUE ET DE LA PHONOLOGIE

en vue de l'application à la reconnaissance et à la synthèse de la parole

10èmes JOURNÉES D'ÉTUDE SUR LA PAROLE

GRENOBLE - 30 Mai - 1er Juin 1979

Titre : LES DICTIONNAIRES EN FORMES COMPLETES ET LEUR UTILISATION DANS LA TRANSFORMATION LEXICALE ET SYNTAXIQUE CORRECTE DE CHAINES PHONETIQUES.

Auteurs : MM. ANDREEWSKY A.*, BINQUET J.P.**, DEBILI F.*, FLUHR C.*, HLAL Y.*, LIENARD J.S.**, MARIANI J.**, POUDEROUX B.* -

Domiciliation: (*) I.N.S.T.N. - SACLAY et ERA 430 du CNRS

(**) LIMSI - BP 30 - 91406 ORSAY CEDEX

RESUME

Des dictionnaires en formes complètes (170.000 formes) ont été réalisés et utilisés pour obtenir à partir de chaînes phonétiques exactes des chaînes lexicalement syntaxiquement correctes, cela dans des temps de calcul très courts.

Actuellement des expériences sont en cours pour traiter le cas de chaînes phonétiques entachées d'un faible taux d'erreurs.

LES DICTIONNAIRES EN FORMES COMPLETES ET LEUR UTILISATION DANS LA TRANSFORMATION LEXICALE ET SYNTAXIQUE CORRECTE DE CHAINES PHONETIQUES.

MM. ANDREEWSKY A., BINQUET J.P., DEBILI F., FLUHR C ., HLAL Y., LIENARD J.S., MARIANI J., POUDEROUX B.,

---

SUMMARY

Dictionnaries in full forms (170.000 forms) were built and used to obtain from correct phonetical strings, strings which are syntactically correct.
The corresponding algorithms are performants and can be used in real time mode. At present experiments are conducted on phonetical strings with a low error rate.

10èmes JOURNÉES D'ÉTUDE SUR LA PAROLE
GRENOBLE - 30 MAI - 1er JUIN 1979

Répéter le titre : LES DICTIONNAIRES EN FORMES COMPLETES ET LEUR UTILISATION DANS LA TRANSFORMATION LEXICALE ET SYNTAXIQUE CORRECTE DE CHAINES PHONETIQUES

Auteurs: MM. ANDREEWSKY A., BINQUET J.P., DEBILI F., FLUHR C., HLAL Y., LIENARD J.S., MARIANI J., POUDEROUX B.,

I - INTRODUCTION :

On se propose de décrire dans la présente note, une expérience d'aide linguistique à la reconnaissance automatique de la parole, qui permet, par ailleurs de vérifier dans une certaine mesure la pertinence des modèles linguistiques utilisés.

On sait que la reconnaissance de la parole ne peut se faire sans que s'effectuent à un degré plus ou moins important des traitements linguistiques de tous niveaux, tant syntaxiques que sémantiques, qui, d'une façon générale sont intégrés à tout le processus de la compréhension. On peut montrer d'une façon simple que le décryptage purement lexical d'une chaîne phonétique est tout à fait insuffisant pour l'identification de la dite chaîne.

Ainsi en atteste une phrase comme :

appelez le témoin!
qui transcrite en chaîne phonétique puis retranscrite en chaîne lexicale engendre outre la phrase initiale, les expressions :
"ah pelez le thé moins,ou encore, happe les le thème oins", etc ...
(ce problème est traité dans [11, 12] au moyen d'une méthode purement morphologique).

On peut prendre de cette façon n'importe quelle phrase de la langue et obtenir ainsi des chimères linguistiques lexicalement correctes mais syntaxiquement ou sémantiquement incohérentes.

On sait aussi que la dimension sémantique dont on ne tient pas compte dans cet article, peut être d'un secours réel pour certaines situations homonymiques syntaxiquement indiscernables. Ainsi en atteste des phrases comme : "Voilà un drole de pieu " et "vois là un drôle de pieux" ou encore : "le jeu curieux des marionnettes" ou "le jeu curieux des maris honnêtes" que bon nombre de femmes ou d'hommes comprendront à leur manière selon leur mode de pensée propre.

II - LES CONDITIONS DE L'EXPERIENCE

A - Du point de vue des conditions expérimentales acoustiques on a supposé parfaites (1) l'émission du locuteur, la saisie acoustique de cette émission et sa transformation en chaîne phonétique. Cela suppose l'absence de substitutions, d'ajouts, d'effacements. Comme nous allons le voir cette hypothèse de la correspondance parfaite son-phonème, bien que très simplificatrice, conduit encore à des traitements linguistiques très complexes.

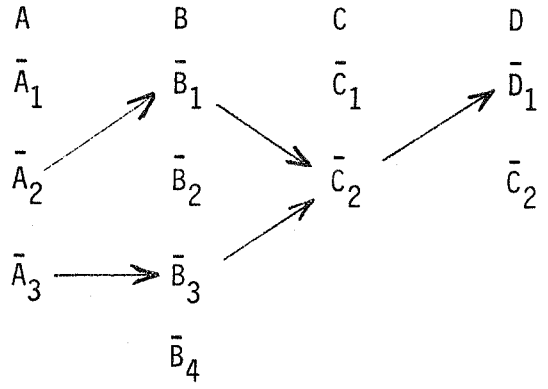
(1) hypothèse d'école qui n'est réalisée encore par aucun laboratoire, cependant actuellement dans les expériences en cours on tient compte des écarts par rapport à cette hypothèse.

B - La taille de l'univers lexical.
Un des buts de l'expérience était de vérifier qu'en utilisant des lexiques volumineux, on ne risquait pas de rendre le problème combinatoirement impossible. Deux préoccupations ont donc été prises en compte. Tout d'abord utiliser un lexique aussi volumineux que possible, (on a pris un dictionnaire de 170.000 formes) et ensuite élaborer les algorithmes dans les conditions du temps réel.

III - ASPECT LINGUISTIQUE DE L'EXPERIENCE.

Pour effectuer l'expérience on a fait appel à la syntaxe positionnelle (et catégorisation grammaticale correspondante) obtenue par une méthode d'apprentissage et décrite dans [1,2 , 5, 7]
Elaborés en vue de pouvoir traiter les ambiguïtés grammaticales du discours qui apparaissent à la suite de l'utilisation d'un même mot dans des fonctions linguistiques différentes, les analyseurs syntaxiques construits (cités plus haut) ont été pensés initialement en vue d'applications tout à fait étrangères à la reconnaissance de la parole. Traitant essentiellement les ambiguïtés qui peuvent être levées par un calcul sur le contexte "proche" de chaque entité lexicale, leur vocation première était de résoudre les problèmes linguistiques qui se posent en documentation automatique dans sa forme la plus élaborée, l'indexation automatique.
En relation avec cette application plusieurs types d'analyseurs syntaxiques ont été construits [1, 2, 5, 7] et dans la présente expérience, seuls ont été utilisés les moins sophistiqués d'entre eux, ceux basés sur l'emploi de matrices de précédence binaire et ternaire [cf.5].

- Cette matrice qui a été obtenue par apprentissage d'abord à partir d'un premier texte assez court, de mille mots environ analysés manuellement, puis d'une suite de textes analysés automatiquement et corrigés manuellement, est utilisée en documentation automatique, comme suit :
chaque phrase est tout d'abord analysée morphologiquement à l'aide d'un dictionnaire qui à chaque mot attribue l'ensemble de ses valeurs grammaticales hors contexte. Ensuite à l'aide de la matrice on établit la pertinence syntaxique de chaque phrase en vérifiant s'il existe au moins un chemin dit de "parcours syntaxique", ce que l'on peut illustrer de la façon suivante : soit quatre mots A, B, C, D et respectivement $\bar{A}_1$, $\bar{A}_2$, $\bar{A}_3$ les valeurs grammaticales de A, $\bar{B}_1$, $\bar{B}_2$, $\bar{B}_3$, $\bar{B}_4$ celles de B, $\bar{C}_1$, $\bar{C}_2$ celles de C, $\bar{D}_1$, $\bar{D}_2$ celle de D. Alors un parcours syntaxique possible validant syntaxiquement la chaîne A B C D sera obtenu si, à l'aide de la matrice de précédence, on peut parvenir à construire un chemin continu qui va d'une des valeurs de A à une des valeurs de D. Deux de ces chemins ssont indiqués sur le schéma, mais un seul suffit pour valider la chaîne A B C D ; Ce type de traitement par chaînage continu permet de tenir compte aussi bien des propriétés globales que locales du discours.



L'aspect local du traitement syntaxique permet de réaliser facilement des analyses dites par défaut (reconnaissance grammaticale d'un mot inconnu du dictionnaire) . Il convient de souligner que pour des langues analytiques (1) comme l'anglais, le bulgare, l'espagnol, le français, ... ces relations syntaxiques sont déjà très résolvantes. Actuellement pour le français elles permettent, la reconnaissance de quelques cent cinquante categories définies en relation avec les diverses expériences linguistiques antérieures. Le nombre de ces catégories est actuellement pratiquement stationnaire.

Pour tester la pertinence linguistique du modèle succintement décrit ici, ainsi que son utilisation pour la reconnaissance de la parole, on a voulu vérifier qu'une chaîne phonétique quelconque compactée, obtenue à partir d'une phrase acceptée de la langue, peut être reconstituée en utilisant les dictionnaires et la syntaxe de ce modèle.

Nous allons maintenant décrire l'expérience effectuée qui comporte les trois points suivants : constitution du texte phonétique, construction du dictionnaire dit universel, et principe de l'algorithme.

IV.- LE TEXTE PHONETIQUE.

Sur un texte correct de 1800 mots on a effectué successivement les opérations suivantes :

- découpage du texte en mots
- phonétisation mots par mots à l'aide d'algorithmes mis au point au LIMSI (*). La phonétisation mot par mots entraine une absence de prise en compte des liaisons et des "e" terminaux.
- les ponctuations ont toutes été conservées
- on a alors éliminé tous les blancs en vue d'obtenir une chaîne phonétique continue appelée texte phonétique.

V. - LE DICTIONNAIRE UNIVERSEL

A l'aide d'un dictionnaire initialement utilisé en documentation automatique, comprenant 22.000 mots en regard desquels figurent les propriétés grammaticales hors contexte de chaque mot, on a construit un dictionnaire de 170.000 formes (féminin, pluriel, conjugaison complète)

(*) B. PROUTS : Traduction phonétique de textes écrits en Français - JEP 1979

(1) c'est-à-dire des langues pour lesquelles les relations entre mots d'une phrase s'obtiennent grâce aux positions respectives de ces mots et de certains mots outils comme les prépositions, les conjonctions, les articles, les pronoms...

de la langue française . Ce dictionnaire a été phonétisé permettant d'obtenir ainsi un dictionnaire dit universel dans lequel en regard de chaque forme on trouve la forme phonétique correspondante obtenue automatiquement , les catégories grammaticales possibles hors contexte ainsi que pour les substantifs et adjectifs, le genre, le nombre; pour les verbes, la personne, le groupe, le temps, la transitivité, la racine.

VI.- PRINCIPE DE L'ALGORITHME

Il comprend les phases suivantes :

- Une recherche lexicale phonétique dans le dictionnaire universel et un contrôle syntaxique parallèle qui est indispensable si l'on veut éviter un nombre considérable de chaînes lexicales. Du point de vue structure du système, l'organisation arborescente du dictionnaire permet de ne faire qu'un seul appel disque au plus pour obtenir tous les mots inclus les uns dans les autres susceptibles de débuter une chaîne de phonème. Par exemple à partir de la chaîne "j'ai mal" écrite en phonétique et compactée, on obtiendra simultanément "j" "geai" "jet", "gemme".

A la suite de cette opération on obtient un ensemble de chaînes syntaxiquement correctes qui,entre deux points de ponctuation,peuvent être au nombre de plusieurs milliers. Toutefois entre deux portions de chaînes inambiguës et communes à tous les chemins,le nombre de chemins est petit et dépasse rarement la vingtaine.

Pour choisir les chemins linguistiquement les plus pertinents on a élaboré une série de critères de choix.

premier critère : Entre deux chemins également possibles, c'est-à-dire syntaxiquement corrects, en premier lieu, on choisit celui qui possède le plus petit nombre de mots [12]

Exemple : l'arrêté accordant le permis (5 mots)
l'arrêté à corps dans le père mis (8 mots)

second critère : Les matrices de précédence utilisées étant fréquentielles, on associe aux différents chemins des poids qui permettent de définir une hiérarchie entre les solutions syntaxiques proposées, lorsque le nombre de mots dans les diverses chaînes candidates est le même.

troisième critère : On envisage de faire appel à la fréquence lexicale qui permet d'effectuer le bon choix par exemple dans le cas suivant :

- les constructions frappées d'alignement et $\left\{\begin{matrix}\text{celles}\\ \text{sels}\end{matrix}\right\}$ situées dans les périmètres ...

On comprend en effet que le substantif étant plus fréquent que le pronom démonstratif, c'est l'option "sels" qui sera syntaxiquement choisie. Tandis que le mot "celles" étant lexicalement plus fréquent que "sels" la fréquence lexicale donnera priorité à l'option "celles".

REMARQUE I. - Notons un cas intéressant, la chaîne "les frais généraux amputés" qui est fournie par l'algorithme d'une manière correcte avec les deux analyses grammaticales suivantes :

Article , Adjectif, Substantif, Participe passé,
Article, Substantif, Adjectif, Participe passé.

REMARQUE 2. - Dans les expériences effectuées les locutions sont traîtées comme un seul mot (il s'agit essentiellement des locutions prépositives et conjonctives ; en vue de, par rapport à, au fur et à mesure que, etc....)

- Enfin des règles d'accord en genre et nombre sont utilisées pour choisir à l'intérieur des chemins syntaxiques corrects les formes dont le masculin-féminin singulier-pluriel sont phonétiquement identiques.

VII- DISCUSSION DE L'EXPERIENCE EFFECTUEE

Sur l'ensemble du corpus de 1800 mots, 75 mots ont donné lieu à des erreurs ou ambiguïtés non résolues, soit environ moins de 5%. Ont été écarté du bilan les erreurs dues au lexique (erreurs d'orthographe dans le dictionnaire) ou des erreurs dues à l'emploi de règles syntaxiques fausses (par exemple, absence de "beurre" en tant que substantif dans le dictionnaire et de ce fait présence dans la syntaxe de la règle impossible article défini * verbe conjugué).

Dans la discussion qui suit, on fait appel à la notion de relation lexicale semantique décrite en peu de mots dans la conclusion et dans 3 .

Du point de vue linguistique, les 75 cas erronés s'analysent comme suit :

- Non reconnaissance des homonymes (13 en tout) du type "plan ⟷ plants, heures ⟷ heurts, ère ⟷ air ⟷ erre ⟷ hère ⟷ haire", etc.... la syntaxe pour séparer ces cas là étant inopérante (ainsi que les accords en genre et en nombre). Par contre les fréquences et les relations lexicales sémantiques peuvent être d'un grand secours.

- non reconnaissance de certains singuliers - pluriels, insolubles sans une analyse très approfondie de l'accord en genre et nombre, ou même essentiellement insolubles : On en trouve 21 en tout.

  Exemple : plans / plan d'execution, demande / demandes de permis de construire, en cas de fausse déclaration / fausses déclarations, etc ...

- non reconnaissance de l'accord pour l'adjectif postérieur ; les règles utilisées ne donnent jamais de mauvais accord mais maintiennent des ambiguïtés. Cela s'est produit 10 fois. Par exemple (voir phrase 1 ; périmètre de protection des monuments historique / historiques).

- Erreurs de caractère syntaxique. Elles sont au nombre de 13 dont 3 ne peuvent être corrigées à l'aide des fréquences de mots.

  Exemples : Préfet de la scène (au lieu de Seine)
  les baisses ont équipé (au lieu de "baies sont équipées")
  les autres erreurs: "et sels" au lieu de "et celles" ; "shah cané" au lieu de "chaque année" ; "poux rassurés" au lieu de "pour assurer" peuvent être levées par les fréquences.

- Erreurs dues à l'emploi du critère "chaîne la plus courte à syntaxe correcte" - 4 erreurs en tout.

  les deux suivantes :

  - obstrué par des poussières et décors étrangers (au lieu de

"des corps étrangers")
- un temps froid et venté (au lieu de éventé) sont une exception à ce critère.

Elles ne peuvent être levées par les fréquences mais elles le sont par des relations lexicales sémantiques.

- Enfin on trouve 4 ambiguïtés dues à des structures de listes (cf.P.8) phrases 3,4, et 5 ; 2 erreurs dues à la non reconnaissance du sujet et 4 ambiguïtés dues à la mauvaise reconnaissance du référent du pronom "leur".

CONCLUSION :

L'expérience effectuée a montré :
- qu'il existait d'ores et déjà des analyseurs syntaxiques très puissants qui permettent d'envisager une aide efficace à la reconnaissance de la parole dès l'instant que le taux de reconnaissance acoustique est suffisamment important.
- que cette aide linguistique fonctionne sur des lexiques ayant l'importance normale d'une langue naturelle et sans contraintes sur la syntaxe utilisée.
- que la performance, du point de vue temps calcul de l'unité centrale est inférieur à la minute pour les 1.800 mots, et du point de vue temps réel, dans les conditions d'exploitation du centre de calcul (en multiprogrammation), est de l'ordre de 90 mots minutes, ce qui correspond à une dictée normale. Mais ce temps pour un système comme celui d'ORSAY permettrait simultanément de traiter au même débit plusieurs dizaines de dictées.
- cette expérience, ainsi que les expériences en cours, ont montré la nécessité de conduire d'une manière conjointe les problèmes acoustiques et linguistiques. En particulier on peut penser que la mise en oeuvre d'algorithmes linguistiques peut dans certains cas contourner les difficultés techniques liées aux problèmes acoustiques.

Il convient cependant de souligner, que toutes les méthodes linguistiques ne sont pas également adéquates tant pour résoudre les problèmes liés à la reconnaissance de la parole, que ceux relatifs à l'indexation automatique ou à la traduction automatique.

Entre autres, les méthodes utilisant des arborescences sont certainement moins souples, moins adaptées et moins rapides que celles que nous avons utilisées. De plus, elles sont difficilement conciliables avec les méthodes dites d'apprentissage. Ce problème de l'apprentissage est d'ailleurs essentiel. Les expériences menées à l'aide des divers analyseurs syntaxiques sur des corpus différents [1, 2, 5] , ont montré que les syntaxes des sous langues d'une langue naturelle peuvent être très différentes (jusqu'à être parfois contradictoires). Il s'ensuit que l'apprentissage tant au locuteur qu'à la sous langue employée est à envisager très sérieusement pour un certain nombre d'applications comme l'enregistrement de textes de même profil mis à la disposition d'un locuteur déterminé.

Enfin, en vue d'améliorer l'aide à la décision pour les phrases données comme correctes par l'ensemble des filtres grammaticaux et syntaxiques, on envisage actuellement l'utilisation d'un dictionnaire lexical sémantique, construit automatiquement à partir d'un corpus déterminé.

Par exemple entre les chaînes :

"Il a posé un seau sur le sol", "il apposait un sceau sur une lettre"
Le dictionnaire lexical sémantique permettra de prendre une décision grâce aux relations ci-dessous obtenues automatiquement par filtrage syntaxique de gros corpus :

apposer ⟶ un sceau; poser ⟶ un seau

sceau ——— sur ⟶ lettre
seau ——— sur ⟶ sol

Ces relations peuvent par ailleurs être pondérées par des fréquences.
Le mode de construction et la syntaxe d'utilisation de ce dictionnaire sont décrits dans la communication I F I P Août 1977. [3].

Notons enfin qu'une expérience analogue a été effectuée par Miclet et Rougeot, mais que faute d'un nombre suffisant de catégories grammaticales elle n'a pu conduire à des résultats aussi concluants.[13].

L'expérience a permis de tester la qualité des analyseurs syntaxiques employés et a montré dans quels cas les paramètres sémantiques étaient indispensables en particulier sous la forme de fréquences lexicales, et de relations lexicales-sémantiques.

- Pour se rapprocher des conditions réelles de la saisie vocale on étudie actuellement au LIMSI plusieurs méthodes dont certaines basées sur l'apprentissage en vue d'améliorer la qualité de la sortie de l'analyseur du signal acoustique.

BIBLIOGRAPHIE

1 - A. ANDREEWSKY, C. FLUHR - Expérience de constitution d'un programme d'apprentissage pour le traitement automatique du langage - Note C.E.A.1606 (1) et (2), dec. 1972 - Nov. 1973.

2 - A. ANDREEWSKY, C. FLUHR - Apprentissage - Analyse automatique du langage application à la documentation - Dunod, Doc. Linguistique quantitative, n° 21, 1973.

3 - A. ANDREEWSKY, F. DEBILI, C. FLUHR - Computational learning of semantic relations for the generation and automatic analysis of content - IFIP, Congress TORONTO, Août 1977.

4 - A. ANDREEWSKY, F. DEBILI, C. FLUHR, J.S. LIENARD, J. MARIANI - Une Expérience d'aide linguistique à la reconnaissance automatique de la parole - Note C.E.A.N-2055 - Oct. 1978.

5 - M. BAUDRY, B. DUPEYRAT - Analyse du signal vocal. Utilisation des extréma du signal et de leurs amplitudes. Détection du fondamental et recherche des formants. Actes des 7èmes Journées du GALF, Nancy Mai 1976.

6 - F. DEBILI - Traitements syntaxiques utilisant des matrices de précédence fréquantielles construites automatiquement par apprentissage. Thèse Doct. Ing. - PARIS VII, Sept. 1977.

7 - B. DUPEYRAT - Reconnaissance de la parole. Méthode des passages par zéro du signal. Reconnaissance automatique de voyelles isolées. Thèse 3ème Cycle - PARIS VI, 1975.

8 - C. FLUHR - Algorithme à apprentissage et traitement automatique des langues. - Thèse d'Etat - Juin 1977, ORSAY

9 - J.P. HATON - Contribution à l'analyse, la paramètrisation et la reconnaissance automatique de la parole. - Thèse d'Etat, Université de NANCY I, Janvier 1974.

10 - J.P. HATON - Reconnaissance analytique de la parole aux niveaux acoustique morphologique, lexical et syntaxique. - Sept. 1976.

11 - J.S. LIENARD - Analyse, synthèse et reconnaissance de la parole - Thèse d'Etat, avril 1972.

12 - J.J. MARIANI - Contribution à la reconnaissance dela parole utilisant la notion de spectre différentiel. - Thèse Doct. Ing. , Mars 1977.

13 - L. MICLET, B. ROUGEOT - Transcription phonétique, orthographique des phrases françaises - Note tech. CEI/CSI/23, août 1972 - CNET, Lannion

14 - J. MARIANI - Formalisation du lexique et des règles phonologiques dans le système ESOPEØ - 10e J.E.P. du GALF - GRENOBLE - Mai-Juin 1979.

DANS LA PLANCHE CI-DESSOUS LES ACCENTS NE FIGURENT PAS, MAIS SONT PRIS EN COMPTE DANS LE TRAITEMENT.

1- LES EXEMPTIONS PREVUES PAR LE PRESENT ARRETES NE SONT PAS APPLICABLES AUX TRAVAUX CONCERNANT LES CONSTRUCTIONS FRAPPEES D'ALIGNEMENT ET SEL/SELLE/SELS SITUE/SITUEE/SITUEES/SITUES DANS LE PERIMETRE DE PROTECTION DES MONUMENTS HISTORIQUES/HISTORIQUE ET DES SITES CLASSES.

2- POUR LES CONSTRUCTIONS EDIFIEES SUR LE TERRITOIRE DE LA VILLE DE PARIS, LA CONSULTATION S'EFFECTUE AU LIEU, JOUR/JOURS ET HEUR/HEURES/HEURS/HEURT/HEURTS FIXE/FIXEES/FIXES PAR ARRETES DU PREFET DE LA CENE/SCENE.

3- LE PRESIDENT DE LA CONFERENCE PEUT ENTENDRE, POUR LES AFFAIRES QUI LES CONCERNENT, TOUTE/TOUTES AUTORITE/AUTORITES OU PERSONNE/PERSONNES COMPETENTE/COMPETENTES POUR EMETTRE UN AVIS SUR CES AFFAIRES.

4- DES L'AFFICHAGE A LA MAIRIE D'UN EXTRAIT DE LA DECISION OCTROYANT LE PERMIS-DE-CONSTRUIRE ET JUSQU'A L'EXPIRATION D'UN DELAI DE UN AN ET UN MOIS APRES CET AFFICHAGE, TOUTE PERSONNE INTERESSEE PEUT CONSULTER LES PIECES SUIVANTES DU DOSSIER, DEMANDE/DEMANDES DE PERMIS-DE-CONSTRUIRE, PLAN/PLANS/PLANT/PLANTS DE SITUATION, PLAN DE MASSE/PLANS DE MASSE, PLAN:PLANS/PLANTS DES FACADES, ARRETES/ARRETE ACCORDANT LE PERMIS-DE-CONSTRUIRE ET EVENTUELLEMENT, ARRETES/ ARRETE DE DEROGATION, CONTRAT/CONTRATS OU DECISIONS/DECISION EN MATIERE D'INSTITUTION DE SERVITUDE DITE DE COURS COMMUNE/COMMUNE - OU DE MINORATION DE DENSITE SURE DES FONDS VOISINS.

5- LES TRAVAUX NON EXEMPTES PAR L'ARTICLE PREMIER ET CEUX QUI NE SONT PAS ASUJETTIS A LA PROCEDURE NORMALE DU PERMIS-DE-CONSTRUIRE NE PEUVENT ETRE ENTREPRIS QU'APRES COMMUNICATION DU PLAN DE MASSE AU CHEF DE SERVICE DEPARTEMENTAL DE L'URBANISME ET DE L'HABITATION DANS LE DELAI DE VAIN/VAINS JOUR/JOURS AVANT LEUR/LEURS EXECUTION/EXECUTIONS OU LA PASSATION DES MARCHES.

10èmes JOURNÉES D'ÉTUDE SUR LA PAROLE

GRENOBLE - 30 MAI - 1ER JUIN 1979

RECHERCHE LEXICALE EN RECONNAISSANCE DE LA PAROLE. RESOLUTION PAR UNE METHODE D'ANALYSE SYNTAXIQUE ISSUE D'UNE INFERENCE GRAMMATICALE AUTOMATIQUE.

BAUDRY, M., DUPEYRAT, B., LEVY, R.

C.E.A. - C.E.N./SACLAY - SES/SIR
B.P. N° 2
91190 GIF SUR YVETTE (FRANCE)

RESUME

La recherche lexicale utilise les résultats d'une analyse phonétique :

1 - Préclassement en voyelles, consonnes, fricatives et occlusives non voisées.

2 - Identification dans chaque classe.

L'apprentissage automatique cherche à passer de la chaîne lexicale (en classes phonétiques) à la chaîne de l'analyseur phonétique. Les règles de récriture produites tiennent compte des erreurs en incluant un contexte droit et gauche. Elles ne sont pas récursives. Toutes les règles sont acceptées et pondérées par leurs occurences.

L'analyse syntaxique peut produire un grand nombre de chaînes aussi on se limite à l'application maximum de trois règles successives. Parmi les chaînes produites on peut trouver plusieurs mots du lexique. Le choix est fait par une procédure utilisant les résultats de l'identification phonétique.

La grammaire reflète les caractéristiques de l'analyseur phonétique. On constate que le nombre de règles tend vers une limite compatible avec le temps réel quand le nombre de mots augmente. On peut alors introduire un nouveau mot dans le lexique sans apprentissage complémentaire.

L'application d'une règle peut permettre un retour arrière sur les hypothèses de l'analyseur phonétique.

LEXICAL RESEARCH IN SPEECH RECOGNITION. PROCESSING BY A PARSING METHOD BASED ON AUTOMATIC INFERENCE.

BAUDRY, M., DUPEYRAT, B., LEVY, R.
C.E.A. - C.E.N./SACLAY - SES/SIR
B.P. N° 2 - 91190 GIF SUR YVETTE (FRANCE)

SUMMARY

The lexical research uses the results of the phonetic analysis :

1 - Pre-classifying in vowels, consonants, unvoiced fricatives and bursts.

2 - Identification into each class.

The automatic learning works by moving from the lexical string (in phonetic classes) to the string of the phonetic analysis. The produced rewriting rules take into account the failures by including a right and left context. They are not recursive. All the rules are validated and weighted according to their occurence. The syntactic analyser can produce a large set of strings thus it has been restrained to a maximum of three successive rules. Among the produced strings, several lexical words can be found. Final choice is made through a process using the results of the phonetic identification.

Grammar reflects the characteristics of the phonetic analyser. It may be noticed that the number of rules tends to a limit which is compatible with real time processing as the volume of words increases. At this stage, a new word can be added to the lexic without any learning.

The application of a rule allow a back-track to the hypothesis of the phonetic analyser.

The above system can be applied to the recognition of any pattern given in unidimensional representation.

10èmes JOURNÉES D'ÉTUDE SUR LA PAROLE
GRENOBLE - 30 MAI - 1er JUIN 1979

RECHERCHE LEXICALE EN RECONNAISSANCE DE LA PAROLE. RESOLUTION PAR UNE METHODE D'ANALYSE SYNTAXIQUE ISSUE D'UNE INFERENCE GRAMMATICALE AUTOMATIQUE.

BAUDRY, M., DUPEYRAT, B., LEVY, R.
C.E.A. - C.E.N./SACLAY - SES/SIR
B.P. N° 2 - 91190 GIF SUR YVETTE (FRANCE)

I - INTRODUCTION

Nous présentons un système de reconnaissance de mots isolés. Seule la recherche lexicale sera traitée ici, l'analyse phonétique du signal de parole ayant déjà fait l'objet d'autres publications (BAUDRY, M., 1978 ; BAUDRY, M., DUPEYRAT, B., 1979).

Une idée directrice de l'étude a été de considérer les résultats de la segmentation suffisamment stables pour qu'un apprentissage de ses imperfections permette de les corriger, c'est-à-dire de les interpréter. Cet apprentissage consiste en l'inférence automatique de la grammaire ayant comme règles de récriture toutes les règles issues de la comparaison entre la segmentation du mot prononcé et la segmentation phonétiquement correcte de ce même mot. L'analyse syntaxique d'une segmentation proposée par le système d'analyse phonétique du signal consiste, par dérivations successives, à remonter à l'axiome "MOT LEXICAL". En cas d'ambiguité, c'est-à-dire quand l'analyse syntaxique d'une segmentation débouche de plusieurs façons possibles sur l'axiome, l'opération réalisée ne sera qu'une pré-classification, le choix définitif devant être effectué par les modules supérieurs : syntaxico-sémantique .... L'originalité de cette étude tient à plusieurs aspects dans la résolution du problème posé :

1°) Le temps réel est conservé malgré l'utilisation d'un mini-ordinateur assez lent.

2°) L'utilisation de techniques d'analyse syntaxique.

3°) La propriété d'insertion de mots nouveaux dans le lexique sans apprentissage complémentaire.

Ceci est possible dès que l'inférence grammaticale issue de l'apprentissage automatique devient caractéristique de l'analyseur phonétique. La grammaire devient alors indépendante du lexique choisi, reflétant précisément les possibilités discriminantes de l'analyseur phonétique.

Le système réalisé peut être utilisé pour d'autres applications pourvu que le codage des formes à reconnaitre soit unidimensionnel.

II - METHODE D'ANALYSE SYNTAXIQUE

1 - Données

Dans son état actuel le module d'analyse phonétique discrimine quatre classes différentes : les voyelles (VO) ; les consonnes (CO) ; les fricatives non voisées (FR) et les plosives non voisées (SI). Il propose dans chaque

classe trois phonèmes hiérarchisés. Nous ne nous servirons pas de cette information dans la mesure où elle n'a d'intérêt que si la segmentation phonétique est parfaite, ce qui n'est presque jamais le cas.

Exemple de données du module de segmentation.

Le mot "gɔlf " a été prononcé, puis segmenté de la façon suivante :

CO /d/g/b/ VO /o/ə/ɔ/ FR /s/f/ʃ/ VO /o/ɔ̃/ə/ CO /v/r/z/

2 - Approche méthodologique

Le premier traitement consiste à effectuer une comparaison entre la chaîne proposée à la segmentation et la chaîne lexicale d'orthographe phonétique correcte. Cette comparaison est en fait une recherche des symboles stables composant ces chaînes. Une chaîne est composée de symboles $\alpha_i \in \{VO, CO, FR, SI\}$. Dans une expérimentation particulière, six prononciations consécutives du mot lexical "bɛrtə" ont donné les segmentations suivantes :

1 -	CO VO CO VO SI VO CO	
2 -	CO VO SI VO CO	
3 -	CO VO SI VO CO	chaînes proposées par l'analyseur phonétique du signal vocal.
4 -	CO VO CO VO SI VO	
5 -	VO CO SI VO CO	
6 -	CO VO CO VO SI VO CO	
	CO VO CO SI VO	chaîne lexicale idéale.

Les causes ayant provoqué ces erreurs sont bien connues pour être des difficultés majeures de la segmentation :

1°) Présence du R entraînant soit sa non détection, soit une double détection aberrante.

2°) La diphonème consonne-consonne et à plus forte raison r-consonne ou consonne-r.

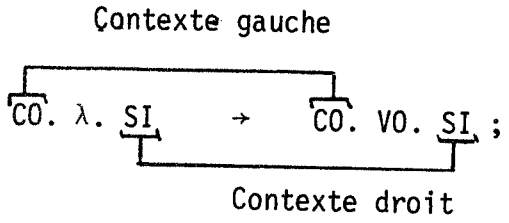
3°) La détection de fin de mot.

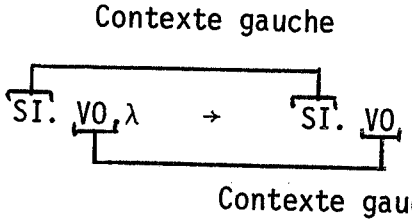
Malgré ces erreurs, trois phonèmes sont toujours détectés à certains endroits du mot : la première voyelle, la plosive, la dernière voyelle. Nous pouvons faire apparaître le squelette du mot "bɛrtə" :

$$P_1.VO.P_2.P_3.SI.VO.P_4$$

ou $P_1 \in \{\lambda, CO\}$; $P_2.P_3 \in \{\lambda, CO\ VO, CO\}$; $P_4 \in \{\lambda, CO\}$

λ est la chaîne vide. C'est autour de ces points d'ancrages d'un mot, que nous allons construire les productions rendant compte des contraintes de contextes des points d'ancrage. Ainsi, la segmentation de l'exemple n° 1 dérive de la chaîne lexicale correcte par l'application de deux règles de récriture :





La formalisation de la grammaire donne :

"S" est l'axiome MOT LEXICAL

V_n = {MOT LEXICAL, VY, CN, FC, OS} où VY, CN, FC et OS sont les notations non terminales de VO, CO, FR et SI

V_t = {VO, CO, FR, SI}

P est l'ensemble des productions de la grammaire que nous venons de définir et qui sont engendrées par l'apprentissage automatique.

Soit CO VO SI VO une chaîne de symboles $\alpha_i \in V_t$ proposée par le système de segmentation pour le mot "bɛrtə" prononcé. Cette chaîne dérive de la chaîne /CO/VO/CN/SI/VO/ par application de la règle :

/VO/CN/SI → /VO/SI/

Ainsi que nous l'avons indiqué plus haut, nous tenons compte des contextes droit et gauche de la règle :

$$C_1 \; A \; C_2 \rightarrow C_1 \; B \; C_2$$

ou $A \in V_n$; $B \in V^*$; C_1 et $C_2 \in V_n$

et B peut éventuellement être la chaîne vide λ. Les résultats donnés par l'analyseur phonétique montrent qu'un phonème déjà récrit ne le sera pas une deuxième fois, autrement dit il n'existe pas de productions récursives :
VO → VO CO

On impose VY → VO CO

Il s'agit d'une grammaire de type 0 avec l'utilisation d'un contexte droit et gauche.

Nous pourrons tirer des exemples, des six segmentations du mot "bɛrtə", les productions suivantes :

Exemple 1 : règle a) CN SI → CO VO SI
règle b) SI VY → SI VO CO

Exemple 2 : règle c) VO CN SI → VO SI
règle b) de nouveau

Exemple 3 : identique au 2

Exemple 4 : règle a) de nouveau

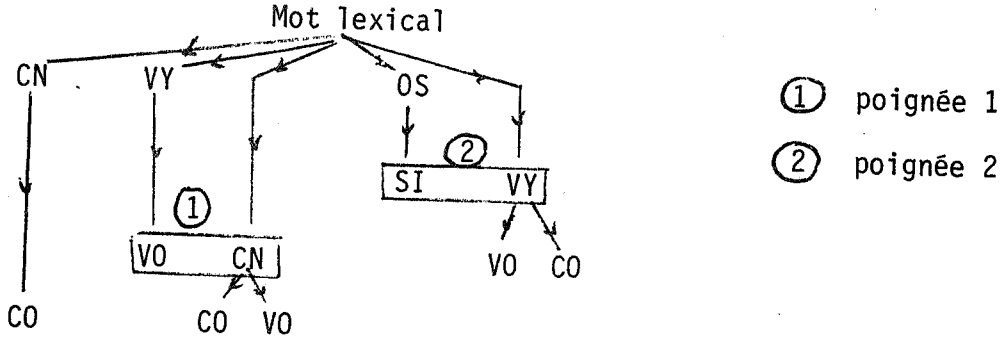
Exemple 5 : règle d) CN VO CO → VO CO

Exemple 6 : identique au 1

Dans cet ensemble d'apprentissage du mot "bɛrtə" P contient les quatre règles de récriture énoncées au-dessus ainsi que toutes les productions directes d'un symbole non terminal en symbole terminal :

règle e) VY → VO
règle f) CN → CO
règle g) FC → FR
règle h) OS → SI

Ainsi, l'arbre syntaxique de la segmentation n° 1 donnera le schéma suivant :



3 - L'analyse syntaxique

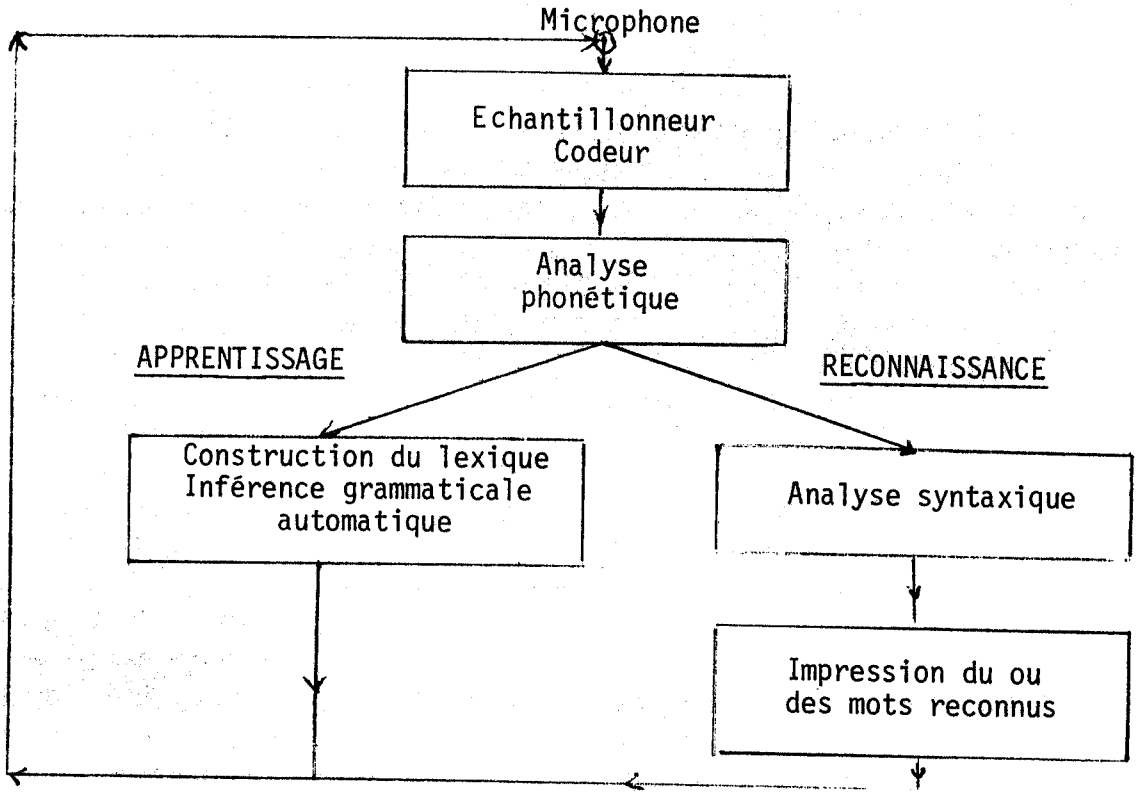
En fait, pour des raisons de programmation et de diminution de combinaisons possibles, l'analyse syntaxique s'effectuera de gauche à droite et de bas en haut.

L'analyse se fait à partir de la chaîne segmentée en recherchant les règles applicables d'abord en début de mot puis en fin de mot et enfin en milieu de mot. En effet, les erreurs de segmentation changent selon la position dans le mot, en dehors de tout contexte phonétique particulier. Ceci impose une utilisation sélective des règles. Nous obtenons ainsi trois types de productions, productions de début, fin et milieu de mot. De plus, une chaîne de segmentation, qui dérive d'une chaîne lexicale ayant plus de 2 symboles de plus ou de moins, oblige à abandonner la chaîne lexicale parce que peu probable. Toutes les règles engendrées par l'apprentissage automatique ne seront pas conservées. Seules celles qui ont une fréquence significative le seront, les autres ne sont représentatives que d'aberrations du codeur ou d'ambiance trop bruitée.

L'analyse syntaxique offre en fin de traitement zéro, un ou plusieurs mots lexicaux.Quand aucun mot lexical n'est trouvé c'est que la segmentation initiale n'est représentative d'aucun mot appris.

III - APPLICATION

1 - Organigramme fonctionnel du système



2 - Description modulaire

a) le module d'apprentissage

Deux sous-programmes composent ce module:i) le recalage de la chaîne segmentée sur la chaîne lexicale du mot prononcé.

Exemple : le mot prononcé est "viktɔr" la segmentation a donné CO VO SI VO CO. Nous établissons un tableau de recalage :

CO VO SI SI VO CO	chaîne lexicale
CO VO SI VO CO	chaîne de segmentation
1 2 3 Ø 4 5	tableau de recalage
1 2 3 4 5 6	

L'indice du tableau de recalage donne le numéro du symbole de la chaîne lexicale et le contenu le numéro du symbole de la chaîne de segmentation auquel il est recalé.

ii) deuxième sous-programme construction des règles de récriture à partir du tableau de recalage, et gestion du lexique pour introduction de mots nouveaux.

b) le module de reconnaissance

Deux sous-programmes le composent, l'analyseur syntaxique dont le principe a été déjà développé et l'identificateur des chaînes lexicales.

3 - Résultats

Les expériences ont été faites sur plusieurs vocabulaires différents :

a) l'alphabet téléphonique augmenté des chiffres : 36 mots

Sur 143 essais (environ quatre fois chaque mot) on a obtenu :

26 mauvaises reconnaissances : absence du mot prononcé parmi les candidats (18 %)
117 bonnes pré-classifications (82 %)
5 candidats en moyenne par pré-classification
160 règles de récritures.

b) l'alphabet radio : 26 mots

Sur 130 essais (cinq fois chaque mot) on a obtenu :

20 mauvaises reconnaissances (15,4 %)
110 bonnes pré-classifications (84,6 %)
3 candidats en moyenne
100 règles de récritures.

c) l'alphabet téléphonique radio et les chiffres : 60 mots

Sur 120 essais (deux fois chaque mot) on a obtenu :

21 mauvaises reconnaissances (17,5 %)
99 bonnes pré-classifications (82,5 %)
7 candidats en moyenne
160 règles de récritures.

IV - CONCLUSION

Nous avons présenté un système de pré-classification de mots. Sur un lexique de 60 mots l'analyse syntaxique délivre 5 candidats possibles. Il demeure encore un grand nombre d'échecs (18 % des cas) qui est cependant constamment réduit par l'amélioration du système d'analyse phonétique.

L'analyseur syntaxique présenté permet aussi de déceler les erreurs fréquentes de l'analyseur phonétique et aussi de remonter au son et éventuellement de recommencer l'analyse phonétique avec d'autres seuils. Il s'agit d'un véritable contrôle sur le son par le biais de l'analyseur syntaxique. Le nombre de règles développées à l'apprentissage est borné. Quand le lexique augmente l'ensemble des productions suit une courbe rapidement asymptotique. Dès lors, l'introduction de mots nouveaux se fait sans apprentissage complémentaire.

REFERENCES

BAUDRY, M., 1978, Etude du signal vocal dans sa représentation amplitude temps. Algorithmes de segmentation et de reconnaissance de la parole. - Thèse d'Etat, 15 juin 1978.

BAUDRY, M., DUPEYRAT, B., 1979, Speech segmentation and recognition using syntactic methods on the direct signal. - ICASSP 79, Washington 2-4 avril.

FU, K.S., 1974, Syntactic methods in pattern recognition. - Academic Press, New-York.

de MORI, R., 1977, Communication and cybernetics. - K.S. FU, Editor, Springer-Verlag, vol. 14.

10èmes JOURNÉES D'ÉTUDE SUR LA PAROLE

GRENOBLE - 30 MAI - 1ER JUIN 1979

RECONNAISSANCE DE MOTS QUASI-NATURELS ET PHONO-CODÉS

JEAN A. DREYFUS-GRAF

5, Avenue de la Grenade, CH-1207- GENEVE

RESUME

Les machines actuelles, qui cherchent à reconnaître la parole, modélisent surtout l'émetteur (exitation et conduit vocal) et ne s'occupent pas suffisamment du récepteur (oreille et cerveau). Or, l'émetteur fonctionne d'une manière linéaire et intégrale, en première approximation, tandis que le récepteur obéit essentiellement à des lois non-linéaires (logarithmiques ou de puissance) et différentielles. Le présent exposé essaye de formuler de telles lois pour les appliquer à la reconnaissance de phrases continues, puis à celle de mots quasi-naturels et phono-codés, ainsi qu'à leurs conversions. Les paramètres spectraux, logarithmiques et différentiels, caractérisent les actions et les pentes de formants et anti-formants, ainsi que les corrections par facteurs de modulation. Les algorithmes de reconnaissance utilisent les divers niveaux de contraintes linguistiques : spectral, phonétique, phonologique, lexical, sémantique et contextuel.

Le phono-décodeur SOKINA-ERUY reconnaît 6 phonèmes centraux, ch,ô,k,i,n,â, ou jusqu'à 10 phonèmes, avec ê,r,ou,u. Le phonocode SOTINA permet de formuler les nombres, de zéro à l'infini d'une manière logique, internationale et directement intelligible aux machines. Divers groupes de phonèmes sont rattachables aux phonèmes centraux, permettant la conversion automatique de mots quasi-naturels (présentant des avantages mnémoniques) en mots ou nombres phonocodés.

RECOGNITION OF QUASI-NATURAL AND PHONO-CODED WORDS

JEAN A. DREYFUS-GRAF

SUMMARY

Present speech regognizing systems mainly model the emitter (excitation and vocal tract), but neglect the receiver (ear and brain). The emitter functions act, however, in a linear and integral manner, as a first approximation, whilst the receiver follows fundamentally non-linear and differential laws. The present paper attempts to formulate such laws in order to develop speech-recognition machines for :
1. continuous sentences, 2. quasi natural words, 3. phono-coded numbers, 4. the conversion of quasi-natural into phono-coded words. Logarithmic and differential spectral parameters characterize the actions and slopes of formants and anti-formants as well as the correction factors for modulation degrees. The recognition algorithms utilize the various constraint levels : spectral, phonetic, phonological, lexical, syntactic, semantic and contextural.
The SOKINA-ERUY phono-decoder recognizes 6 central phonemes, /ʃ/, /o/, /k/, /i/, /n/, /a/, or up to 10 phonemes, with /ɛ/, /R/, /u/, /y/. With the SOTINA phono-decoder, the numbers from zero to infinity can be formulated in a manner which is logical, international and immediately understandable by machines. Various phoneme-groups are attachable to the central phonemes, allowing the automatic conversion of quasi-natural words (which present mnemonic advantages) in to phono-coded words or numbers. Tables 7 and 10 show some examples, in 2 or 3 lines : (1) quasi-natural words, (2) phono-coded words, (3) corresponding numbers.

TABLEAU 7 Exemples de mots quasi-naturels (1), avec les mots phonocodés correspondants (2), convertis par le phono-décodeur SOKINA-E (voir Tableau 6.b.)

(1)	à gauche	à droite	en haut	en bas	start	stop	in	off
(2)	AKOS	AKOAK	AO	AKA	SKAAK	SKOK	IN	OS
(1)	à l'ouest	à l'est	au nord	au sud	marche	halte	s.o.s	
(2)	A-OESK	A-ESK	ONOO	OSIK	NAAS	A-K	ESOES	
(1)	one	two	three	four	five	six	seven	eight	nine	zero
(2)	OAN	KO	S-I	SOO	SAIS	SIKS	SESEN	EIK	NAIN	SI-O

TABLEAU 10 Exemples de mots quasi-naturels (1), prononcés par l'homme, avec les mots phono-codés (2) et les nombres correspondants reconnus par la machine (3).

a. Chiffres unitaires combinables

(1)	Lolo	au-lit	auto	Odile	ôta	osons	aussi	hocha	honni	homard
(2)	O-O	O-I	OTO	OTI	OTA	OSO	OSI	-OSA	-ONI	-ONA
(3)	00	01	02	03	04	05	06	07	08	09

b. Dizaines combinables

(1)	colon	colis	coco	copie	coca	cochon	cosy	cocha	Toni	tonna
(2)	TO-O	TO-I	TOTO	TOTI	TOTA	TOSO	TOSI	TOSA	TONI	TONA
(3)	20	21	22	23	24	25	26	27	28	29

c. Centaines combinables

(1)	mikado	Nikita	Isaïe	capital	Tokyo	Kyoto	Dakota
(2)	NITATO	NITITA	ISAI	TATITA-	TOTIO	TIOTO	TATOTA
(3)	842	834	171	434	230	302	424

d. Mots de passe

(1)	Nabuchodonosor	inhibition	Nicosie	Panama	capacité
(2)	NATITOTONOSO-	INITISIO	NITOSI	TANANA	TATASITI
(3)	9 3 2 2 0 5	1 8 3 6 0	8 2 6	4 9 9	4 4 6 3

e. Numéros de téléphone (ou serrure)

	abonné: CB, Grenoble			JDG, Genève					
(1)	vas-y saute ami y passons			Toto qui c'est qui t'a dit tout					
(2)	SASI	SOTANI	ITASO	TOTO	TI	SI	TI	TA	TI TO
(3)	7 6	5 4 8	1 4 5	2 2	3	6	3	4	3 2

	abonné: Observatoire, Salève				Administr. Cant. Genève					
(1)	on conquit dimanche et on sauça				touche à tout hihihi					
(2)	OTOTI	TINA	SIO	SOSA	TO	SA	TO	I	I	I
(3)	0 2 3	3 9	6 0	5 7	2	7	2	1	1	1

f. Constantes universelles

e = base des logarithmes naturels

											G= gravitation					
(1)	ton	chat	inique	au	nid	inique	au	nid			s'y	sécha	mouillé			
(2)	TO	SA	I	NI	TO	NI	I	NI	TO	NI	SI	SI	SA	NO	I	I
(3)	2 ,	7	1	8	2	8	1	8	2	8	6 ,	6	7	.10^{-11}		

10èmes JOURNÉES D'ÉTUDE SUR LA PAROLE
GRENOBLE - 30 MAI - 1ER JUIN 1979

RECONNAISSANCE DE MOTS QUASI-NATURELS ET PHONO-CODÉS

JEAN A. DREYFUS-GRAF

INTRODUCTION

En dépit des moyens importants qui ont été mis en oeuvre (par exemple par Carnegie Mellon University [1] ou I.B.M. [2]),il n'existe jusqu'à présent, sur notre planète, aucune machine capable de reconnaître la parole humaine, indépendamment de la voix du locuteur et d'un nombre restreint de mots pré-enregistrés, qui sont soumis à des contraintes artificielles.

Pourtant l'oreille humaine reconnaît aisément des millions de phrases et de mots, articulés par des milliers de voix différentes, avec des accents et des degrés de voisement variables, pouvant aller d'une voix de stentor au chuchotement. Il semble donc que les machines actuelles, qui modélisent surtout l'émetteur de parole (excitation et conduit vocal) ne s'occupent pas suffisamment du récepteur de parole (oreille et cerveau). Or, l'émetteur fonctionne d'une manière linéaire et intégrale, en première approximation, tandis que le récepteur obéit à des lois essentiellement non-linéaires (logarithmiques ou de puissance [3]) et différentielles, en première approximation. On peut supposer que les valeurs des analyses spectrales sont contrôlées par de multiples boucles cybernétiques, puisqu'elles sont largement indépendantes de la dynamique, de la hauteur de voix, de l'accent et du degré de voisement. On peut aussi admettre que les décisions logiques sont basées sur des comparaisons différentielles simples, telles que : par paires de valeurs, a, b, avec seulement 3 réponses admises, "a plus grand, égal ou plus petit que b". Ceci signifie 3 niveaux +1,0,-1, ou 0,1,2 = 1 tit = log.dual.3 = 1,585 bit. Nous allons essayer de formuler de telles lois non-linéaires pour les appliquer à la reconnaissance de phrases continues, puis à celle de mots quasi-naturels et phono-codés [4], ainsi qu'à leurs inter-conversions.

2. PARAMETRES SPECTRAUX, LOGARITHMIQUES ET DIFFERENTIELS

Selon la partie a) du Tableau 1, l'évolution caractéristique d'un phonème, dans la bande n, est représentée par la différence Lng = Ln-Lg entre le logarithme Ln de son énergie et le logarithme Lg d'un niveau de référence global. L'évolution présente une pente initiale Pin= dLng/dt1, une action quasi-stationnaire Ang= Lng.t2 (positive ou négative, selon qu'il s'agit d'un formant ou d'un anti-formant), et une pente finale Pfn= dlng/dt3, dont les durées respectives sont t1, t2, t3. La discrimination est ternaire, c'est-à-dire qu'elle ne distingue que 3 niveaux différentiels, -1,0,+1, auxquels on associe les 3 niveaux 0,1,2 d'une unité ternaire (ternary digit), = 1 tit = 1,585 bit. Au lieu de la différence Ln-Lg on pourrait utiliser les différences Ln-L(n-1), Ln-L(n-2) ..., pourvu que, d'une manière ou d'une autre, les valeurs absolues soient remplacées par des différences, présentant 3 niveaux caractéristiques.

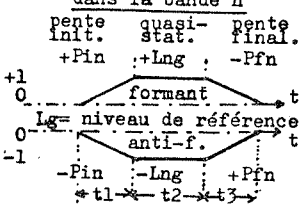
Selon la partie b), il s'agit de choisir le nombre N minimum de bandes de fréquences F1 à Fn, compatibles avec les formants et anti-formants des phonèmes à discriminer. On en montre 2 exemples : N= 14 pour les fréquences acoustiques (90-6000 Hz), et N= 11 pour les fréquences téléphoniques (250-3250 Hz).

Selon les parties c), d), e), on retient 2 tit = 3,17 bit pour les énergies logarithmiques globales, 22,2 bit (ou 17,43) pour les différences formantiques de l'exemple 1 (ou 2), et 12,34 bit pour les degrés de modulation [5] voisement, hauteur, mélodie, friction, roulement [6], qui en effectuent les corrections.

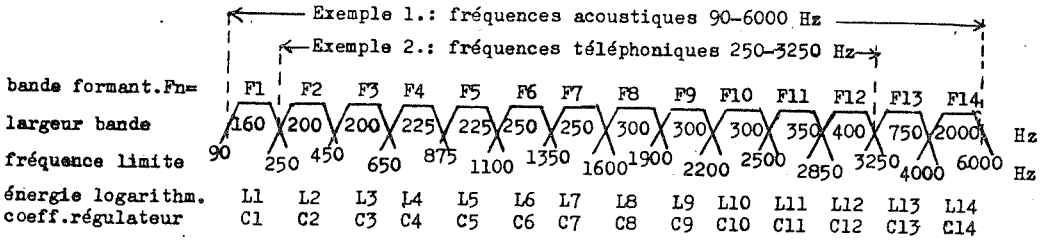
TABLEAU 1 PARAMETRES SPECTRAUX, LOGARITHMIQUES ET DIFFERENTIELS

n = numéro de bande de formant ou anti-formant Fn
Ln (dB) = log. de l'énergie E (J) dans la bande n
Ln.Cn = Ln régulé par le coefficient égalisateur Cn
Lg (dB) = log. global des énergies régulées = $\sum$ Ln.Cn / n = réf.
Cg = coefficient global de régulation = $\sum$ Cn / n
Pg (W) = dLg/dt = pente de Lg (initiale=pos., finale=nég.)
Lng = Ln-Lg (dB) = différence d'énergie log. (niveau de réf. Lg)
Pin (W) = dLng/dt1 = pente initiale (pos. formant, nég. anti-form.)
Pfn (W) = dLng/dt3 = pente finale (nég. formant, pos. anti-formant)
(Js) = action quasi-stationnaire de Lng = Lng.t2

a) Evolution d'un phonème dans la bande n



b) Bandes de formants et anti-formants



c) Energies logarithmiques globales G (3 niveaux 0 1 2 = 1 tit)

paramètre		symb.	tit
G1	quasi-station.régulée(réf.)	Lg	1
G2	dynamique réglante (coeff.)	Cg	1
G3	pentes	Pg	(1)
G4	silences(internes, ponctuat. externes)	Sg	(1)
		total	2

e) Degrés de modulation M (ternaires)

paramètre	(3 niveaux 0 1 2 = 1 tit)	symb.	tit	bit
M1	voisement (80-370 Hz)	Mv	1=	1,585
M2	hauteur(hom.0,fem.1,enf.2)	Mh	1=	1,585
M3	mélodie (64 quarts-tons)	Mm		6
M4	friction (subform.60 Hz)	Mf	1	1,585
M5	roulement(" 30 Hz)	Mr	1	1,585
		total		12,34

d) Différences formantiques D (ternaires) Lng = Ln - Lg
(3 niveaux 0 1 2 = -1 0 +1 = 1 tit)

paramètre		symb.	tit
D1	L1-Lg	L1g	1
D2	L2-Lg	L2g	1
D3	L3-Lg	L3g	1
D4	L4-Lg	L4g	1
D5	L5-Lg	L5g	1/5

pentes Pin = Pi1-Pi14 (D15-D28) (14)

paramètre		symb.	tit
D6	L6-Lg	L6g	1
D7	L7-Lg	L7g	1
D8	L8-Lg	L8g	1
D9	L9-Lg	L9g	1
D10	L10-Lg	L10g	1/5

pentes Pfn = Pf1-Pf14 (D29-D42) (14)

paramètre		symb.	tit
D11	L11-Lg	L11g	1
D12	L12-Lg	L12g	1
D13	L13-Lg	L13g	1
D14	L14-Lg	L14g	1/4
		report	10
		total	14

Exemple 1: W = 22,2 bit
Exemple 2: II = 17,43 "

3. DEBITS D'INFORMATION DES PARAMETRES SPECTRAUX

Selon le Tableau 2., et en admettant une fréquence d'analyse Fa= 100 Hz (ou une fenêtre de 10 ms), le débit d'information des paramètres spectraux devient : 3771 bit/sec pour l'exemple 1. (avec N= 14 bandes), resp. 3294 bit/sec pour l'exemple 2. (avec N= 11 bandes). Ces valeurs sont du même ordre que celles des synthétiseurs de parole (vocoders) modélisant l'émetteur de parole (1200 à 9600 bit/sec), mais leurs interprétations sont différentes puisqu'elles correspondent à la modélisation du récepteur de parole. On déterminera expérimentalement la fréquence d'analyse Fa optimum, qui doit se situer entre 100 et 200 Hz.

TABLEAU 2 — Débit d'information (bit/sec) des paramètres spectraux (fréquence d'analyse Fa = 100 Hz)

	Exemple 1.:"direct"90-6000 Hz fréqu.échantillonn.:Fe=12000Hz log.dual. 4096 niveaux= 12 bit débit: 12000x12= 144000 bit/sec				Exemple 2.:"téléph" 250-3250 Hz fréqu.échantillonn.Fe= 6000 Hz log.dual. 256 niveaux= 8 bit débit: 6000x8= 48000 bit/sec			
a)+b) nombre bandes fréq.	N = 14				N = 11			
	paramètre No	unité d' informat. bit	Fa Hz	débit d' informat. bit/sec	paramètre No	unité d' informat. bit	Fa Hz	débit d' informat. bit/sec
c) énergies log.globales d) degrés de modulation e) différences formant.	G1-G4 M1-M5 D1-D14 (D15-D42)	3,17 12,34 22,2	100 100 100 total	317 1234 2220 3771	G1-G4 M1-M5 D1-D11 (D12-D33)	3,17 12,34 17,43	100 100 100 total	317 1234 1743 3294

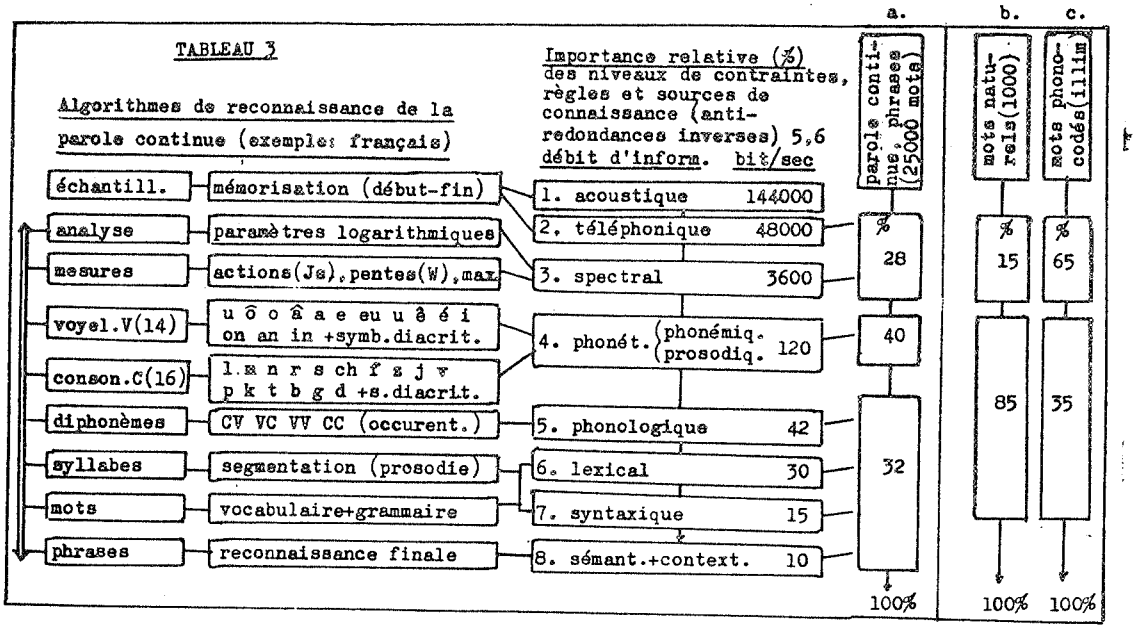
4. ALGORITHME DE RECONNAISSANCE DE LA PAROLE CONTINUE

Les paramètres spectraux décrits peuvent être obtenus par des technologies analogiques, digitales ou mixtes. A partir de ces paramètres, on peut développer divers algorithmes concernant la parole continue (phrases), découpée (mots) ou phono-codée (phonèmes simplifiés).

Le Tableau 3. schématise un exemple d'algorithme concernant la parole continue, dans le cas de la langue française, comprenant quelque 25000 mots usuels. Il pourrait s'appliquer, entre autres, à un nouveau "Phonétographe 1979". Cet algorithme utilise les divers niveaux de contraintes, règles et sources de connaissance linguistiques, tels que : 1. acoustique, 2. téléphonique, 3. spectrale, 4. phonétique (phonémique+prosodique), 5. phonologique, 6. lexical, 7. syntaxique, 8. sémantique et contextuel. A partir du niveau 3. (paramètres spectraux), il recherche d'abord les actions Ang (Js) et les pentes (ou puissances) Pin, Pfn (W) maximales, permettant d'identifier les voyelles et consonnes dominantes (parmi les 14 voyelles et 16 consonnes admises pour la langue française). Ensuite, il identifie progressivement les autres phonèmes, les diphonèmes, les syllabes et mots répertoriés pour aboutir à la phrase cherchée, ceci en parcourant la boucle des niveaux 3. à 8. autant de fois que nécessaire pour lever les ambiguités.

5. IMPORTANCES RELATIVES DES NIVEAUX LINGUISTIQUES

Peut-on évaluer d'emblée l'importance relative des diverses contraintes linguistiques) Une première approximation en est fournie par "l'analyse des anti-redondances inverses" [7]. En effet, selon le Tableau 3, partie droite, le débit d'information diminue progressivement de 144000 bit/sec à 10 bit/sec quand on descend l'escalier des niveaux 1. à 8. La Colonne a. indique que les niveaux 1. à 4. totalisent 68%, n'accordant que 32% aux niveaux supérieurs.



6. RECONNAISSANCE DE LA PAROLE PAR L'HOMME ET PAR LA MACHINE

Le Tableau 4 montre que ces proportions correspondent environ à la reconnaissance de la parole par l'homme, mais non à celle par les machines mentionnées au début de l'exposé. En effet, en se basant sur l'intelligibilité humaine aux "syllabes dépourvues de sens" ("phonatomes"), qui est de 98% en direct, l'homme reconnaît les phonèmes avec un taux d'erreur de 0,7% [8][9], tandis que les machines utilisées, par exemple, par Carnegie Mellon University [1] ou IBM [10] présentent des taux d'erreur de 46%, resp. 36%, au niveau des phonèmes. Quant au taux d'erreur de 0,1% revendiqué par IBM [11], au niveau des mots (1011), il est illusoire car il correspond en fait à la reconnaissance d'une phrase artificielle parmi 100. et non à celle d'un mot naturel parmi 1011. En effet, dès que les mots se rapprochent de conditions naturelles, leurs taux d'erreur remontent à 33,1% [2], ceci bien que le nombre de mots ait été réduit à 486, et celui des phrases à 20.

TABLEAU 4 — TAUX D'ERREUR (%) dans la RECONNAISSANCE DE LA PAROLE

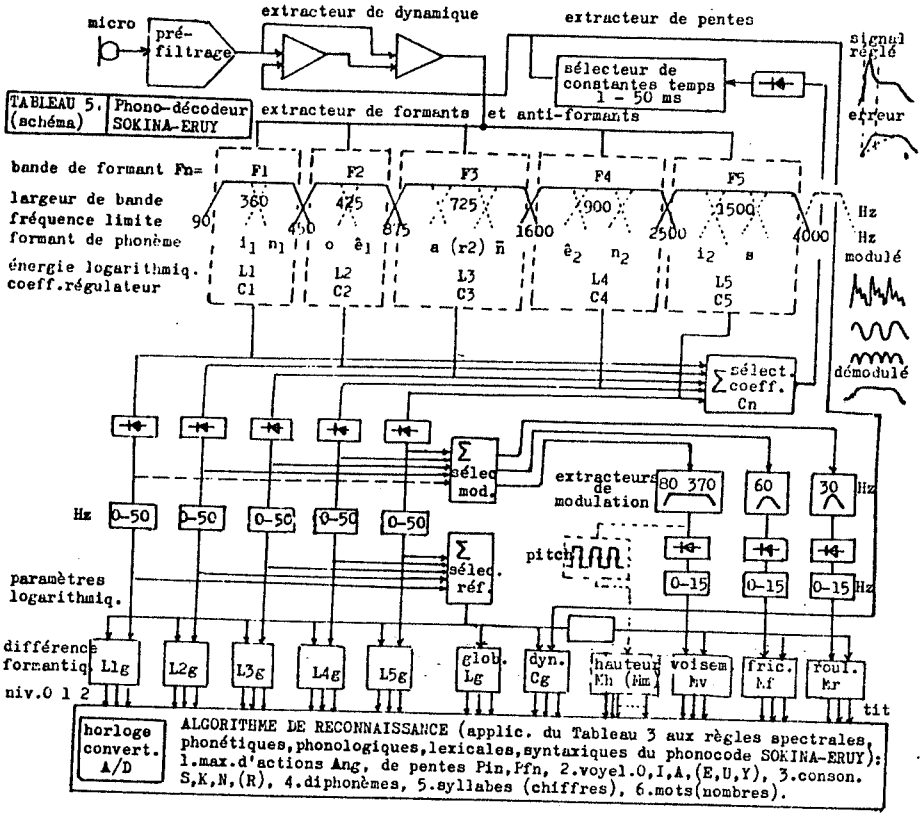
		PAR L'HOMME		PAR LA MACHINE				
	laboratoire	C.N.E.T.		Carn.Mellon U.	I.B.M.	L.B.M.	I.B.M.
	référence	[8] [9]		[1]	[10]	[11]	[2]
bande de fréquence : direct 100-6000 Hz téléphoniq. 250-3250 Hz	éléments à reconnaître	syllabes dépourvues de sens (illimité)		30 phrases par locuteur, 7/ 1011 mots	363 phrases 8/ 250 mots	100 phrases 7/ 1011 mots	20 phrases 486 mots
	langage	phonatomes		grammaire artificiel.AIX051	artific. N.Raleigh	artificiel AIX051	semi-naturel
	mio instr/s	-		17 - 30	30	-	-
	mémoire ou ordinateur	-		256 K . 36 bit	360/91 370/160	-	-
	temps réel	1 x		10 - 50 x	25 x	-	-
niveau de contrainte, règle ou source de connaissance	élément reconnu	direct %	téléph. %	direct %	téléph. %	direct %	direct %	direct %
1-3. acoust.+téléph. +spectral	-	-	-					
4. phonétique	phonèmes	(0,7)	(4,2)	46	-	38	-	-
5. phonologique	syllabes	2	12	-	-	11	-	-
6. lexical.	mots	(6)	(30)	3	-	3	0,1	33,1
7-9. syntax.+sémant. +contextuel	phrases	(0)	(2)	5	11	19	1	(100)

7. PHONO-DECODEUR SOKINA-ERUY

Le Tableau 5 montre une application des lois non-linéaires et différentielles d'un récepteur de parole à un phono-décodeur SOKINA-ERUY, c'est-à-dire à une machine reconnaissant les 6 classes de phonèmes internationaux O(=o), I(=i), A(=a), S(=ch,s), K(=k,t), N(=n,m), mais qui est extensible à 4 classes supplémentaires, E(=ê), U(=ou), Y(=u), R(= r roulé). Un phonocode, tel que SOTINA [12] permet de construire des langages logiques, applicables principalement aux nombres, de zéro à l'infini, et qui permettent de commander les machines, indépendamment de la voix du locuteur et du nombre de mots. Jusqu'à présent 2 réalisations préliminaires seulement ont été expérimentées : un prototype "CHARLES" [13] pour SOKINA et un modèle [14] pour SOFENA.

L'application des lois non-linéaires et différentielles peut être réalisée en technologie analogique, digitale ou hybride. C'est sous cette dernière forme qu'est présenté le schéma du Tableau 5. Celui-ci comprend un logarithmiseur-séparateur de dynamique [15] un extracteur de formants et anti-formants (avec N= 5 bandes, F1 à F5), trois sélecteurs (coefficients Cn, constantes de temps, niveau de référence), et 4 extracteurs de modulation (hauteur ou pitch, voisement, friction, roulement). Utilisant des comparateurs à 3 niveaux. 0,1,2. on obtient 11 paramètres ternaires, logarithmiques et différentiels, c'est-à-dire : 2 paramètres globaux Lg, Cg, 5 paramètres formantiques L1g à L5g, et 4 paramètres correcteurs (de modulation) Mh, Mv, Mf, Mr.

Ces paramètres constituent le niveau 3. d'un algorithme de reconnaissance similaire à celui du Tableau 3, mais que se borne à identifier les 6 ou 10 phonèmes du phonocode SOKINA-ERUY.



8. MATRICE PHONETIQUE

Le Tableau 6 montre la matrice phonétique SOKINA-ERUY, qui est issue des 11 paramètres spectraux décrits par le Tableau 5. Chaque point de la matrice représente l'une des valeurs 0,1,2, qui doit être déterminée statistiquement. La matrice comprend 11 x 10 = 110 points d'actions A (Js) quasi-stationnaires, et 2x6 x 10 = 120 points de pente P (W) transitoires.

En admettant une fréquence d'analyse Fa = 100 Hz, ou une fenêtre de 10 ms, le débit d'information de la matrice est 11 tit = 17.4 bit x 100 Hz = 1740 bit/sec, c'est-à-dire environ la moitié de celle qui correspondait à la parole naturelle, selon le Tableau 3. On déterminera expérimentalement la fréquence d'analyse Fa optimum, qui doit se situer entre 100 et 200 Hz.

Selon la rangée a., du Tableau 6, le phono-décodeur SOKINA-ERUY reconnait les 10 phonèmes centraux (soulignés), ch,ô,k,i,n,â,ê, r_1, ou,u. mais il peut y rattacher, par exemple, d'autres phonèmes. tels que : s,f,- o,on,- t,p,- m,- a,an,- é,- eu, qui facilitent l'établissement de mots quasi-naturels correspondant à des mots phono-codés.

TABLEAU 6 Matrice phonétique SOKINA-ERUY — . = niveau 0,1,2

débit d'imformation: 11 tit= 17,4 bit x 100 Hz = 1740 bit/sec

symbole internat.	S	O	K	I	N	A	E	R	U	Y
val.phonét.centr.	ch	ô	k	i	n	â	ê	r	ou	u
énergie glob.										
quasi-station. Lg	.	.	.	.	.	.	.	.	.	.
dynamique Cg	.	.	.	.	.	.	.	.	.	.
pente ΔPg	.	.	.	.	.	.	.	.	.	.
modulations										
voisement Mv	.	.	.	.	.	.	.	.	.	.
hauteur Mh	.	.	.	.	.	.	.	.	.	.
(mélodie) (Mm)										
friction Mf	.	.	.	.	.	.	.	.	.	.
roulement Mr	.	.	.	.	.	.	.	.	.	.
action qu.st.										
90-450 Lg1	.	.	.	.	.	.	.	.	.	.
450-875 Lg2	.	.	.	.	.	.	.	.	.	.
875-1600 Lg3	.	.	.	.	.	.	.	.	.	.
1600-2500 Lg4	.	.	.	.	.	.	.	.	.	.
2500-4000 Lg5	.	.	.	.	.	.	.	.	.	.
pentes										
90-450 Pi1 Pf1	.	.	.	.	.	.	.	.	.	.
450-875 Pi2 Pf2	.	.	.	.	.	.	.	.	.	.
875-1600 Pi3 Pf3	.	.	.	.	.	.	.	.	.	.
1600-2500 Pi4 Pf4	.	.	.	.	.	.	.	.	.	.
2500-4000 Pi5 Pf5	.	.	.	.	.	.	.	.	.	.

phonèmes centraux et rattachables au phono-décodeur										
a. SOKINA-ERUY	ch s f	ô o on	k t p	i	n m	â a an	ê é	r_1	ou	u eu
b. SOKINA-E	ch s f	ô ou o on	k t p	i u	n m	â a an	ê é			
c. SOKINA	ch s f	ô ou o on	k t p	i u é ê	n m	â a an				
d.phonèm.français convertibles en phonocode SOKINA par machine non spécialisée	ch j s z f v (r_3)	ou ô o e on (r_1)	k g t d p b	i u é ê eu	n m (1) (r_4)	à a an in (r_2)				

TABLEAU 7 Exemples de mots quasi-naturels (1), avec les mots phonocodés correspondants (2), convertis par le phono-décodeur SOKINA-E (voir Tableau 6.b.)

(1)	à gauche	à droite	en haut	en bas	start	stop	in	off
(2)	AKOS	AKOAK	AO	AKA	SKAAK	SKOK	IN	OS
(1)	à l'ouest	à l'est	au nord	au sud	marche	halte	s.o.s	
(2)	A-OESK	A-ESK	ONOO	OSIK	MAAS	A-K	ESOES	
(1)	one	two	three	four	five	six	seven	eight	nine	zero
(2)	OAN	KO	S-I	SOO	SAIS	SIKS	SESEN	EIK	NAIN	SÔ-O

TABLEAU 8 PHONOCODE NUMERIQUE SOTINA (ou SOKINA): règles

- chiffre = voyelle O,I, ou voyelle O,I,A + consonne S,T,N
- A (initial ou après O,I) = "avec" (évite répétitions)
- segmentation: fin de chiffre = fin de voyelle
  fin de nombre =fin de voyelle+pause>300ms
- NO (initial ou final) = puissance posit. ou négat. de 10
- virgule = pause 200 - 300 ms.(ou mot à double-consonne)

0	1	2	3	4	5	6	7	8	9	10	11	12	13...	20	21	22
0	I	TO	TI	TA	SO	SI	SA	NI	NA	IO	II	ITO	ITI...	TOO	TOI	TOTO

23...	30	31	32...	1.000.000.000 = 10^9	10^{-9}	10^{-12}
TOTI..	TIO	TII	TITO.	LANAO=1avec9zéros=	NANO	NOITO

TABLEAU 9 REGLES DE CONVERSION (FRANCAIS-SOKINA) DE MOTS QUASI-NATURELS EN NOMBRES PHONO-CODES

1. Chacune des voyelles présentes (14) est répartie parmi l'une des 3 classes O,A,I (voir Tableau 6)
2. La voyelle A,initiale ou après une autre voyelle,est supprimée
3. Nombre de voyelles retenues par mot = nombre de chiffres par nombre (contrôle du nombre de chiffres prescrit)
4. Chacune des consonnes présentes (16) est répartie parmi l'une des 3 classes S,K,N (voir Tableau 6)
5. Chaque consonne finale (du mot),ou précédée par une autre consonne, est supprimée (consonnes multiples non admises)
6. Chaque voyelle (O,I) ou consonne (S,K,N) + voyelle (O,A,I) retenue,est convertie en son chiffre SOTINA correspondant.
7. (faculatif) les consonnes instables,l,r, sont supprimées.

9. CONVERSION DE MOTS QUASI-NATURELS PAR LE PHONO-DECODEUR SOKINA-E

Le Tableau 7. montre, à titre d'exemples, quelques uns des mots quasi-naturels (1), tels que des ordres directionnels français ou des chiffres anglais, ainsi que les mots phono-codés correspondants (2), convertis par le phono-décodeur SOKINA-E. Celui-ci reconnaît, selon le Tableau 6, rangée b., les 7 phonèmes centraux (soulignés), ch,ô,k,i,n,â,ê, et il peut y rattacher 12 phonèmes supplémentaires, s,f,- ou,o,on,- t,p,- u,- m,- a,an,- é.

10. CONVERSION DE MOTS QUASI-NATURELS EN NOMBRES PHONO-CODÉS (SOTINA ou SOKINA)

Le Tableau 8 rappelle les principales règles de formation du phonocode numérique SOTINA ou SOKINA. (S et K sont interchangeable, mais K est plus favorable aux transmissions téléphoniques, tandis que T se prête mieux à la mémorisation des chiffres 0 à 9). Ce phonocode permet de formuler les nombres de zéro à l'infini d'une manière logique, internationale et directement intelligible aux machines.

Selon la rangée c. du Tableau 6, le phono-décodeur SOKINA reconnaît les 6 phonèmes centraux (soulignés), ch,ô,k,i,n,â, et il peut y rattacher 13 phonèmes supplémentaires, s,f,- o,ou,on,-t,p,-, u,é,ê,- m,- a,an.

Selon la rangée d., une machine non-spécialisée (reconnaissant la parole continue) peut rattacher aux 6 classes SOTINA tous les phonèmes d'une langue naturelle (par exemple, 14 voyelles et 16 consonnes pour le français).

Le Tableau 9 montre les règles de conversion de mots quasi-naturels (par exemple français) en nombres phono-codés (SOTINA). Le remplacement de nombres phono-codés par des mots quasi-naturels présente des avantages mnémoniques. En effet, si la musique des nombres phonocodés est plus facile à mémoriser que celle des nombres chiffrés usuels, elle est cependant moins facile à retenir que des mots quasi-naturels, rattachables aux repères linguistiques des associations d'idées. La valeur mnémonique et "parlante" des mots est d'autant plus grande que ceux-ci sont soumis à des contraintes linguistiques plus importantes, telles que : syllabiques, lexicales, syntaxiques, sémantiques et contextuelles [16].

Le Tableau 10 montre quelques exemples de conversion de mots quasi-naturels (1) en nombre phono-codés (2), puis en nombres chiffrés usuels (3). On considère des unités, dizaines ou centaines combinables, puis des mots de passe, des numéros de téléphone ou de serrure, et des constantes universelles. Il est évident que le numéro de téléphone de Grenoble "76 548 145" est plus difficile à mémoriser que la phrase "vas-y saute ami y passons". De même le nombre "2,71287128" (base "e" des logarithmes naturels) est plus difficile à retenir que la phrase "ton chat inique au nid inique au nid".

Les règles de conversion du Tableau 9. doivent être adaptées spécifiquement aux divers groupes de phonèmes rattachables, indiqués par les rangées a.-d. du Tableau 6 , selon qu'il s'agira d'un phono-codeur SOKINA-ERUY, SOKINA-E, ou SOKINA, ou bien d'une machine complète. Le phonétographe III (1960), qui continue à être cité comme pionnier [17] [18], en attendant de devenir une pièce de musée, pourrait dès lors renaître dans une nouvelle génération de machines, comprenant la parole naturelle.

TABLEAU 10 Exemples de mots quasi-naturels (1), prononcés par l'homme, avec les mots phono-codés (2) et les nombres correspondants reconnus par la machine (3).

a. Chiffres unitaires combinables

(1)	Lolo	au-lit	auto	Odile	ôta	osons	aussi	hocha	honni	homard
(2)	0-0	0-I	OTO	OTI	OTA	OSO	OSI	-OSA	-ONI	-ONA
(3)	00	01	02	03	04	05	06	07	08	09

b. Dizaines combinables

(1)	colon	colis	coco	copie	coca	cochon	cosy	cocha	Toni	tonna
(2)	TO-0	TO-I	TOTO	TOTI	TOTA	TOSO	TOSI	TOSA	TONI	TONA
(3)	20	21	22	23	24	25	26	27	28	29

c.Centaines combinables

(1)	mikado	Nikita	Isaïe	capital	Tokyo	Kyoto	Dakota
(2)	NITATO	NITITA	ISAI	TATITA-	TOTIO	TIOTO	TATOTA
(3)	842	834	171	434	230	302	424

d. Mots de passe

(1)	Nabuchodonosor	inhibition	Nicosie	Panama	capacité
(2)	NATITOTONOSO-	INITISIO	NITOSI	TANANA	TATASITI
(3)	9 3 2 2 0 5	1 8 3 6 0	8 2 6	4 9 9	4 4 6 3

e. Numéros de téléphone (ou serrure)

	abonné: CE, Grenoble			JDC, Genève					
(1)	vas-y	saute ami	y passons	Toto	qui c'est	qui t'a	dit	tout	
(2)	SASI	SOTANI	ITASO	TOTO	TI SI	TI TA	TI TO		
(3)	7 6	5 4 8	1 4 5	2 2	3 6	3 4	3 2		

	abonné: Observatoire, Salève			Administr.Cant.Genève			
(1)	on conquit	dimanche	et on sauça	touche à	tout	hihihi	
(2)	OTOTI	TINA	SIO SOSA	TO	SA TO	I I I	
(3)	0 2 3	3 9	6 0 5 7	2	7 2	1 1 1	

f. Constantes universelles

	e = base des logarithmes naturels									G= gravitation					
(1)	ton	chat	inique	au	nid	inique	au	nid		s'y	sécha	mouillé			
(2)	TO	SA	I NI	TO	NI	I	NI	TO	NI	SI	SI	SA	NO	I	I
(3)	2 ,	7	1 8	2	8	1	8	2	8	6 ,	6	7	.10		-11

REFERENCES

[1] LOWERRE, REDDY, Speech Understanding Systems, The Harpy System. Summary of Results, Carnegy-Mellon University, August 1977.

[2] BAHL, BAKER, COHEN, JELINEK, LEWIS, MERCER, (I.B.M.) Recognition of continuously read natural corpus, IEEE-ICASSP, Tulsa 1978, pp.422-424

[3] DREYFUS-GRAF, Cybernétique auditive, Revue d'Accoustique, No 14, 1971

" " L'oreille comme boîte grise. Conférence, Paris, avril 1970.

[4] " " Parole codée (phonocode), Revue d'Accoustique, No 21, 1972

[5] " " Reconnaissance de parole et segmentation, 7èmes JEP, Nancy 1976

[6] " " Phonétographe et sub-formants, Bull.Techn. PTT,Bern, No 2, 1957

" " Phonétographe : Présent et futur, " " " No 5, 1961

[7] " " Analyses des redondances de systèmes symboliques et degrés de voisement, 8èmes JEP, Aix-en-Provence, 1977

[8] DREYFUS-GRAF, Caractéristiques comparatives de systèmes reconnaissant la parole, 9èmes JEP, CNET-GALF, Lannion 1978

[9] LORAND, L'influence d'un bruit blanc sur la qualité d'une communication téléphonique, CNET, Lannion, Note Technique ETA/1, 1969

[10] JELINEK, Continuous Speech Recognition by Statistical Methods, Proc.. IEEE, April 1976, pp.532-556

[11] BAHL, BAKER, COHEN, COLE, JELINEK, LEWIS, MERCER, Automatic recognition of continuously spoken sentences from a finite state grammar,IEEE - ICASSP, Tulsa, 1978, pp.418-421

[12] DREYFUS-GRAF, Codes phonétiques (phonocodes) et télécommunications, 1er Congrès Européen d'Accoustique, FASE 75, Paris, 1975

[13] COURBON, CARTIER, LORAND, Reconnaissance automatique de la parole, "Charles", Revue d'Accoustique, Vol.10, No 43, 1977, pp 440-446

[14] VIGNERON, Reconnaissance des échantillons de parole continue, en temps réel, et par microprocesseur, Thèse, Université de Nancy, 1978

[15] DREYFUS-GRAF, Recognition of Natural and of Artificial Speech, (Phonocodes), IEEE-ICSCP, Newton-Boston, 1972, pp. 306-310

[16] DREYFUS-GRAF, Les chiffres peuvent-ils parler ? Table périodique des éléments. en phonocode SOTINA. Revue OSAEC, No 5, Lausanne, février 1979,pp.4-5

[17] FLANAGAN, Computers that Talk and Listen : Man-Machine Communication by Voice, Proc.IEEE, April 1976, pp.411,415 "speech typewriter" [35]

[18] HATON, LIENARD, La reconnaissance de la parole, La Recherche, No 99, Vol. 10, pp 329-332, Paris, avril 1979.

10èmes JOURNÉES D'ÉTUDE SUR LA PAROLE

GRENOBLE - 30 MAI - 1ER JUIN 1979

LES LEXIQUES AUTOMATIQUES POUR LE TRAITEMENT DE LA PAROLE : LA VERSION III DU LEXIQUE DU PROJET A.R.I.A.L*

G. GOUARDERES

Laboratoire CERFIA** - Université Paul Sabatier TOULOUSE

RESUME

Nous attribuons aux lexiques deux rôles principaux :

- le premier consiste à associer à chaque mot (ou unité lexicale) une classe grammaticale, un ou plusieurs sens et différents emplois,

- le second permet de simplifier la grammaire (et les règles) en catalogant de façon exhaustive les exceptions et les ambiguités. Depuis 1976, nous avons développé, pour divers modules du projet ARIAL* une série de versions spécifiques de lexiques, dérivés du lexique de base, lui même en constante évolution depuis cette date. Nous proposons, ici, un système de lexiques automatiques adapté aux applications en sciences humaines et plus particulièrement au traitement des grands vocabulaires. Il comprend :

- un lexique de référence pour le français parlé (en constante évolution : 6000 entrées actuellement, soient 37 000 formes morphosyntaxiques reconnues),

- un outil conversationnel permettant aux utilisateurs de générer leurs propres lexiques d'applications.

Ce système est opérationnel depuis 1978. Nous donnons quelques exemples de lexiques dérivés qui illustrent les performances actuelles du système.

* A.R.I.A.L. : Analyse et Reconnaissance de l'Information Acoustique et Linguistique.

** C.E.R.F.I.A. : Cybernétique des Entreprises Reconnaissance des Formes Intelligence Artificielle.

LES LEXIQUES AUTOMATIQUES POUR LE TRAITEMENT DE LA PAROLE : LA VERSION III DU LEXIQUE DU PROJET A.R.I.A.L.

G. GOUARDERES

SUMMARY

ABSTRACT :

To complete the communication of Professor PERENNOU, G., we describe an "evolutive automatic lexicons system" and its use in the field of speech recognition.

The lexicons (like dictionnaries), can play a double part in every problems of studies and automatic processing of natural languages :

- the first part consists in assigning a grammatical class, one or several meanings and different uses to every word (or lexical unity).

- the second part allows to simplify the grammar (and the rules) by a exhaustive list of exceptions and ambiguities.

We define briefly (Fig.1) the places and parts of lexicons in a comple simulated recognition system for continuous speech and large vocabulary.

This system is composed of :

- a referent lexicon for spoken french : his dimention an his content is always developed (6000 entries at the beginning, more than 37 000 identified morphosyntactic forms).

By means of specific operators, we generate application oriented lexicons and subs-lexicons.

For exemple : application oriented lexicons (derivate lexicons) of French Braille, Acoustic, phonetic, syllabic etc ...) (fig. 2 p. 7).

We, also, consider the whole structure of the complete system build by the method introduced by E.F. CODD for the DATA BASE RELATIONAL MODEL.

CONCLUSION :

We present the state of the art of a "tool" for social and linguistic sciences (i-e. Speech recognition) which is not explicitly tailored to a particular discipline. The generation of specific subsystems is very easy for no computer experimented end-users without a radical redesign of the original system. As the sytem is up on two computers (CII-IRIS 80 and Burroughs B 3500) Some samples of the actual performances will be given.

10èmes JOURNÉES D'ÉTUDE SUR LA PAROLE
GRENOBLE - 30 MAI - 1ER JUIN 1979

LES LEXIQUES AUTOMATIQUES POUR LE TRAITEMENT DE LA PAROLE :

G. GOUARDERES

GENERALITES SUR LES LEXIQUES EN RECONNAISSANCE DE LA PAROLE.

LE ROLE DES LEXIQUES : Pour l'étude du message parlé on sépare, habituellement les problèmes par niveaux selon trois grands groupes : structure physique du signal, aspects linguistiques, compréhension du discours.

On situe souvent la place du lexique au niveau linguistique, nous considérons plutôt, qu'il est le trait d'union entre les traitements et les données des différents niveaux (GOUARDERES, G., 1977a - PERENNOU, G., 1978). (Fig. 1p.2)

Il serait illusoire de vouloir représenter en extension tous les éléments de la langue avec leurs attributs. Le lexique est formé d'unités dont les composantes sont, soit immédiates, soit référencées (par règles). Ces unités ne sont pas homogènes, le choix de l'utilisation d'une règle étant lié à son efficacité. Dans cet exposé, le terme de lexique désigne, par opposition au dictionnaire, un ensemble homogène par niveaux (acoustique, phonologique, syntaxique, ...). Chaque élément est codé de façon standard du point de vue de sa nature, de son contenu et de sa structure.

Dans notre système : le lexique de base (ou lexique de référence) contient tous les attributs intrinsèques ou contextuels du mot.

- Un lexique dérivé est une restructuration de tout ou partie des attributs pour l'ensemble des mots du lexique de base.

- Un sous lexique est une restriction des entrées du lexique de référence possédant une ou plusieurs propriétés identiques.

- Un système de lexiques automatiques, comprend un lexique de base (ou de référence) évolutif en mode conversationnel (qui est l'ensemble des représentations standardisées par niveaux) et des lexiques dérivés (ou des sous-lexiques) qui sont codés et organisés de façon automatique à partir du lexique de base, en vue de leur consultation dans une application donnée.

LA PLACE DES LEXIQUES DANS UN SYSTEME

DE RECONNAISSANCE DE LA PAROLE

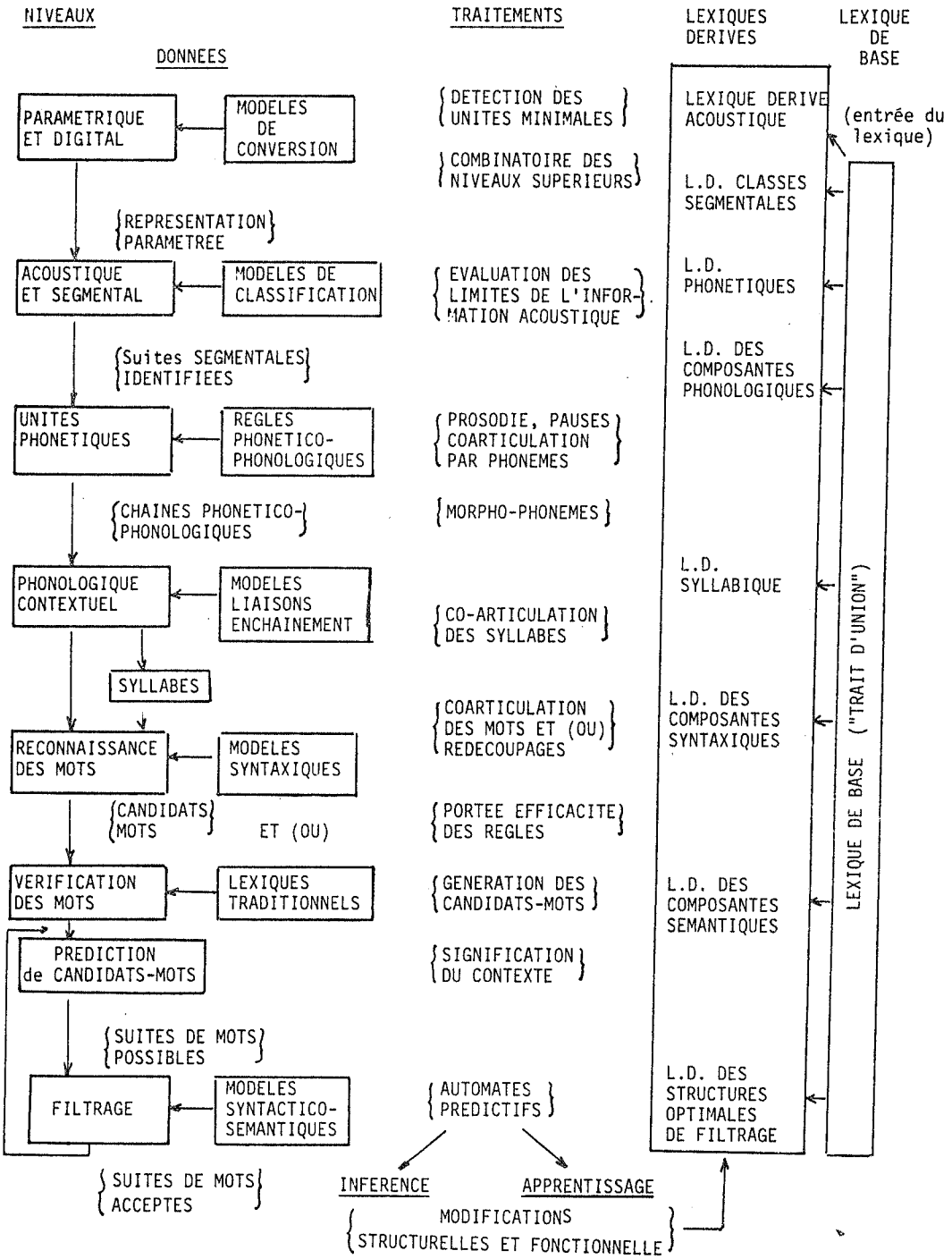


Fig. 1

LE SYSTEME DE LEXIQUES AUTOMATIQUES.

Le lexique de base : le "Corpus" de départ du langage étudié doit constituer un échantillon suffisamment représentatif pour permettre une évolution rapide des modèles et des règles.

Pour le français parlé, nous avons pris comme référence les deux dictionnaires français contemporains les plus utilisés : "Le Petit Robert" et "Le Petit Larousse" soit 50 000 mots environ. Parmi ceux-ci on distingue :

les 1 063 mots du français fondamental (GOUGENHEIM et AL. 1964)
3 000 mots du dictionnaire élémentaire
15 000 mots correspondants à la nomenclature de l'académie française
30 000 mots qui sont des termes spécialisés ou techniques.

Dans ce "Corpus" général on peut estimer à 24 000 le lexique individuel d'un locuteur moyen (c'est-à-dire le nombre de "mots en puissance" dont il dispose).

La version n°1 du lexique, avec des modèles morphosyntaxiques simples pouvait identifier 3 000 formes fléchies à partir des 1 063 mots du français fondamental. Malgré l'ajout d'environ 500 mots ce corpus s'est avéré notoirement insuffisant car :

- le choix des mots relevait d'une étude d'un langage un peu trop scolaire.

- Dans les 1 000 mots les plus fréquents on trouve surtout des "mots-outils". Toutefois pour le locuteur moyen ces mots outils (au nombre de 1 000 environ pour le corpus général) recouvrent la moitié de tout ennoncé. Si bien que l'on a une forte proportion de ces "mots-outils" tant que l'on a pas franchi un certain seuil.

C'est pourquoi, malgré l'amélioration des modèles de la version II (3 000 mots du dictionnaire élémentaire au départ), nous avons choisi pour la version III le "Dictionnaire du Vocabulaire Essentiel". (MATORE, G., 1965) en fonction de la répartition ci-dessous.

- les 100 premiers mots représentent 59 p.100 de tout texte
- les 1 000 mots suivants représentent 27 p.100 de tout texte
- les 3 000 mots suivants représentent 11,5 p.100 de tout texte (version II)
- les 5 000 mots (MATORE) CORPUS DE DEPART DE LA VERSION III
- les 20 000 mots suivants représentent 2,5 p.100 de tout texte.

(Selon les estimations du tableau précédent ceci représente 90% de tout texte pour le locuteur moyen). Cet ouvrage indique également la représentation phonétique (A.P.I.) et précise la ou les définitions de chaque mot.

Nous avons comparé son contenu à ceux des VERSIONS I ET II et complété par des formes irrégulières ou composées.

Exemples : MUET, MUETTE (Adjectif et nom) ; DOCTEUR, DOCTORESSE ; AUX DELA DE ; EST-CE QUE ; NON-SENS ; ... etc

Ce qui porte à 6 000 environ les entrées de la version III et à 37 000 le nombres des formes lexicales générées. (GOUARDERES, G., 1977b).

LA REPRESENTATION D'UN ELEMENT DANS LA BASE:(Description du modèle morpho-syntaxique) (GOUARDERES, G., 1977a).

Nous appellerons radical (graphique ou phonétique) la partie stable d'une unité lexicale(ou d'un mot) qui constituera une entrée dans le lexique de base. A ce radical (ou entrée) on associe :

- des références de règles (Modèles) qui fourniront les désinences des formes fléchies des mots.

- des attributs (ou rubriques) qui définiront les principales caractéristiques graphiques, phonétiques, syntaxiques et sémantiques d'un mot.

Exemple
VEN- -IR
VIEN- -DRAIS (ou VIEND- -RAIS)

Exemple de modèle d'accord des formes graphiques :

L'entrée (i) dans le fichier de base fournit le début d'un mot : RAD (i). La rubrique : Pointeur-graphique (i) donne l'étiquette de la liste des terminaisons possibles de RAD (i) : Soit DES (i,j) cette liste, l'index j caractérise dans cette liste la désinence choisie.

Exemple : extrait des figures 1, 2

i = 5383 RAD (i) = AFFAIBLI Pointeur-graphique (i) = H6.

Si l'on veut la 1e personne du pluriel du présent de l'indicatif, alors j=4 et DES (i,j) = SSONS.

Forme générée : RAD (i) + DES (i,j) = AFFAIBLISSONS.

REMARQUE :

On peut générer simultanément l'équivalent phonétique, acoustique ou maille.

LA COMPOSANTE PHONOLOGIQUE D'ADAPTATION : (PERENNOU, G., 1978).

Nous venons de montrer comment le modèle morpho-syntaxique génère de façon automatique l'invariant phonologique d'une forme graphique, Braille et plus généralement d'une entrée du lexique.

L'insertion des termes du lexique dans le discours en continu utilise actuellement une composante phonologique d'adaptation mise au point par PERENNOU, G., et TEP, G., (1977) en vue de la reconnaissance automatique de la parole (PERENNOU, G., & HATON, J.P. 1978) (TEP, G., 1978). On trouvera les directions pour une nouvelle version dans PERENNOU, G., (1979). Pour le lexique nous resumons la description de cette composante en deux points :

- les règles générales phonologiques permettent de prendre en compte les phénomènes de coarticulation de phonèmes, syllabes et mots. Nous les appelons règles de "soudure". Il existe, en outre, des phénomènes de mutation phonologique contextuels (Etat du locuteur, par exemple) qui peuvent également être décrits par règles.

- Chaque fois que l'application d'une règle est facultative, ou qu'il existe une exception le lexique doit en rendre compte. (Fig. 2 p. 7).

DESCRIPTION DE LA COMPOSANTE SYNTAXIQUE :

Elle est axée, dans la version III sur la catégorie syntaxique. Nous n'avons pas voulu, au niveau de la représentation des attributs d'un mot priviligier une approche conceptuelle de la syntaxe.

Toutefois, la méthode heuristique employée, ainsi que les opérateurs de dérivation généraux du lexique, permettent une utilisation immédiate de la "syntaxe structurale" selon TESNIERE, L..

En effet, les conventions utilisées (codes syntaxiques, catégories verbales etc ..., permettent de distinguer les types d'actants et de circonstants, les valences du verbe et les notions de translations dans un automate prédictif pour les structures de phrases simples.

DESCRIPTION DE LA COMPOSANTE SEMANTIQUE :

Cette composante du lexique est en cours de développement dans la version III. Comme pour la composante syntaxique, nous essayons par une approche heuristique des "qualités" sémantiques (± concret, ± animé, etc ...) de définir une structure sémantique de la phrase simple.

LA STRUCTURE DU LEXIQUE : (GOUARDERES, G., 1979, CHRISMENT, C.Y., 1978).

La structure adoptée permet une représentation "instantannée" de la connaissance. Nous avons abordé le problème de la structuration des données sous le double aspect de leur nature formelle (Algèbre relationnelle) et de leur fonction logique (régularités, lois, usages) (GOUARDERES, G., & AL, 1979). Nous n'insisterons pas sur l'aspect informatique de cette réalisation décrite par ailleurs dans GOUARDERES, G., (1977a, 1977b, 1979). Elle répond aux contraintes suivantes :

1 - Le lexique est "ouvert" et évolue constamment en mode conversationnel.

2 - Les règles et les modèles associés évoluent en conséquence selon le même principe.

3 - L'obtention immédiate de lexiques dérivés est simple et nécessite un minimum de programmation.

4 - La propriété de trait d'union du lexique de base est préservée, aussi bien pour les relations "horizontales" que "verticales" définies précédemment. (Fig. 1 p. 2)

A partir de ce schéma conceptuel (effectivement réalisé par une structure relationnelle au sens de CODD), nous avons développé pour les utilisateurs

potentiels de ces lexiques :

- un lexique de référence pour le français parlé (en constante évolution, plus de 6 000 entrées actuellement, soient 37 000 formes morphosyntaxiques reconnues).

- un outil conversationnel, leur permettant de générer leurs propres lexiques d'applications et les statistiques afférentes aux unités discursives considérées. (Statistiques sur les syllabes phonologiques du français fondamental, par ex.).

CONCLUSION ET PERSPECTIVES.

Les résultats obtenus, tant en reconnaissance de la parole qu'en étude de texte, ou pour la production simultanée Braille-Parole montrent l'utilité du lexique dans l'étude des grands vocabulaires.

Nous pensons que le système de lexiques automatiques proposé est un outil déterminant pour la définition de la stratégie de reconnaissance et des traitements afférents en reconnaissance de la parole.

La prochaine version (version IV) du système de lexiques automatiques prévoit :

1 - Une évolution de la structure vers une base de données réparties.

2 - L'extension du nombre des entrées (10 000) et des modèles (soit plus de 100 000 formes reconnues).

3 - La description complète de la composante sémantique avec les performances des automates associés.

4 - L'extension des opérateurs généraux du système, en particulier ceux d'inférence et d'adaptation à partir de modules d'apprentissage automatique.

LEXIQUE DE BASE

0103ADRESSE	ADR&SE	N1M1N F
0104ADROIT	ADR<A	J3M8J
0105ADROITEMENT	ADR<ATEM4	D
0106ADULTE	ADULTE	J1L1K
0107ADVERBE	ADV&RBE	N1M1N M
0108ADVERSAIRE	ADV&RS&RE	N1M1N
0109AERIEN	A& RYJ4M5J	
0110AEROPORT	A&ROPOR	N1M1N M
0111AFFAIBLI	AF&BLI	H6X1V
0112AFFAIRE	AF&RE	N1M1N F
0113AFFECTION	AF&KSY2	N1M1N F

LEXIQUE DERIVE SYLLABIQUE

AERIEN	A&RY5	A & RY5	MASC SIN	J	J4	M5
AERIENNE	A&RY&NE	A & RY& NE	FEM SING			
AERIENS	A&RY5Z+	A & RY5 Z+	MASC PLU			
AERIENNES	A&RY&NE4+	A & RY& NE 4+	FEM PLUR			
AEROPORT	A&ROPOR	A & RO POR	SING	N	N1	M1
AEROPORTS	A&ROPORZ+	A & RO POR Z+	PLUR			
AFFAIBLIS	AF&BLIZ+	A F& BLI Z+	1PS IND PRES	V	M6	A1
AFFAIBLIS	AF&BLIZ+	A F& BLI Z+	2PS IND PRES			
AFFAIBLIT	AF&BLIT+	A F& BLI T+	3PS IND PRES			
AFFAIBLISSONS	AF&BLIS2Z+	A F& BLI S2 Z+	1PP IND PRES			
AFFAIBLISSEZ	AF&BLIS6Z+	A F& BLI S6 Z+	2PP IND PRES			
AFFAIBLISSENT	AF&BLISET+	A F& BLI SE T+	3PP IND PRES			
AFFAIBLIRAI	AF&BLIR7	A F& BLI R7	1PS IND FUTU			
AFFAIBLIRAS	AF&BLIRAZ+	A F& BLI RA Z+	2PS IND FUTU			
AFFAIBLIRA	AF&BLIRA	A F& BLI RA	3PS IND FUTU			
AFFAIBLIRONS	AF&BLIR2Z+	A F& BLI R2 Z+	1PP IND FUTU			
AFFAIBLIREZ	AF&BLIR6Z+	A F& BLI R6 Z+	2PP IND FUTU			
AFFAIBLIRONT	AF&BLIR2T+	A F& BLI R2 T+	3PP IND FUTU			
AFFAIBLISSAIS	AF&BLIS7Z+	A F& BLI S7 Z+	1PS IND IMPA			
AFFAIBLISSAIS	AF&BLIS7Z+	A F& BLI S7 Z+	2PS IND IMPA			

LEXIQUE DERIVE ACOUSTIQUE

VCFVCFVCVF	ADVERSAIRES	ADV&RS&RE4+	P N	ADVERSAIRE
VVCCV	AERIEN	A&RY5	MS J	
VVCCVCV	AERIENNE	A&RY&NE	FS J	AERIEN
VVCCVF	AERIENS	A&RY5Z+	MP J	AERIEN
VVCCVCVF	AERIENNES	A&RY&NE4+	FP J	AERIEN
VVCVOVC	AEROPORT	A&ROPOR	MS N	
VVCVOVCF	AEROPORTS	A&ROPORZ+	MP N	AEROPORT
VFVCCVF	AFFAIBLIS	AF&BLIZ+	1PS V PERI	AFFAIBLIR
VFVCCVF	AFFAIBLIS	AF&BLIZ+	2PS V PERI	AFFAIBLIR
VFVCCVO	AFFAIBLIT	AF&BLIT+	3PS V PERI	AFFAIBLIR
VFVCCVFVF	AFFAIBLISSONS	AF&BLIS2Z+	1PP V PERI	AFFAIBLIR
VFVCCVEVF	AFFAIBLISSEZ	AF&BLIS6Z+	2PP V PERI	AFFAIBLIR
VFVCCVFVO	AFFAIBLISSENT	AF&BLISET+	3PP V PERI	AFFAIBLIR
VFVCCVCV	AFFAIBLIRAI	AF&BLIR7	1PS V FUTI	AFFAIBLIR
VFVCCVCVF	AFFAIBLIRAS	AF&BLIRAZ+	2PS V FUTI	AFFAIBLIR
VFVCCVCV	AFFAIBLIRA	AF&BLIRA	3PS V FUTI	AFFAIBLIR
VFVCCVCVF	AFFAIBLIRONS	AF&BLIR2Z+	1PP V FUTI	AFFAIBLIR
VFVCCVCVF	AFFAIBLIREZ	AF&BLIR6Z+	2PP V FUTI	AFFAIBLIR
VFVCCVCVO	AFFAIBLIRONT	AF&BLIR2T+	3PP V FUTI	AFFAIBLIR
VFVCCVFVF	AFFAIBLISSAIS	AF&BLIS7Z+	1PS V IMPI	AFFAIBLIR
VFVCCVFVF	AFFAIBLISSAIS	AF&BLIS7Z+	2PS V IMPI	AFFAIBLIR
VFVCCVFVO	AFFAIBLISSAIT	AF&BLIS7T+	3PS V IMPI	AFFAIBLIR
VFVCCVFCVF	AFFAIBLISSIONS	AF&BLISY2Z+	1PP V IMPI	AFFAIBLIR
VFVCCVECVF	AFFAIBLISSIEZ	AF&BLISY6Z+	2PP V IMPI	AFFAIBLIR
VFVCCVFVO	AFFAIBLISSAIENT	AF&BLIS7T+	3PP V IMPI	AFFAIBLIR
VFVCCVC	AFFAIBLIR	AF&BLIR+	V INFI	AFFAIBLIR
VFVCCVFVO	AFFAIBLISSANT	AF&BLIS4T+	V PAPE	AFFAIBLIR

Fig. 2

REFERENCES

CAUSSE, B. , GOUARDERES, G. , TEP, G. , 1976, L'utilisation du lexique pour la génération des phrases parlées rapport CERFIA.

CAUSSE, B. , GOUARDERES, G. , TRUQUET, M. , 1979, (à paraître), "French Braille (Grade 2) and synthetic voice". -Computerized Braille production-today and tomorow". London conference.

CHRISMENT, C.Y. , 1978 Systèmes de gestions de bases de données - caractéristiques générales. Cours M.I.A.G.E. Université Paul Sabatier Toulouse III.

GOUARDERES, G. , 1977a, Organisation d'un lexique en vue de l'analyse de la parole en continu diplôme ingénieur C.N.A.M.

GOUARDERES, G. , 1977b, Saisie et représentation sur ordinateur d'un lexique pour le français parlé. -Rapport C.E.R.F.I.A.

GOUARDERES, G. & ALL , 1979 (à paraître), Un lexique évolutif pour les sciences humaines. Congrès A.F.C.E.T. Toulouse.

GOUGENHEIM, G. , ALL. , 1964, L'élaboration du français fondamental.

MATORE, G. , 1963, Dictionnaire du vocabulaire essentiel.

PERENNOU, G. , ALL. , 1977, "About some lexical analysis problems in connected speech recognition". Rapport CERFIA.

PERENNOU, G. , HATON, J.P. , 1978, Reconnaissance automatique de la parole. 8e Ecole d'été informatique de l'A.F.C.E.T.

PERENNOU, G. , 1979 Voir communication dans ce même congrès.

TEP, G. , 1978, Contribution à l'étude phonologique d'un système d'analyse de la parole continue. -Thèse 3e cycle Université Paul Sabatier Toulouse.

WHITE, G.M. , 1978, Large vocabulary speech recognition and dictionary speech strategie. Fourth International joint conference on Pattern Recogntion. Kyoto-Japan.

10èmes JOURNÉES D'ÉTUDE SUR LA PAROLE

GRENOBLE - 30 MAI - 1ER JUIN 1979

FORMALISATION DU LEXIQUE ET DES REGLES PHONOLOGIQUES DANS LE SYSTEME ESOPE Ø

MARIANI Joseph-Jean

Laboratoire d'Informatique pour la Mécanique et
les Sciences de l'Ingénieur (L.I.M.S.I.-C.N.R.S.)
B.P. 30, 91406 ORSAY Cedex, FRANCE

RESUME

Le système ESOPE Ø est orienté vers la reconnaissance de parole continue en temps réel, et pour des vocabulaires limités. Il utilise des outils linguistiques simples et performants comme aide à la reconnaissance.

La facilité d'emploi de ces outils linguistiques a permis une rapide adaptation du système à diverses applications tant dans la reconnaissance de phrases, que dans le dialogue ou la détection de mots.

On présente ici la formalisation du lexique, et l'organisation de matrices reflétant à la fois les performances du système de décodage phonétique et certaines réalités phonologiques de la langue.

FORMALIZATION OF THE LEXICON AND PHONOLOGICAL RULES IN ESOPE Ø

MARIANI Joseph-Jean
L.I.M.S.I. (C.N.R.S.)
B.P. 30, 91406 ORSAY Cedex, FRANCE

SUMMARY

The speech "understanding" system ESOPE Ø is oriented towards continuous speech dialog, in real time, for small vocabularies. It uses simple linguistic methods to correct the phoneme recognition. As there methods are easy to use, they allow for a fast adaptation of the system to many applications in sentence recognition, dialog or word spotting.

We present here the formalization of the lexicon, and the organization of matrixes that show both the quality of the phoneme decoding and some phonological rules of the french language.

10èmes JOURNÉES D'ÉTUDE SUR LA PAROLE
GRENOBLE - 30 MAI - 1ER JUIN 1979

FORMALISATION DU LEXIQUE ET DES REGLES PHONOLOGIQUES DANS LE SYSTEME ESOPE Ø

MARIANI Joseph-Jean
L.I.M.S.I. (C.N.R.S.)
B.P. 30, 91406 ORSAY Cedex

INTRODUCTION

Dans la mesure où la reconnaissance automatique des phonèmes qui composent une phrase donne actuellement des résultats insuffisants (inférieurs en général à 70% de phonèmes correctement reconnus alors qu'un taux de 85% semble nécessaire à un auditeur humain (1977, J.S. LIENARD, J.J.MARIANI, G. RENARD)), le niveau lexical est le niveau d'articulation privilégié dans les systèmes de reconnaissance automatique de parole.

ROLE DU NIVEAU LEXICAL

C'est lui qui transforme les traits acoustiques ou les phonèmes reconnus par les étages inférieurs, en mots lexicaux, avec l'aide éventuelle d'informations venues des étages supérieurs (syntaxe, sémantique, pragmatique). Il constitue ainsi dans la plupart des systèmes, une barrière très nette entre les niveaux inférieurs et les niveaux supérieurs, que les informations venues de ces niveaux respectifs ne franchissent pas (figure 1). En effet les niveaux supérieurs n'ont pas besoin de savoir que le phonème "U" vient d'être reconnu et les niveaux inférieurs de savoir que le mot prononcé est un adverbe. Par contre il leur est nécessaire de connaître la valeur syntaxique, sémantique ou pragmatique du mot pour les uns, son contenu phonétique ou acoustique pour les autres. Ces informations figurent donc dans le lexique, avec éventuellement des informations de type phonologique (liaisons, altérations) (figure 1).

LE ROLE DE LA SYNTAXE SUR LE LEXIQUE

Le rôle de la syntaxe dans les systèmes de reconnaissance de parole est très étendu. Précisons tout de suite que le plus souvent le niveau sémantique est confondu avec le niveau syntaxique puisque dans la plupart des systèmes existants, le sens des mots est directement lié au choix du lexique et de l'articulation, représentée syntaxiquement, des éléments lexicaux dans la phrase,

en fonction du contexte d'action (ou application).

Le rôle donc de l'étage syntactico-sémantique est double (figure 2) :

a) aider à la reconnaissance des mots en limitant le nombre de mots reconnaissables à chaque instant. Ceci est valable aussi bien dans la reconnaissance acoustique de mots (dite reconnaissance globale) où le mot est reconnu globalement grâce à son image acoustique présente dans le lexique, ce qui nécessite un apprentissage, que dans la reconnaissance phonétique de mots (dite reconnaissance analytique) où le mot est reconnu grâce à son contenu phonétique introduit manuellement ou obtenu automatiquement à partir de la forme graphèmique du mot et d'un programme de traduction phonétique (1979, B. PROUTS). Et c'est également aussi valable pour une reconnaissance de mots isolés, où chaque mot isolé doit être suivi d'un silence d'environ 0.3 s. qui délimite l'entité lexicale, que pour la reconnaissance de parole continue où la délimitation est faite à la fin de groupes de sens, de phrases ou de groupes de phrases.

b) aider à la compréhension des mots en transmettant au niveau pragmatique la valeur syntaxique (au sens large) des mots reconnus, ou, dans de nombreux systèmes, commander directement l'exécution d'une tâche lorsqu'une instruction (ensemble de plusieurs mots) a été reconnue.

Dans la mesure où les études menées dans les sciences cognitives sont loin d'aboutir, et si l'on ajoute à cela la difficulté supplémentaire amenée par l'imprécision de la reconnaissance phonétique, on verra facilement l'importance du chemin restant à parcourir, et on découvrira aisément l'abus de langage consistant à parler de "compréhension de la parole" ("speech understanding") là où il n'y avait qu'aide syntaxique à la reconnaissance des mots.

LE ROLE DU NIVEAU PRAGMATIQUE SUR LE LEXIQUE

Le niveau pragmatique intervient dans quatre directions (1979, J.J. MARIANI, J.S. LIENARD, G. RENARD) :

- établir le dialogue en répondant au locuteur ou en lui posant des questions par synthèse vocale (1978, G. MERCIER, P. QUINTON, R. VIVES),
- déclencher éventuellement l'exécution d'une tâche et donner, ou demander, des informations vocales concernant le déroulement de cette tâche,
- prédire la syntaxe et le lexique utilisables par le locuteur humain lors de sa prochaine intervention vocale,(1975 , PIERREL)
- dans l'éventualité où le système de reconnaissance a un niveau sémantique effectif, choisir le sens des mots en fonction du contexte d'application,

du dialogue passé et de l'état de la tâche.

ROLE DES NIVEAUX ACOUSTIQUE ET PHONETIQUE SUR LE LEXIQUE

Deux types de reconnaissance sont essentiellement discernables : ceux nécessitant un apprentissage des entités lexicales (prononciation de ces entités et étiquetage préalables à leur reconnaissance), ceux ne nécessitant qu'un apprentissage des entités phonétiques, et l'élaboration des règles permettant de pallier les erreurs de reconnaissance (matrices de confusion par exemple). Dans la première catégorie, le stockage des informations se fait au niveau d'entités acoustiques, avec ou sans segmentation, ou d'entités phonétiques, après reconnaissance de phonèmes ou pseudo-phonèmes. Ces informations sont accompagnées de l'étiquette du mot prononcé.

FORMALISATION DU LEXIQUE EN FONCTION DE LA SYNTAXE ET DU TYPE DE RECONNAISSANCE

Il sera question ici essentiellement de la formalisation du lexique dans le système ESOPE Ø, mais également, très brièvement, dans d'autres systèmes de communication parlée expérimentés au L.I.M.S.I.

LEXIQUE DANS UN SYSTEME DE DIALOGUE PAR RECONNAISSANCE GLOBALE DE MOTS ISOLES

Le rôle de la syntaxe est ici prépondérant que ce soit pour uniquement prédire les mots acceptables (commande vocale à bord d'avion), ou, au choix, soit prédire ces mots soit vérifier leur validité et guider leur correction (programmation vocale en FORTRAN (1979, F. NEEL)). Dans le premier cas, la syntaxe choisie est arborescente, décrite par une structure de liste, chaque noeud pouvant être constitué par un sous-arbre. L'apprentissage de cette syntaxe peut être fait par l'opérateur assisté de l'ordinateur qui formalise l'arbre et détermine alors le vocabulaire non terminal, le vocabulaire terminal et demande à l'opérateur de prononcer les mots de ce dernier vocabulaire. La syntaxe détermine donc directement le numéro des mots acceptables. Dans certains états, elle présente le menu des expressions que le pilote peut prononcer sur un écran graphique, et chaque mot reconnu peut être synthétise et/ou affiché sur l'écran. Le lexique comporte donc le numéro du mot, son écriture graphèmique, son écriture phonétique, son contenu acoustique. Dans le second cas, la syntaxe décrite par un automate porte sur les catégories syntaxiques des mots, et cette catégorie figure donc dans le lexique. Dans les deux cas, le contenu acoustique est codé sur 120 octets (1979, J.S. LIENARD) et le lexique comporte

une centaine de mots.

LEXIQUE DANS LE SYSTEME DE RECONNAISSANCE DE PAROLE CONTINUE ESOPE Ø

1. Les règles phonétiques

La reconnaissance de la parole continue est faite ici après reconnaissance des phonèmes qui composent la phrase. Pour chaque segment détecté, on considère les quatre phonèmes les mieux reconnus, et on constitue ainsi un treillis phonétique où figurent le nom des phonèmes et leur note de reconnaissance (1978, MARIANI, J.J., LIENARD, J.S.). Le taux de reconnaissance du phonème exact dans le treillis est de l'ordre de 80%. On peut définir trois types d'erreur : confusion (le bon phonème n'est pas reconnu en première position), élision (un segment correspond à deux phonèmes), ajout (deux segments correspondent à un phonème). Pour pallier ces erreurs, des règles ont été écrites à partir des performances du système, et des réalités de la langue (confusion entre phonèmes et élisions dans la langue courante, phonèmes différents d'un trait distinctif). Ces règles sont décrites par une matrice de confusion (30 x 30 phonèmes) qui mesure la dissemblance entre le phonème correct et les phonèmes rencontrés dans le treillis, et une matrice d'élision (30 x 30 phonèmes) qui détermine la possibilité d'élision dans une suite de deux phonèmes. Ces matrices évoluent comme le système : de nouvelles règles sont créées lorsqu'on rencontre les suites de phonèmes correspondantes, des règles disparaissent, ou sont modifiées avec les progrès du module de reconnaissance phonétique (figure 3).

2. Détermination du lexique et de la syntaxe par le niveau pragmatique

Dans le dialogue de parole continue, le niveau pragmatique a pour but de prédire la syntaxe et le lexique que peut utiliser le locuteur pour poser une question, ou répondre à une question posée par l'ordinateur.

Chaque lexique et chaque syntaxe sont donc définis à partir de la tâche, et de l'étape du dialogue au sein de la tâche. Il existe une bibliothèque de sous-langages, où chaque lexique et chaque syntaxe sont répertoriés et appelables par leur numéro. Ainsi, dans une application de dialogue téléphonique, la syntaxe change suivant que le système attend la demande de renseignement du locuteur, ou la confirmation de la reconnaissance, le lexique restant le même (figure 6) (Application commune à d'autres équipes (1975, PIERREL, 1978, MERCIER) .

3. Formalisation du lexique en fonction de la syntaxe utilisée dans ESOPE Ø

Dans le système de reconnaissance de la parole continue ESOPE Ø, le but

était de pouvoir fonctionner en dialogue parlé avec l'ordinateur, et en temps réel (moins d'une seconde de traitement pour reconnaître une phrase). La reconnaissance phonétique a donc été simplifiée et les aides linguistiques ont été choisies pour être facielemnt apprises afin d'atteindre ce résultat. Le rôle de la syntaxe dans les systèmes de "compréhension" de parole continue étant essentiellement actuellement de pallier les erreurs de reconnaissance phonétique, il n'est pas nécessaire de posséder la syntaxe exacte des phrases prononçables mais il suffit de posséder une grammaire permettant de générer un sur ensemble du langage. C'est la raison qui nous a conduit à choisir une syntaxe de type matrice de précédence portant sur des catégories syntaxiques, bien que sur les langages utilisés jusqu'à présent, où la taille des vocabulaires est faible, des types de syntaxe plus contraignante seraient facilement utilisables. Par contre pour de très grands vocabulaires (170.000 formes) et pour une syntaxe voisine de la langue naturelle, la facilité d'apprentissage des syntaxes locales nous a permis l'obtention de résultats très intéressants (1979, ANDREEWSKY, A., BINQUET, J.P., DEBILI, F., FLURH, C., HLAL, Y., LIENARD, J.S., MARIANI, J.J., POUDEROUX, B.). Dans ESOPE Ø, la possibilité de succession de deux catégories syntaxiques est donnée par un choix binaire, qui peut éventuellement indiquer la nécessité d'un accord en genre et en nombre entre les catégories syntaxiques (1978, J.J. MARIANI, J.S. LIENARD). (figure 4)

Le lexique comporte l'écriture graphémique du mot, avec le nombre de graphèmes, son écriture phonétique, avec le nombre de phonèmes, sa catégorie syntaxique, éventuellement le genre et le nombre du mot et des informations phonologiques concernant la liaison avec le mot suivant : possibilité d'une liaison en (z), (n), (t), etc, possibilité de l'élision d'un E muet, nécessité pour le mot suivant de commencer par une consonne, ou par une voyelle (pour les mots suivant les articles). Ces informations phonologiques, ainsi que l'écriture phonétique du mot ont été données à la main. La réalisation d'un programme de traduction phonétique très performant (1979, B. PROUTS) va automatiser très prochainement ce processus à partir de l'écriture graphémique du mot. La taille des vocabulaires varie entre 12 et 50 mots, le facteur de branchement de la syntaxe entre 4 et 12 suivant le type de phrases à reconnaître ((sujet),(verbe),(complément) , suite de chiffres, nombres de 0 à 100, de 0 à 1 milliard, expressions arithmétiques, conversation téléphonique).

4. Stratégie de reconnaissance

Le système utilise une stratégie prédiction-vérification (figure 5). On

conserve les meilleures phrases (best few) sans faire de retour arrière (no-backtracking). La reconnaissance se fait de gauche à droite et à chaque instant donc, on limite le nombre de phrases acceptées et on rejette les phrases ayant une note de reconnaissance insuffisamment bonne. A la fin de la reconnaissance, la phrase qui a la meilleure note, et qui est syntaxiquement correcte, est choisie. La figure 7 donne un exemple de segments de phrases conservés à plusieurs étapes de la reconnaissance d'une phrase, avec les mots possibles par la suite.

CONCLUSION

Les résultats donnés en (1979, ANDREEWSKY, A. ...) présentent bien le dilemme présent dans la reconnaissance de la parole continue : faut-il développer des algorithmes très sophistiqués, coûteux en temps et en complexité, aux niveaux supérieurs, pour pallier l'insuffisance de la reconnaissance phonétique, ou porter les efforts uniquement sur la reconnaissance phonétique sachant qu'une reconnaissance parfaite permettrait, dès à présent, une orthographication exacte à 95% au niveau syntaxique pour la langue quasi-naturelle.

Dans la mesure où la qualité de la reconnaissance phonétique nous interdit l'apprentissage des concepts par la machine sur support vocal, faut-il se substituer aux sémanticiens dans l'étude de la formalisation des concepts alors que la difficulté déjà grande est ici élevée au carré ?

L'attrait et la puissance du dialogue parlé, qui peuvent être ressentis dans les systèmes fonctionnant déjà en conversationnel et en temps réel, nous poussent cependant dans cette voie manifestement déraisonnable.

REFERENCES

LIENARD, J.S., MARIANI, J.J., RENARD, G., Intelligibilité de phrases synthétiques altérées : application à la transmission phonétique de la parole, 9^{e} Congrès International d'Acoustique, Madrid, 1977.

PROUTS, B., Traduction de textes écrits en français, 10^{e} J.E.P., Grenoble, 1979

MARIANI, J.J., LIENARD, J.S., RENARD, G., Speech recognition in the context of two-way immediate man-machine interaction, IEEE-ICASSP, Washington, 1979.

MERCIER, G., QUINTON, P., RIVES, R., KEAL : un système pour un dialogue oral avec une machine, Congrès AFCET, Théorie et Technique de l'Informatique, Paris, 1978.

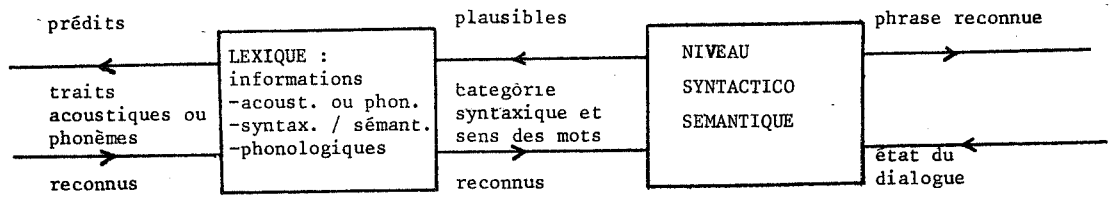
NEEL, F., Programmation vocale en Fortran, Rapport interne LIMSI, 1979.

LIENARD, J.S., Speech characterization from a rough spectral analysis, IEEE-ICASSP, Washington, 1979.

MARIANI, J.J., LIENARD, J.S., ESOPE Ø : un programme de compréhension automatique de la parole procédant par prédiction-vérification aux niveaux phonétique, lexical et syntaxique, Congrès AFCET-IRIA, Reconnaissance des Formes et Traitement des Images, 1978.

ANDREEWSKY, A., BINQUET, J.P., DEBILI, F., FLUHR, C., HLAL, Y., LIENARD, J.S., MARIANI, J.J., POUDEROUX, B., Les dictionnaires en formes complètes et leur utilisation dans la transformation lexicale et syntaxique correcte de chaînes phonétiques, 10e J.E.P., Grenoble, 1979.

PIERREL, J.M.; Contributions à la compréhension automatique du discours continu , Thèse de spécialité , Université NANCY 1 , 1975



Figures 1 & 2 :-Rôle des niveaux lexical et syntactico-sémantique
-The function of the lexical and syntactic/semantic levels

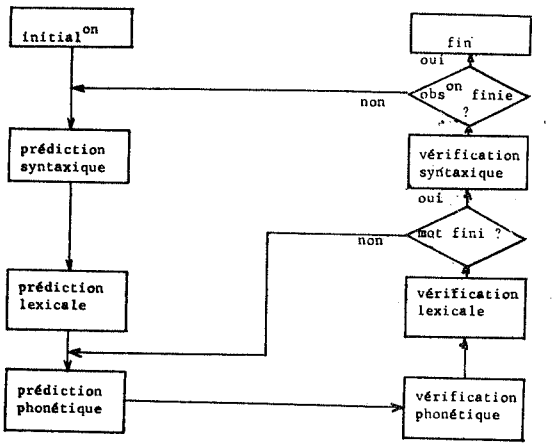
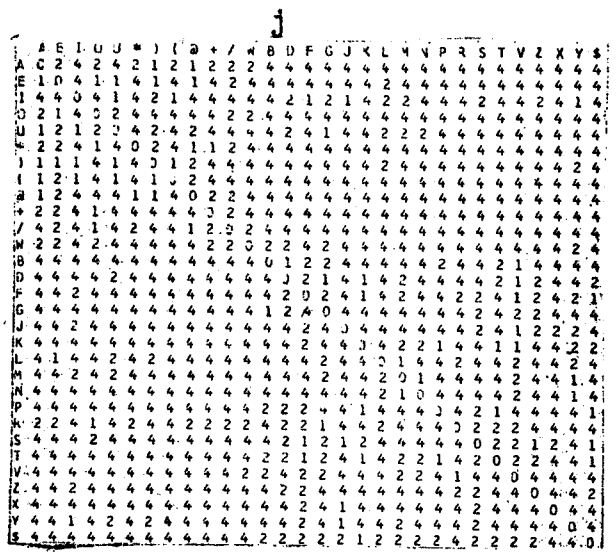
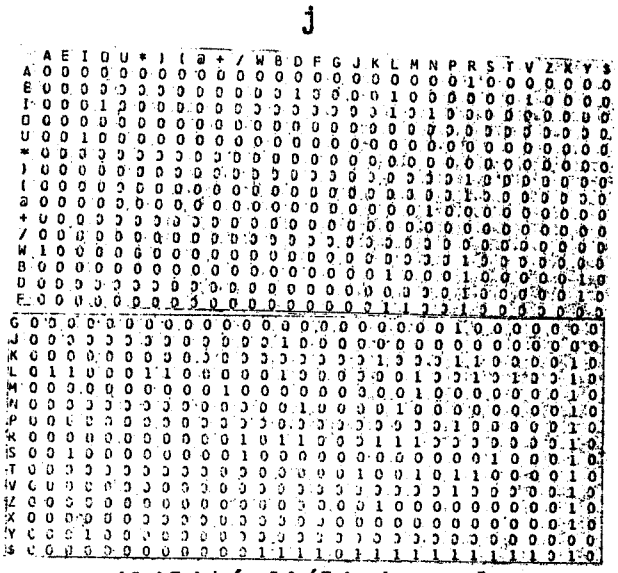


FIGURE 5 :-Stratégie utilisée dans le système ESOPEØ
-Strategy in the system ESOPEØ

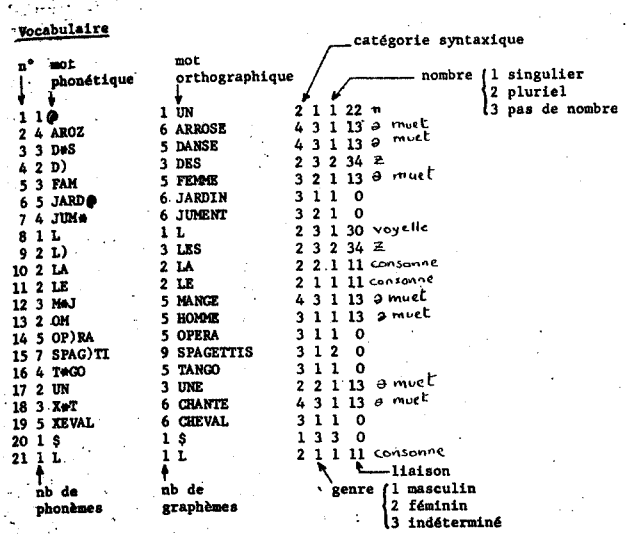


simi litude entre le phonème i (prédit) et le phonème j (reconnu)



possibilité d'élision du phonème i suivi du phonème j

Figure 3 :-Matrices de confusion et d'élision
-similarity and elision matrixes



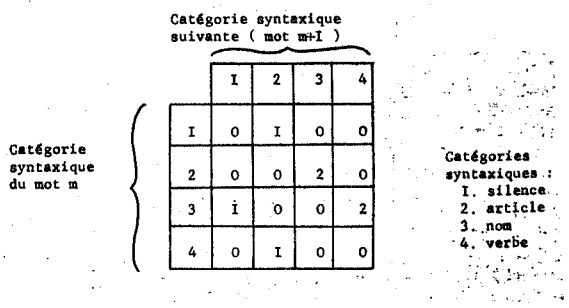
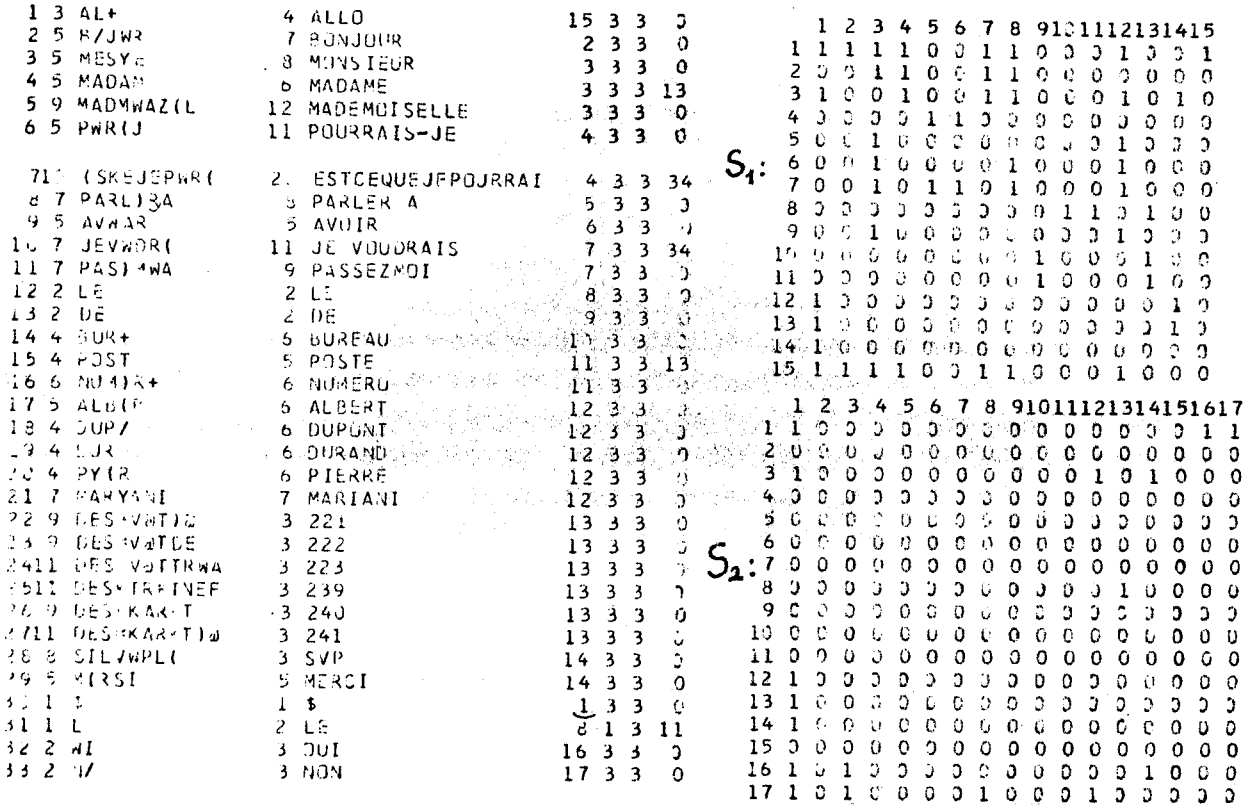


Figure 4 :-Exemple de lexique et de syntaxe pour une application sujet-verbe-complément
-The lexicon and the syntax in the recognition of SVC sentences.



phrase reconnue	réponse synthétisée	prédits synt.	prédits lex.	phrase prononcée
...	...	S1	1	pourrais-je parler à dupont ?
pourrais-je parler à dupont	est-ce que vous avez demandé dupont ?	S2	1	oui ,s'il vous plait.
oui , s'il vous plait .	je vous le passe	S1	1	allo !
allo	qui demandez-vous?	S1	1	je voudrais le poste 240 .
je voudrais le poste de durand	voulez-vous parler à durand ?	S2	1	Non,le 240 !
non,le 240	est-ce que vous avez demandé le 240 ?	S2	1	Oui,monsieur .
oui,madame	je vous passe votre correspondant			

Figure 6 :-Exemple de lexique et de 2 syntaxes ,et rôle de la pragmatique dans une application de standard téléphonique .
-Lexicon,syntax and pragmatics in a telephone dialog .

	phrase retenue	note	mots prédits
étape 1	est-ce	270	que je pourrais
	pourrais-je	290	(parler à / avoir)
	allo	320	DEBUT DE DEMANDE
ét. 2	pourrais-je parler	500	à
	allo allo	590	DEBUT DE DEMANDE
	Allo monsieur	640	(5 noms ou DEBUT DE DEMANDE)
ét. 3	Pourrais-je parler à durand	820	FIN ou SVP
	allo allo le 200	930	FIN ou SVP
	allo allo le bureau	960	N° ou de +
	allo monsieur durand	970	(FIN ou SVP ou DEBUT DE DEMANDE)

Figure 7 :-Exemple de reconnaissance d'une phrase avec conservation des meilleures hypothèses
-Example of a sentence recognition

10èmes JOURNÉES D'ÉTUDE SUR LA PAROLE

GRENOBLE - 30 MAI - 1ER JUIN 1979

Formalisation d'un lexique en logique du premier ordre pour la reconnaissance automatique de la parole.

MELONI H.

Groupe Intelligence Artificielle - FACULTE DES SCIENCES DE LUMINY - CASE 901, 70 Route Léon Lachamp - 13288 MARSEILLE Cedex 2

RESUME

Nous présentons une formalisation d'un lexique, qui s'intègre dans un système de traitement des contraintes linguistiques pour la reconnaissance automatique de la parole. Chaque mot est représenté par une formule atomique de logique du premier ordre (clause unaire du langage PROLOG), l'accés étant fait par sa catégorie lexicale, dans le cadre d'une analyse descendante sur les catégories. Les informations qui le caractérisent sont les suivantes :

- Un ensemble de traits syntactico-sémantiques qui conditionnent ses occurrences dans les divers contextes de la phrase

- La séquence de ses phonèmes ainsi que les marques des modifications contextuelles qu'elle peut subir

- Sa description orthographique et ses variations.

Formalisation d'un lexique en logique du premier ordre pour la reconnaissance automatique de la parole.

MELONI H.

SUMMARY

Whitin the framework of automatic recognition of continuous speech, we present here a formalization of a lexical component, which is a subpart of general treatment of linguistic constraints.

A word is represented by a first-order predicate logic atomic formula. The access is given by the lexical category with a top-down analysis.

The characteristic informations of a word are the following :

- a set of syntactic and semantic features predicting the appearance of a given word in every context

- the sequence of its phonemes and the possible contextual alterations

- its orthographic description with its variations.

10èmes JOURNÉES D'ÉTUDE SUR LA PAROLE
GRENOBLE - 30 MAI - 1er JUIN 1979

Formalisation d'un lexique en logique du premier ordre pour la reconnaissance automatique de la parole.

MELONI H.

INTRODUCTION

Nous présentons une formalisation du lexique en logique du premier ordre, en vue de la reconnaissance automatique de la parole continue. Après avoir situé la composante lexicale dans notre système de traitement des informations linguistiques, nous examinerons le codage des caractéristiques (que l'on ne peut consigner dans des règles générales) propres à chacun des mots du dictionnaire, ainsi que l'utilisation qui en est faite.

SITUATION DU LEXIQUE DANS LE SYSTEME DE TRAITEMENT DES CONTRAINTES LINGUISTIQUES

Nous avons réalisé un système permettant la description de grammaires ainsi que la programmation de diverses stratégies d'analyse des chaînes terminales (remontante, descendante, gauche-droite, droite-gauche, accès direct sur des terminaux particularisés). Dans ce système, nous avons programmé un analyseur de phrases d'un sous-ensemble du français qui utilise une combinaison de toutes ces techniques (MELONI, H., 1978). Une des conséquences de la méthode employée, est l'optimisation de la recherche d'un mot dans le lexique par l'intervention de toutes les informations connues à cette étape de l'analyse. Les mots sont sollicités de manière descendante, ce qui impose un accès par la catégorie lexicale qui leur est associée.

Chaque mot est représenté par une clause unaire du langage PROLOG (ROUSSEL, P., 1975), (MELONI, H., 1976) dont le prédicat d'appel est le nom de la catégorie lexicale à laquelle appartient ce mot. Les informations concernant chacun des éléments du lexique sont codées sous forme de termes (arbres construits à partir de l'opérateur binaire droite-gauche ".").

Exemples : ART(M.S.DEF,LE.L.E(E)):..

NOM(M.*N.NH.NA.CC.NL, (MORCEAU.X).(M.O.R.S.AU)./$):..

VERBE(P.H.T.AT.NA.PR(SUR.DANS,NH.NA.CC.*L),METTENT.(M.AI).T.E(E)):..

Lors de la consultation du lexique, l'analyseur demande la résolution d'un prédicat (du nom de la catégorie) auquel sont affectées, dans les paramètres, les informations contextuelles disponibles (genre, nombre, traits syntaxiques et sémantiques, etc...). L'unification de ces paramètres avec ceux représentés dans les clauses qui constituent la catégorie sollicitée a pour effet, d'une part, d'interdire l'accès aux éléments qui ne peuvent s'accorder avec les informations transmises et, d'autre part, de permettre la récupération des caractéristiques propres du mot.

CODAGE DES INFORMATIONS LEXICALES CONCERNANT LA SYNTAXE ET LA SEMANTIQUE

Nous affectons, à chaque mot, un ensemble de traits syntactico-sémantiques qui contribuent à sa caractérisation et limitent les contextes de son occurrence par des restrictions d'accords entre les termes reliés syntaxiquement. La nature, le nombre et l'efficacité des traits sont étroitement dépendants de l'univers dans lequel ils opèrent. En conséquence, nous n'avons conservé que les plus généraux, étant entendu que, lors de la définition précise d'un univers, il conviendrait d'y adapter ces informations. Chaque trait est représenté par un identificateur différent suivant qu'il doit être affecté positivement ou négativement; s'il peut recevoir indifféremment l'une ou l'autre valeur, il sera caractérisé par une variable.

Exemple : +HUMAIN, -HUMAIN et $\pm$ HUMAIN seront respectivement codés
H, NH et *H

A travers quelques exemples, nous donnons un aperçu de quelques uns des traits utilisés.

Exemples : NOM(F. *N.NH.CC.NL, (RECOMPENSE.S).(R.EI.K.ON.P.AN).S.E(E)):..
NOM(M.P.NH.AB.NL, TRAVAUX.(T.R.A.V.AU)./$):..

Le premier argument du prédicat "NOM" reflète la liste des traits qui caractérisent le nom codé. Les identificateurs ont la signification suivante :

S → SINGULIER P → PLURIEL F → FEMININ M → MASCULIN

H → + HUMAIN NH → - HUMAIN A → + ANIME NA → - ANIME

CC → + CONCRET AB → - CONCRET L → + LIEU NL → - LIEU

On trouvera dans (GROSS,M., 1977) une étude complète concernant la syntaxe du nom.

Exemples :

ART(*G.P.DEF, LES.L.EI.L(Z)):..
DEM(M.S,CET.S.AI.L(T)):..
QUANTR(C,PLUS.P.L.U.S):..

L'identificateur "C" correspond au trait + COMPTABLE, tandis que "DEF" caractérise + DEFINI

Exemples :

VERBE(*N.NH.NT.NAT,(COULE.NT).(K.OU).L.E(E)):..
VERBE(*N.H.NT.AT.PR(A,H.A.CC.NL),(SOURI.T.ENT).(S.OU.R.I)./$):..
VERBE(*N.*H.T.NAT.NA,(MANGE.NT).(M.AN).J.E(E)):..

Dans le codage des verbes, nous prenons en compte les informations qui concernent le sujet et les divers compléments lorsqu'ils existent. Le premier exemple indique que, pour le verbe représenté, le nombre n'est pas connu ("*N"), le sujet aura impérativement le trait (-HUMAIN), et qu'il ne possède aucun complément. Le second exemple implique un sujet ayant le trait (+HUMAIN), l'absence de complément d'objet direct et la présence d'un complément indirect (introduit par la préposition "A") qui possède les traits (+HUMAIN, +ANIME, +CONCRET, -LIEU). Le troisième exemple suppose un sujet affecté du trait (±HUMAIN) et un complément d'objet direct qui a le trait (-ANIME).

Les multiples informations concernant le verbe sont étudiées de façon systématique par M. GROSS (1968) ainsi que dans (BOONS, J.P. et al., 1976) et (BOONS, J.P. et al., 1977).

DESCRIPTION PHONEMIQUE DES MOTS

Les suites terminales que nous souhaitons analyser peuvent être constituées d'unités telles que des phonèmes, des faisceaux de traits distinctifs, d'indices caractéristiques des traits, de propriétés acoustiques de ces indices, ou bien, d'une combinaison de ces divers éléments. L'information lexicale relative à la prononciation du mot sera codée à l'aide de la séquence des phonèmes qui le caractérisent. Un ensemble de règles doit permettre, à partir de ces unités, de déduire les suites terminales correspondantes. L'exploration de ces termes peut être faite par la suite, de gauche à droite, de droite à gauche ou bien en accédant directement à un phonème particulier du mot (voyelle accentuée dans certains contextes). Cela nous a conduits, d'une part, à factoriser le terme représentant la suite des phonèmes de façon à pouvoir atteindre immédiatement la voyelle accentuée et, d'autre part, à limiter chaque entrée lexicale à une séquence phonémique unique.

Exemples :

NOM(M.S.NH.AB.NL,TRAVAIL.(T.R.A.V.A).Y):..

NOM(M.P.NH.AB.NL,TRAVAUX.(T.R.A.V.AU)./§):..

VERBE(*N.*H.T.NAT.NA,(MANGE.NT).(M.AN).J.E(E)):..

Les mots "travail" et "travaux" ont des entrées lexicales différentes tandis que "mange" et "mangent" sont accessibles simultanément.

Des régles générales doivent permettre de rendre compte de certaines modifications phonologiques telles que l'assimilation consonnantique ou l'harmonie vocalique ; cependant, certains phonèmes sont susceptibles d'apparaître, de disparaître ou d'être modifiés dans certains contextes (liaisons, enchaînements, élisions, etc...) sans que cela obéïsse toujours à des règles valables dans tous les cas. Aussi, nous avons noté de façon particulière, les phonèmes qui peuvent subir de telles altérations.

Exemples:

DEM(*G.P.ND,DES.D.EI.L(Z)):..

ADJ(M.*N,(PETIT.S).(P.E.T.I).L(T)):..

NOM(F.*N.NH.NA.CC.NL,(CHOSE.S).(CH.O).Z.E(E)):..

Lorsqu'un phonème est dominé par l'identificateur "L" (L(Z),L(T)) cela signifie qu'il ne pourra apparaître que dans un contexte favorable de liaison (des amis, petit ami). De même, un phonème dominé par "E" (E(E)) devra disparaître, si l'environnement dans lequel il est placé permet un enchaînement ou une élision (l'ami, petite amie).

REPRESENTATION ORTHOGRAPHIQUE DES MOTS

Afin de donner une transcription écrite de la phrase reconnue, nous avons ajouté la description graphique des mots, aux informations lexicales. Les modifications subies lors des accords en genre et en nombre (pluriels irréguliers) ainsi que celles issues des conjugaisons des verbes, ont été représentées dans la description. Nous avons choisi, pour des raisons d'efficacité de noter toutes ces modifications, bien que seules les irrégularités soient pertinentes.

Exemples :

NOM(F.*N.NH.NA.CC.*L,(EAU.X).(VL.AU)./ß):..

PRON(F.*N,(ELLE.S).VL.AI.L.E(E)):..

VERBE(*N.H.NT.AT.PR(A,H.A.CC.NL),(SOURI.T.ENT).(S.OU.R.I)./ß):..

CONCLUSION

Les informations qui concernent un mot du lexique recouvrent toutes les fonctions qui incombent à ses diverses occurrences dans la phrase. Certaines sont clairement et définitivement connues, tandis que d'autres (notamment celles concernant la sémantique) sont assujetties à l'univers dans lequel on travaille. Le codage de ces paramètres dans le lexique, dépend donc directement de l'usage qui en est fait par le système de traitement des contraintes linguistiques, dans le cadre de la reconnaissance automatique de la parole.

REFERENCES

BOONS,J.P., GUILLET,A., LECLERE,C., 1975, La structure des phrases simples en français. Constructions intransitives; Librairie DROZ, GENEVE-PARIS.

BOONS, J.P., GUILLET,A., LECLERE,C., 1976, La structure des phrases simples en français - Constructions transitives; Rapport de recherches n° 6.

GROSS,M., 1968, Grammaire transformationnelle du français : syntaxe du verbe; Larousse.

GROSS, M., 1977, Grammaire transformationnelle du français : syntaxe du nom; Larousse.

MELONI, H., 1976, PROLOG : mise en route de l'interpréteur et exercices ; Rapport interne du G.I.A. de LUMINY.

MELONI, H., 1978, Système de traitement des contraintes linguistiques en reconnaissance automatique de la parole continue; Rapport interne du G.I.A. de LUMINY.

ROUSSEL, P., 1975, PROLOG : Manuel de référence et d'utilisation.

10èmes JOURNÉES D'ÉTUDE SUR LA PAROLE

GRENOBLE - 30 MAI - 1ER JUIN 1979

GRANDS LEXIQUES ET TRAITEMENTS PHONOLOGIQUES : UNE STRUCTURE DE COMPOSANTE PHONOLOGIQUE ADAPTEE AU TRAITEMENT AUTOMATIQUE.

G. PERENNOU - G. TEP

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

RESUME

Le traitement automatique de la parole, en présence de grands vocabulaires, exige que soient associés aux mots des invariants phonologiques aussi économiques que possible afin de minimiser le volume de stockage nécessaire.

Dans notre projet A.R.I.A.L. nous disposons déjà d'une composante phonologique associée à un lexique opérationnel de 6000 morphèmes (correspondant à 40.000 formes fléchies distinctes au niveau phonémique).

Le but de l'article est de décrire la structure d'une nouvelle composante phonologique séparant mieux les facteurs individuels des lois générales et donnant un statut plus précis à la syllabe.

Le processus phonologique est présenté comme une succession de trois cycles de transformations à l'intérieur desquels on procède de gauche à droite.

Le premier cycle, venant après l'épellation, traite des règles de joncture.

Le second produit une syllabation et permet le traitement des timbres syllabiques.

Le troisième fait intervenir le locuteur. Il permet de traiter des élisions, des liaisons et des enchaînements compte tenu des frontières de niveaux divers introduits par une composante syntaxique.

GRANDS LEXIQUES ET TRAITEMENTS PHONOLOGIQUES : UNE STRUCTURE DE COMPOSANTE PHONOLOGIQUE ADAPTEE AU TRAITEMENT AUTOMATIQUE

G. PERENNOU - G. TEP

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

SUMMARY

In automatic speech recognition using large vocabularies one must associate to the words phonological invariants as thrifty as possible to reduce the volume of the required storage.

In our A.R.I.A.L. project we have already at disposal one phonological component joined to an operational lexicon including 6000 morphemes (corresponding to 40.000 flexed forms distinctive at the phonemic level).

In this paper we intend to describe the structure of a new phonological component, parting better individual factors from general rules and giving to the syllabe a more accurate status.

Phonological process will be presented like a series of three cycles of transformation in which we proceed from left to right.

The first cycle coming after the spelling is dealing with boundaries rules. The second one gives a syllabication and allows the treatment of some morphophonemes. The third cycle makes the locutor interfere. It enables the treatment of "elisions, liaisons et enchaînements" taking into account the boundaries of different levels brought in by a syntactic component.

10èmes JOURNÉES D'ÉTUDE SUR LA PAROLE
GRENOBLE - 30 MAI - 1ER JUIN 1979

GRANDS LEXIQUES ET TRAITEMENTS PHONOLOGIQUES : UNE STRUCTURE DE COMPOSANTE PHONOLOGIQUE ADAPTEE AU TRAITEMENT AUTOMATIQUE

G. PERENNOU
G. TEP

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

INTRODUCTION

Le traitement automatique de la parole comportant de grands vocabulaires impose une composante phonologique abstraite décrivant les transformations permettant de passer du niveau morphologique au niveau phonétique.

La phonologie du français est encore en pleine évolution. Les traitements des frontières, du SCHWA, le statut de la syllabe,... donnent lieu actuellement à des remaniements, et même à des remises en cause profondes (F. DELL 1973,1978, A. SHANE 1968,1978 - B. DE CORNULIER, 1978 - H. BASBOLL, 1978).

Il ne saurait donc être question dans le cadre du traitement automatique de la parole de décrire , de proposer ou de formaliser un système phonologique complet du français.

Notre objectif se limite ici à indiquer les principaux mécanismes prévus en association avec la prochaine version du lexique du projet ARIAL*. Nous pensons que ceux-ci permettront de prendre en charge les propositions que pourront nous faire les linguistes.

Un autre système fonctionne sur le lexique actuel (G. GOUARDERES, 1979), (système plus particulièrement développé par G. PERENNOU et G. TEP et utilisé par G. TEP (1978) pour la reconnaissance au niveau lexical des suites phonémiques en parole continue).

La nouvelle version vise à une meilleure structuration des mécanismes phonologiques indépendants du locuteur. C'est en fait le résultat de mises au point par les auteurs que nous allons maintenant exposer.

LEXIQUE ET MORPHEMES

Le traitement phonologique envisagé comportera en entrée des suites de morphèmes séparés par des frontières de diverses natures :

- \+ concaténation de deux morphèmes dans un mot
- $\#_1$ frontière entre un mot syntaxique et le mot auquel il se rapporte (ou dans une expression figée)
- $\#_2$ frontière entre adjectif et le nom auquel il se rapporte, ou jugée équivalente,
- $\#_3$ frontière verbe-complément ou équivalentes
- $\#_4$ frontière de conjonction et de syntagme ou jugée équivalente.
- $\#_5$ ponctuation.

* Analyse de Reconnaissance des Informations Acoustiques, Linguistiques.

Nous supposons que ces frontières ont été indiquées convenablement au niveau syntaxique. A noter que nous nous écartons ici de la tendance qui est de prévoir seulement deux types de frontières # et ## (cf. A. SHANE ou F. DELL), ce qui ne permet pas de traiter les liaisons et les enchaînements.

Exemple :

le #$_1$ enfant +Sg #$_4$ écout + impf + 3p Sg #$_3$ en #$_1$ silence #$_4$ et #$_4$ avec #$_1$ ...

Pour plus de clarté nous avons désigné une entrée lexicale par la représentation graphique correspondante. En réalité ces termes renvoient à des descriptions phonologiques se trouvant dans le lexique et qui se substituent aux entrées avec effacement de +.

Exemple :

lə #$_1$ ɑ̃ fɑ̃,t #$_4$ ekutɛ,t #$_3$ ɑ̃,n #$_1$ silɑ̃sə #$_4$ e #$_4$ avɛk #$_1$ atɑ̃siɔ̃,n

Ce sont les chaînes de ce type qui font l'objet du traitement phonologique

LA REPRESENTATION MORPHONOLOGIQUE

Les morphèmes peuvent correspondre à plusieurs morphes (ex : "aller" correspond à v-, al-, -). Chacun de ceux-ci est utilisé suivant le contexte. Nous nous intéressons simplement ici à la représentation d'un morphe qui sera un mot écrit avec l'alphabet morphonologique suivant :

$\mathcal{M} = \mathcal{P} \cup \mathcal{C} \cup \mathcal{V} \cup \mathcal{O}$

$\mathcal{P}$ = symboles A.P.I. désignant des phonèmes,
$\mathcal{C}$ = S F T K ...
$\mathcal{V}$ = O Ö E A ...
$\mathcal{O}$ = , . ~ ≃ * ±

Cette représentation d'un mot "m" se notera $\Gamma(m)$.

Exemple :	"rouge"	s'écrira	ruʒə
	"petit"	-	pətI,t
	"moyen"	-	mwajɛ~n
	"brun"	-	bry≃n
	"beau"	-	bɛ*l
	"cheval"	-	ʃəva±l
	"appel_	-	ap.El

PREMIER CYCLE DE TRAITEMENT

Nous utilisons une terminologie standard pour la description des règles de réécriture. Ajoutons que :

; est un 'ou' disjonctif Λ est la chaîne vide
V → est une voyelle A.P.I. ; O ; Ö ; E ; A ...
V' → V ; j ; ɥ ; w
C → une consonne A.P.I. ; F ; S ; T ; K ...
C' → C ; j ; ɥ ; w ;

N → m ; n ; ɲ ; ŋ
Ṽ → ɑ̃ ; ɛ̃ ; œ̃ ; ɔ̃

Principe général : les règles qui peuvent être appliquées le sont avec priorité de gauche à droite.

Traitement des consonnes latentes

$$,C \longrightarrow \begin{cases} \Lambda / \_\_ \#_i \ (i \geq 4) \ ; \ \_\_ \#_i C \ (\forall_i) \ ; \ \_\_ ,C \\ C / \_\_ V \\ C- \text{ sinon} \end{cases} \qquad (L)$$

Exemple d'emploi

Supposons que l'on ait :
Γ(fem) = ə Γ(plur)=,z Γ(petit) = pəti,t
Γ(grand) = grɑ̃,T

La représentation de "petit ami" est pəti,t #₂ ami qui après application de (L) devient : pətit#₂ami

De même :
"grand ami" ⟶ grɑ̃,T#₂ami $\xrightarrow{(L)}$ grɑ̃T_#₂ami
"petit+fem" ⟶ pəti,tə $\xrightarrow{(L)}$ pətitə
"grand+fem" ⟶ grɑ̃,Tə ⟶ grɑ̃Tə
"petit+plur chat+plur" ⟶ pəti,t,z #₂ ʃa,t,z #₅ ⟶ pəti#₂ʃa#₅
"petit+plur ami+plur" ⟶ pəti,t,z #₂ ami,z #₅ ⟶ pəti,z_#₂ami#₅
(à noter que si l'on opère de droite à gauche on obtient pəti,t #₂ ami #₅)
"petit+fem+plur ami+fem+plur" ⟶ pəti,tə,z #₂ amiə,z #₄ ⟶ pətitəz_#₂ amiə#₄

Sonorisation des consonnes de jonction

On rend compte ici des alternances telles que neuf-neuve.

$$\begin{bmatrix} F \\ S \\ T \\ K \end{bmatrix} \longrightarrow \begin{bmatrix} v \\ z \\ d \\ g \end{bmatrix} \quad / \ \_\_V' \qquad (S)$$

T ⟶ t sinon

Exemples d'emploi :
grɑ̃Tə $\xrightarrow{(S)}$ grɑ̃də (grande)
grɑ̃T_#₂ami $\xrightarrow{(S)}$ grɑ̃t_#₂ami

Dénasalisation

Les terminaisons de morphes en consonnes nasales, donnent lieu à deux types de transformation : le premier est annoncé par ~ et permet de traiter les alternances telles que bon-bonne et le second de celles du type commun-commune.

$$V\sim N \longrightarrow \begin{cases} \tilde{V} & /\_\_\_ \#_i (i\geq 3)\ ;\ \_\_\_ \#_i C (\forall i)\ ;\ \_\_\_\_, C \\ VN & /\_\_\_ V' \\ VN- & \text{sinon} \end{cases} \qquad \text{(DN1)}$$

$$V\simeq N \longrightarrow \begin{cases} V^{\sim}N- & /\_\_\_\_ \#_i V' (i\leq 2) \\ VN & /\_\_\_\_ V' \\ V^{\sim} & \text{sinon} \end{cases} \qquad \text{(DN2)}$$

$$y^{\sim} \rightarrow \tilde{\text{œ}}\ ;\ i^{\sim} \rightarrow \tilde{\varepsilon}\ ;\ \varepsilon^{\sim} \rightarrow \tilde{\varepsilon}\ ;\ a^{\sim} \rightarrow \tilde{a}\ ;\ \text{ɔ}^{\sim} \rightarrow \tilde{\text{ɔ}} \qquad \text{(DN3)}$$

Exemples : "ancien"

$\tilde{a}sj\varepsilon\sim n\#_2 ami \xrightarrow{(DN1)} ansj\varepsilon n\_\#_2 ami$ (ancien ami)

$\tilde{a}sj\varepsilon\sim n\text{ə} \xrightarrow{(DN1)} \tilde{a}sj\varepsilon n\text{ə}$

$\tilde{a}sj\varepsilon\sim n\#_2 li,t\#_5 \longrightarrow \tilde{a}sj\tilde{\varepsilon}\#_2 li\#_5$

"commun"

$komy\simeq n\#_2 akɔr,T\#_5 \xrightarrow{(DN2)} komy^{\sim}n\_\#_2 akɔr\#_5 \xrightarrow{(DN3)} kom\tilde{\text{œ}}n\_\#akɔr\#_5$

Autres alternances aux jonctures

On peut vouloir traiter au moyen de règles les alternances telles que beau-bel plutôt que par polymorphie. La règle suivante rend compte des phénomènes de ce type qui affectent quelques mots du français.

$$V*C' \longrightarrow \begin{cases} VC' & /\_\_\_\_ V'\ ;\ \_\_\_\_ \#_i V (i\leq 2) \\ V* & \text{sinon} \end{cases} \qquad \text{(M1)}$$

$$V\underline{*}C' \longrightarrow \begin{cases} V^{*} & /\_\_\_\_, Z \\ VC' & \text{sinon} \end{cases} \qquad \text{(M2)}$$

$$\varepsilon^{*} \longrightarrow \begin{cases} \emptyset & /j\_ \\ 0 & \text{sinon} \end{cases} \qquad \text{(M3)}$$

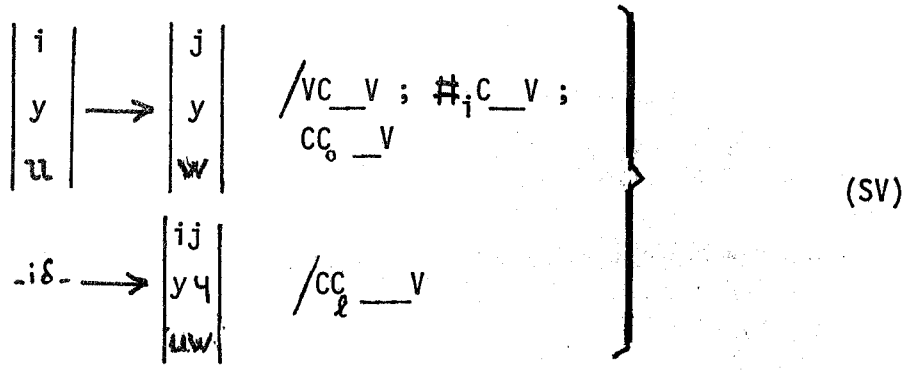
$\text{ɔ}^{*} \longrightarrow u\ ;\quad a^{*} \longrightarrow o$

Exemples

$b\varepsilon*l,z\#_5 \longrightarrow b\varepsilon^{*},z\#_5 \longrightarrow bO,z\#_5 \longrightarrow bO\#_5$ (beau)

Traitement des voyelles fermées

Les voyelles fermées donnent lieu à l'insertion ou la substitution de semi-voyelles conformément à la règle suivante :



où : C_0 désigne une consonne occlusive et C_ℓ une consonne liquide.

Exemples: "épier" ⟶ epje,r
"skier" ⟶ skje,r
"strier" ⟶ strije,r
"peuplier" ⟶ pøplije,r

DEUXIEME CYCLE DE TRANSFORMATION : SYLLABATION ET TIMBRES SYLLABIQUES

Le deuxième cycle de transformation est le premier niveau de syllabation qui doit permettre le traitement des phonèmes dont le timbre dépend de la nature de la syllabe.

Cette première syllabation met en lumière un certain nombre de problèmes aux frontières de mots qui motiveront les traitements d'élision, de liaison et d'enchaînement. Des impératifs d'économie et de rythme interviendront aussi pour décider de l'élision interne du ə dont dépend également le statut de certaines syllabes.

Voici maintenant l'algorithme de syllabation utilisé à ce niveau (G. PERENNOU, 1978 - 2).

(i) un niveau 4 est attribué aux voyelles et élément de V
" 3 " aux semi-voyelles
" 2 " aux liquides
" 1 " aux autres consonnes et aux éléments de C,
Il est fait abstraction des ə qui n'ont donc pas de niveau.

(ii) une frontière est placée entre deux voyelles consécutives,

(iv) si un phonème de niveau i est dans le contexte de niveau k- ℓ, on place une frontière devant ce phonème quand $k \geq i < \ell$.

(v) Une frontière est placée après chaque #

Exemple : $\#_5 \mid yn\partial \#_1 \mid s \mid tri \mid jyr\partial \#_2 \mid ak \mid si \mid d\tilde{a} \mid tel\partial \#_5 \mid$
4 1 1 1 2 4 3 4 2 4 1 1 4 1 4 1 4 2

(vi) les syllabes sans voyelle ni ə qui peuvent apparaître en début de mot sont éliminées par suppression de la frontière de droite.

Exemple :

$\#\mid s \mid tri \mid jyr\partial \# \longrightarrow \#\mid stri \mid jyr\partial \#$

Il est maintenant possible d'utiliser les règles qui donnent leur timbre à certains phonèmes. Par exemple :

E ⟶	ɛ	en syllabe accentuée fermée	(T1)
0 ⟶	ɔ	en syllabe fermée par r et l	(T2)
Ö ⟶	œ	en syllabe fermée	(T3)
	ø	en finale fermée par z ou en finale ouverte	
Œ ⟶	ɛ	en syllabe fermée	(T4)
	ə	en syllabe ouverte	
A ⟶	ɛ	en syllabe ouverte	(T5)
	a	en syllabe fermée	

Exemples :

pÖr ⟶ pœr ; pÖrÖ ⟶ pørø ou pœrø (peur-peureux)
mAr ⟶ mɛr ; mAreə ⟶ mareə (mer-marée)

Il est bien clair que les traitements de cet ordre sont en général dépendants du locuteur et du dialecte. Les règles ci-dessus ont cependant un caractère quasi-obligatoires. Elles doivent être complétées selon le cas par des règles moins certaines en général mais qui peuvent dans chaque cas particulier devenir presque systématique. Nous renvoyons aux traités spécialisés pour cette question (par exemple : P. FOUCHE (1969), PR. LEON (1966)).

TROISIEME CYCLE DE TRANSFORMATIONS : ELISIONS ET ENCHAINEMENTS

La syllabation précédente fait apparaître des pseudo-syllabes n'ayant pour toute voyelle qu'un "ə". D'autres ont deux voyelles à cause de ce même "ə". Or une syllabe française doit comporter une voyelle et une seule.

Il y aura donc une tendance générale :

élider "ə" ou le réaliser comme Ö selon les besoins de la syllabation.

Exemple : on réalisera en général :

#|yn#|stri| jyr#|ak|si| dã|tɛl#| (ə → ∧)
#|as| trÖ#|dÖ#|La #|nyi #| (astre de la nuit) (ə → Ö)

Dans le même esprit on verra apparaître un "ə" dans #|ur|sÖ#|blã (ours blanc) là où la graphie ne met pas de "e".

En dehors de cette tendance il y aura d'autres facteurs provoquant des modifications syllabiques :

1. Complexité des syllabes
2. Liaison des mots
   - par liaison C- # V
   - par enchaînement C#V
3. Les élisions se produisant dans les séquences C(C)(ə)#C
4. Les nécessités du temps et l'affaiblissement de la pénultième.

Dans le cadre de cet article et compte tenu des objectifs indiqués en introduction nous ne pouvons traiter ces problèmes faisant intervenir des facteurs individuels.

Nous renvoyons à (FOUCHE, 1969 et P.R. LEON, 1971) pour une étude plus détaillée de ces problèmes.

Bornons nous à montrer que la chaîne fournie peut servir de point de départ pour un autre cycle de transformations telles que : élisions, liaisons et enchaînements. Considérons par exemple un locuteur qui applique les règles simples suivantes :

$$\left.\begin{array}{l} C\_ \longrightarrow \wedge \quad / \_\_ \#_i \quad (i \geqslant 3) \\ C\_\#_i| \longrightarrow |C\_\#_i \end{array}\right\} \quad (L')$$

$$\left.\begin{array}{l} C_1C_2'(ə)\#_i| \longrightarrow |C_1C_2'\#_i \;/\; \_\_ V' \text{ et } i \neq 5 \text{ et } niv C_1 \geqslant niv C_2 \\ C'(ə)\#_i| \longrightarrow |C'\#_i \;/\; \_\_ V' \end{array}\right\} \quad (E)$$

$$\left.\begin{array}{l} C_1C_2(ə) \longrightarrow C_1|C_2Ö \;/\; \text{en pénultième si } niv C_1 \geqslant niv C_2 \\ \text{—}iõ\text{—} \longrightarrow |C_1C_2Ö \;/\; niv C_1 < niv C_2 \end{array}\right\} \quad (I)$$

$$ə \longrightarrow \wedge \quad (EL1)$$

Le locuteur appliquant les règles avec priorité de gauche à droite et dans l'ordre indiqué prononcera alors :

porte-plum'

port'-manteau

* les notr' aussi (notrəz-$\#_3$| $\xrightarrow{(L')}$ notrə$\#_3$|

$\xrightarrow{(E)}$ no|tr $\#_3$)

etc...

CONCLUSION

Nous pouvons maintenant préciser au moyen du diagramme suivant la structure de la composante phonologique abstraite retenue en association avec le lexique.

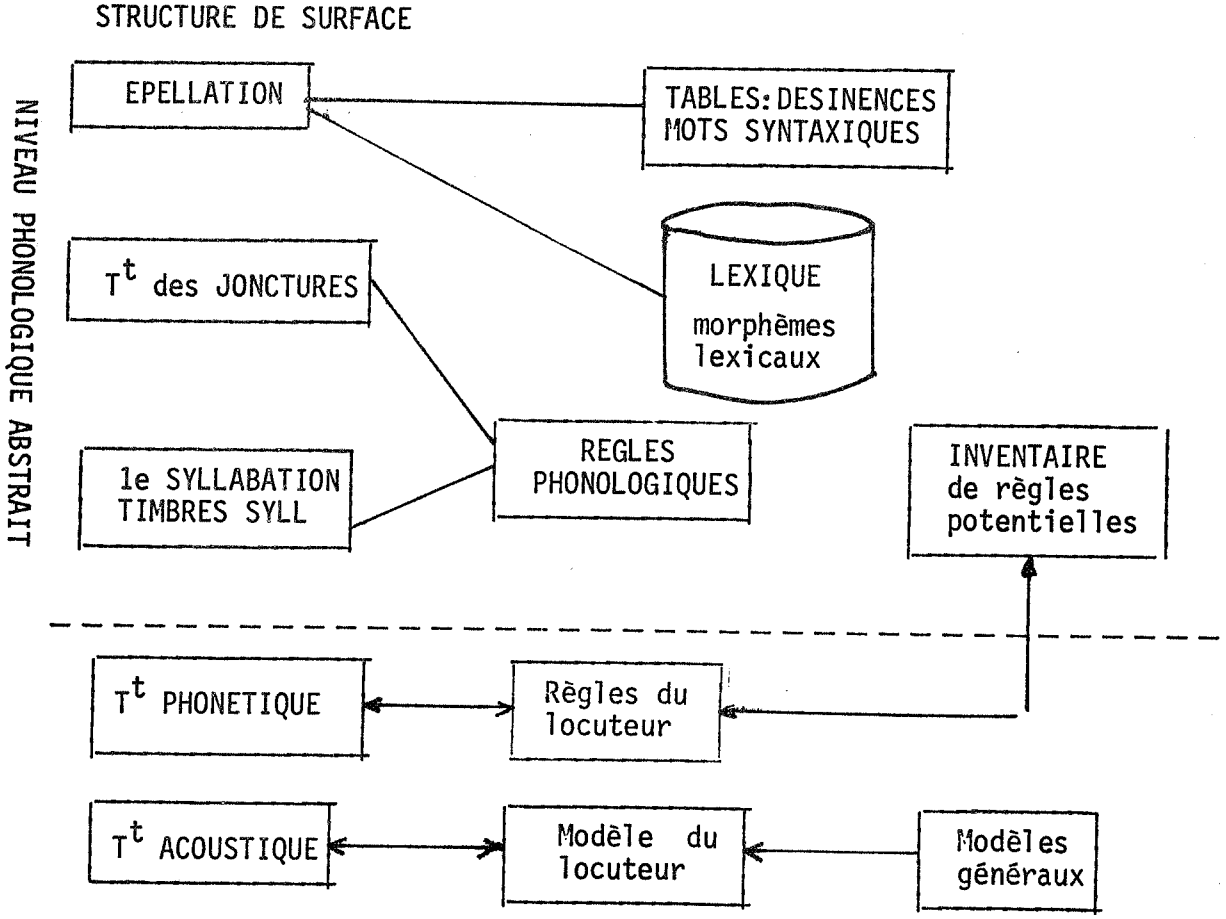


Diagramme de la structure du système phonologique et de sa situation par rapport aux niveaux phonétique et acoustique.

Cette composante est en cours d'implantation mais elle s'appuie sur l'acquis de la première version déjà opérationnelle et qui a permis, d'une part de proposer un modèle de génération de chaînes de phonèmes en vue de la synthèse de la parole,
d'autre part de construire un analyseur lexical de chaînes de phonèmes.

Sa contribution sera de réduire dans des proportions considérables le volume du lexique et de rendre ainsi possibles des traitements en mémoire centrale de l'ordinateur qui ne le seraient pas avec un vocabulaire de quelques milliers de mots représentés avec toutes leurs variantes et flexions (gain d'un facteur 10 au niveau abstrait et d'un facteur supérieur au niveau phonétique - soit un gain total d'un facteur de quelques centaines).

Dans cette composante apparaissent des "cadres" qui ne pourront être complétés qu'avec l'aide des phonéticiens et linguistes. Nous pensons qu'il y a là une collaboration fructueuse à poursuivre.

REFERENCES

[1] Etude de phonologie française. Ed. C.N.R.S.. 1978.

BASBOLL, H., 1978,Boundaries and ranking rules in french phonology. Dans [1]

CORNULLIER de, B., 1978, Syllabe et suite de phonèmes en phonologie du français. dans [1]

DELATTRE, P., 1958, Studies in French and comparative phonetics. Mouton et Cie, 1966.

DELL, F., 1973, Les règles et les sons, Paris, Hermann.

DELL, F., 1978, Epenthèse et effacement de SCHWA dans des syllabes contiguës en français. Dans [1]

GOUARDERES, G., 1977, Organisation d'un lexique en vue de l'analyse de la parole en continu.

GOUARDERES, G., Les lexiques automatiques pour le traitement de la parole : version 3 du lexique du projet A.R.I.A.L.

FOUCHE, P., 1969, Prononciation française. Ed. Klincksieck.

LEON, PR., 1966, Prononciation du français standard. Didier.

LEON, PR., 1971, Essai de phonostylistique. Didier.

PERENNOU, G., 1978, Note interne C.E.R.F.I.A. sur la syllabation.

PERENNOU, G., 1978, Note interne C.E.R.F.I.A. sur la composante phonologique.

SHANE, A., 1968, French phonology and morphology. Cambridge. Mass, MIT Press.

SHANE, A., 1978, L'emploi des frontières de mots en français. Dans [1]

TEP, G., 1978, Thèse de 3ème cycle. Contribution à l'étude phonologique d'un système d'analyse et de synthèse de la parole continue.

10èmes JOURNÉES D'ÉTUDE SUR LA PAROLE

GRENOBLE - 30 MAI - 1ER JUIN 1979

LE NIVEAU LEXICAL DANS LE SYSTEME MYRTILLE II : REPRESENTATION DU LEXIQUE ET TRAITEMENTS ASSOCIES

Jean-Marie PIERREL, Jean-François MARI, Jean-Paul HATON
Equipe de Traitement du Signal et Reconnaissance de Formes
Centre de Recherche en Informatique de Nancy
Université de Nancy I
C.O. 140 - 54037 - NANCY CEDEX

RESUME

Le système MYRTILLE II est un système de compréhension du dialogue parlé en cours de réalisation au CRIN. Nous présentons ici les divers aspects du niveau lexical de traitement :

- représentation du lexique : elle comporte en fait deux aspects. L'un concerne la représentation de la transcription phonétique des racines des mots ; les altérations phonologiques sont traitées sous forme de procédures et ne sont donc pas précompilées. Le deuxième aspect concerne la représentation des fonctions syntaxiques et sémantiques des mots.
- techniques de recherche lexicale qui sont de deux types dans MYRTILLE II :
  . vérification de mots (vérification des hypothèses émises à d'autres niveaux) ;
  . localisation d'un mot dans le continuum de parole.

L'article présentera divers exemples pratiques liés aux points précédents.

THE LEXICAL PROCESSING LEVEL IN MYRTILLE II SPEECH UNDERSTANDING SYSTEM

Jean-Marie PIERREL, Jean-François MARI, Jean-Paul HATON

SUMMARY

MYRTILLE II is a speech Understanding System for pseudo-natural language which is under development in our laboratory. In this paper, we present the various aspects of the lexical processing level :

- representation of the lexicon : the lexicon contains the basic phonetic transcription of the words ; the various acoustic forms of these words are produced by phonological rules which are stored under the form of procedures. Syntactic and semantic pieces of information are also stored in the lexicon and are used during the emission of hypotheses.

- word verification and word spotting techniques : we have developed some methods based on dynamic programming that we present together with practical examples.

10èmes JOURNÉES D'ÉTUDE SUR LA PAROLE
GRENOBLE - 30 MAI - 1ER JUIN 1979

LE NIVEAU LEXICAL DANS LE SYSTEME MYRTILLE II : REPRESENTATION DU LEXIQUE ET TRAITEMENTS ASSOCIES

Jean-Marie PIERREL, Jean-François MARI, Jean-Paul HATON

1. INTRODUCTION

La compréhension du discours continu nécessite la prise en compte d'informations de nature très diverse, depuis l'acoustique jusqu'à la sémantique. Le niveau lexical, relatif aux mots, occupe une place privilégiée au confluent des informations acoustiques et phonétiques et des informations "linguistiques" (syntaxe, sémantique, pragmatique), car le passage au niveau du mot est une étape obligée du processus de compréhension du langage. Dans le cas de la parole, le niveau lexical a des caractères spécifiques d'une part à cause de l'indéterminisme inhérent à la localisation des mots dans le semi-continuum sonore et d'autre part, du fait qu'une même unité lexicale peut se réaliser acoustiquement de nombreuses façons, en majeure partie décrite par la phonologie qui apparaît ainsi comme une composante importante d'un lexique de parole.

Dans cet article, nous présentons les divers problèmes posés par la réalisation et l'utilisation d'un lexique, à la lumière des travaux menés dans le cadre du projet MYRTILLE II en cours de réalisation dans notre laboratoire. Le système MYRTILLE II est destiné à comprendre des phrases d'un langage proche du français parlé, dans des univers d'applications nécessitant un vocabulaire de plus de 300 mots. Le système fonctionne suivant le principe bien connu d'Hypothèse-et-Test, généralisé à tous les niveaux de traitement. Pour des langages pseudo-naturels du type de celui que nous traitons, il est impossible de dissocier syntaxe et sémantique. Nous avons ainsi défini une structure de réseau à noeuds procéduraux qui regroupe les deux types d'informations, complétée par un lexique fournissant les informations complémentaires concernant les noeuds.

Dans ces conditions, le rôle du niveau lexical est triple :
- émission des hypothèses,
- test des mots émis comme hypothèses,
- détermination du point de départ de l'analyse d'une phrase lorsque celle-ci n'a pas lieu systématiquement à partir du début.

Dans une première partie, nous décrivons la structure arborescente du lexique et les diverses informations (sémantiques, syntaxiques, phonologiques, prosodiques et phonétiques) que l'on y trouve.

Ensuite, nous présentons les techniques de recherche lexicale mises en oeuvre pour la localisation et la reconnaissance de mots.

2. STRUCTURE DU LEXIQUE ET EMISSION DES HYPOTHESES

Dans le cadre général d'un principe d'hypothèse et Test, les informations

d'ordre lexical interviennent pour une bonne part lors de l'émission des hypothèses et rendent compte d'informations de type syntaxique, sémantique, prosodique ou phonémique.

2.1. Niveau syntaxico-sémantique

Dans le système MYRTILLE II [11], [12], nous avons opté pour une définition hiérarchisée des traits syntaxico-sémantiques liés au lexique. Nous obtenons ainsi une représentation arborescente du vocabulaire bien adaptée au processus d'Hypothèse et Test et décrivant les diverses dépendances syntaxico-sémantiques liées à chaque mot. Voici rapidement présentés les quatre niveaux d'arborescence que nous utilisons :

(i) le premier niveau provoque une partition en trois grandes classes :

- les mots de liaison qui apparaissent de façon explicite dans la définition du langage,
- les mots qui possèdent des fonctions grammaticales particulières : prépositions, adverbes, conjonctions, etc...,
- les terminaux généraux qui regroupent noms, verbes et adjectifs. C'est pour cette classe de mots que nous donnerons des exemples dans la présentation des 3 autres niveaux.

(ii) le second niveau provoque une partition suivant les grandes classes grammaticales : nom, verbe, adjectif, etc...

(iii) le troisième niveau distingue des sous-classes grammaticales en fonction de la construction grammaticale et des dépendances syntaxico-sémantiques qu'elles demandent.

A titre d'exemple, voici les informations fournies par ce niveau dans le cas des terminaux généraux de type NOM. Pour chaque sous-classe, outre des pointeurs vers les autres niveaux de l'arborescence, on y trouve :
- la liste des types de noms (sous-classes) pouvant être complément des noms de cette sous-classe,
- la liste des types d'adjectifs (sous-classes) pouvant qualifier les noms de cette sous-classe,
- la liste des fonctions syntaxiques liées à cette sous-classe.

Pour les terminaux généraux de type verbe, on y trouvera des informations sur les types possibles pour le sujet et les compléments.

(iv) le quatrième niveau correspond aux différents mots ou entités et détaille les propriétés propres à chacun. Ainsi, pour les verbes, on y trouvera, outre des renseignements phonétiques et phonologiques, la liste des types de constructions possibles pour la suite verbale et le type des auxiliaires acceptés par ce verbe.

Les traitements associés à ce niveau de description syntaxico-sémantique du lexique consistent essentiellement à restreindre le nombre d'hypothèses de type nom, verbe et adjectif afin de provoquer le moins de demandes possibles au niveau de reconnaissance phonémique. Ces traitements sont de deux ordres, suivant les informations fournies par la structure syntaxico-sémantique de la phrase.

a) lorsque la structure syntaxico-sémantique a préalablement émis des hypothèses, il s'agit de les restreindre compte tenu des mots déjà reconnus et des dépendances syntaxico-sémantiques qu'ils demandent. Ainsi, si on recherche un complément du nom après avoir reconnu le mot "tempête", on sélectionnera les hypothèses telles que "pluie", "neige", ... et on rejètera "soleil", "température", ... car on peut parler de "tempête de neige" mais non de "tempête de soleil" !

b) lorsque la structure syntaxico-sémantique conduit à une impasse, il s'agit de fournir un point de reprise sur les seuls critères de liaisons syntaxico-sémantiques entre les mots en déterminant les mots pouvant "dominer" les grands composants de la phrase non encore reconnus : verbe pour le groupe verbal, nom ou pronom pour le groupe sujet, nom ou verbe pour la suite du groupe verbal.

2.2. Niveau prosodique et phonémique

Outre les informations syntaxico-sémantiques et la représentation phonémique des mots (cf. § suivant), nous avons choisi d'adjoindre dans le lexique diverses informations qui doivent permettre de restreindre aussi les hypothèses émises sur les mots à reconnaître. Ce sont :

(i) la longueur phonémique qui, compte tenu de la plage de phonèmes à traiter et/ou des marqueurs de segmentation prosodique peut conduire à écarter certaines hypothèses.

(ii) des patrons phonémiques (par exemple pour le mot "petit" "p * t *") qui, par tests comparatifs de présence ou non de plosives ou fricatives, peuvent infirmer ou confirmer certaines hypothèses dans une plage phonémique donnée (filtrage).

(iii) des patrons prosodiques enfin qui eux aussi, par comparaison aux informations prosodiques obtenues sur le signal, peuvent confirmer plus ou moins telle ou telle hypothèse.

2.3. Représentation phonétique des mots

Aux quatre niveaux d'arborescence décrits au paragraphe 2.1., il faut enfin ajouter un cinquième niveau correspondant à la représentation phonétique de chaque mot que nous fournissons sous forme de transcription phonémique des racines de mots et de règles phonologiques.

a) transcription phonémique des racines des mots : Elle est représentée sous forme d'un graphe permettant de rendre compte des phénomènes de substitution, insertion ou omission de phonèmes correspondant soit à des prononciations différentes des mots soit à des erreurs provoquées par les limites du système acoustico-phonétique.

b) informations phonologiques : Elles sont représentées par des procédures et non pas précompilées dans le système. Les principales altérations phonologiques prises en compte sont :

- les accords en genre et nombre des noms et adjectifs,
- les conjuqaisons des verbes,
- les liaisons,
- les fusions ou assimilations de phonèmes compte tenu du contexte.

3. TECHNIQUES DE RECHERCHE LEXICALE

3.1. Introduction

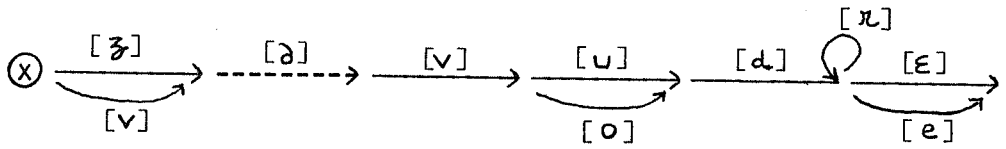
Notre travail de recherche dans le domaine de la localisation et de la vérification de mots s'est jusqu'à présent effectué indépendamment du travail de conception du système MYRTILLE II. La représentation des formes phonétiques de référence et la façon d'y accéder sont celles utilisées dans MYRTILLE I [2], mais nous verrons qu'il n'y a pas d'incompatibilités entre les deux représentations et que les algorithmes de recherche lexicale peuvent être utilisés dans le niveau morphologique du système MYRTILLE II.

Nous allons décrire tout d'abord la représentation phonétique des formes de référence du dictionnaire avant de décrire plus précisément les algorithmes de recherche.

Représentation des formes de référence :

Celle-ci est celle définie par J.P. HATON [1] et utilisée dans le système MYRTILLE I [2].

Une forme est représentée par un graphe (figure I) qui permet de tenir compte de certaines altérations phonologiques et des erreurs courantes du système acoustique.



Représentation lexicale de : je voudrais

Figure I

----- : élision
: substitution
: insertion

Le langage de programmation utilisé (FORTRAN IV) nous a conduit à représenter ce graphe par un tableau TER(LMØT,3).

TER(I,1) représente le ième phonème
si TER(I,1) est négatif, cela signifie que ce phonème peut s'élider
TER(I,2) représente le phonème pouvant se substituer à TER(I,1)
TER(I,3) représente l'éventuel phonème s'insérant avant TER(I,1).

Les notions de racines de mots et de désinence n'apparaissent pas dans cette représentation mais on peut considérer que celle-ci est le résultat de procédures de construction de forme de référence à partir de données du lexique et des niveaux supérieurs du système.

3.2. Localisation de mots

3.2.1. Etude bibliographique

Le problème de la localisation de mots (en anglais word-spotting) a reçu beaucoup moins d'attention que celui de la comparaison de formes de durée sensiblement égale, il peut être vu comme un problème d'extraction de sous-chaîne et différents chercheurs l'ont résolu de la façon suivante :

- R. VIVES [6] du CNET travaille sur des chaînes de phonèmes et extrait

toutes les sous-chaînes communes à la forme de référence et à la chaîne d'entrée. Ceci lui permet de localiser approximativement les frontières de mots puis de comparer les deux formes mises en correspondance [7].

- G. PERENNOU et al. du CERFIA de Toulouse [8] travaillent sur des représentations syllabiques. L'algorithme utilisé parcourt la chaîne syllabique d'entrée de gauche à droite et grâce à une fonction d'adressage dans le lexique détermine les éventuels mots débutant par une paire de syllabes constituant une clé donnée. La comparaison ne porte plus que sur les syllabes restantes. Lors de la constitution du treillis de mots, des phénomènes de liaison et d'élision sont pris en compte.

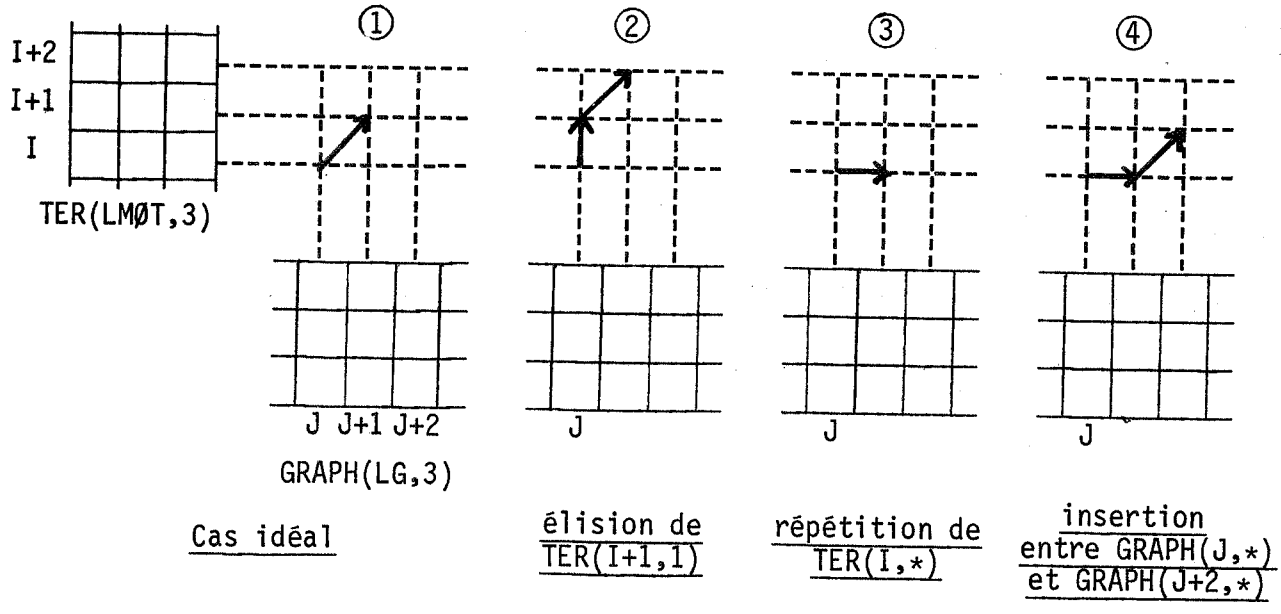
- Il existe aussi une approche par programmation dynamique ; on peut se reporter aux travaux de [9, 10] ainsi qu'aux critiques de cette approche données par [4].

- Une méthode heuristique fondée sur un principe de similitude locale au voisinage de chaque paire de points mis en correspondance a été développée par [4] et [5]. Elle a été utilisée pour la reconnaissance de mots clés dans une représentation du signal sonore sous forme d'une suite d'échantillons de sortie de vocodeurs ou de coefficients d'autocorrélation ; nous l'avons généralisée aux treillis de phonèmes.

3.2.2. Extraction de sous-chaînes à partir d'un treillis phonémique fondée sur un principe de similitude locale

Le problème d'extraction de sous-chaîne peut être vu comme un problème de construction de chemins dans le produit cartésien (forme de référence x treillis de phonèmes d'entrée) comme le montre la figure II.

La forme de référence et le treillis d'entrée sont représentés respectivement par les tableaux TER(LMØT,3) et GRAPH(LG,3). Il existe une relation entre l'allure d'un chemin au voisinage d'un point et l'altération phonologique dans la chaîne d'entrée. En effet, les différents phénomènes de substitution, d'insertion et d'élision à l'intérieur d'un mot se traduisent par les enchaînements suivants de transitions :



Pour autoriser des altérations non prévues dans le lexique, il faut savoir apprécier la qualité des correspondances des dernières paires de phonèmes construites.

En chaque point du réseau, nous définissons une similitude instantanée entre les phonèmes mis en correspondance :

$$V(i,j) = 1 \Leftrightarrow GRAPH(J,*) = TER(I,*)$$

A chaque chemin partiel s'achevant en (i,j), nous associons un score dans lequel les trois dernières similitudes instantanées sont prépondérantes :

$$S(i,j) = 0{,}6\ S(i-a,\ j-b) + 0{,}4\ V(i,j)\ K(a,b)$$

$$S(1,j) = V(1,j)$$

(a,b) représente une transition élémentaire du type (1,0), (0,1), (1,1) et K(a,b) un facteur de pénalisation inférieur à 1, on a :

$$K(1,1) = 1\ ;\ 0 < K(0,1) < 1\ ;\ 0 < K(1,0) < 1.$$

Ce facteur sert à pénaliser les transitions parallèles aux axes.

Appelons Vn, Kn et Sn les valeurs respectives de la similitude instantanée, du facteur de pénalisation et de la similitude locale à la nième étape de construction d'un chemin. Nous avons :

$$Sn = 0{,}216\ Sn\text{-}2 + 0{,}144\ Vn\text{-}2\ Kn\text{-}2 + 0{,}240\ Vn\text{-}1\ Kn\text{-}1 + 0{,}4\ Vn\ Kn.$$

La recherche d'un mot à l'intérieur d'un treillis phonétique se fait en partant du principe que la similitude locale du chemin correspondant est supérieure en chaque point à un seuil donné, on possède ainsi un critère d'abandon de chemin.

Après un certain nombre d'essais, il s'est avéré que cette contrainte de similitude locale était insuffisante pour localiser un mot dans sa globalité. Nous avons ajouté une contrainte relative à la pente de chaque chemin en l'obligeant à rester entre deux valeurs extrêmes, cela signifie que la variation de la vitesse de prononciation d'un mot est aussi bornée.

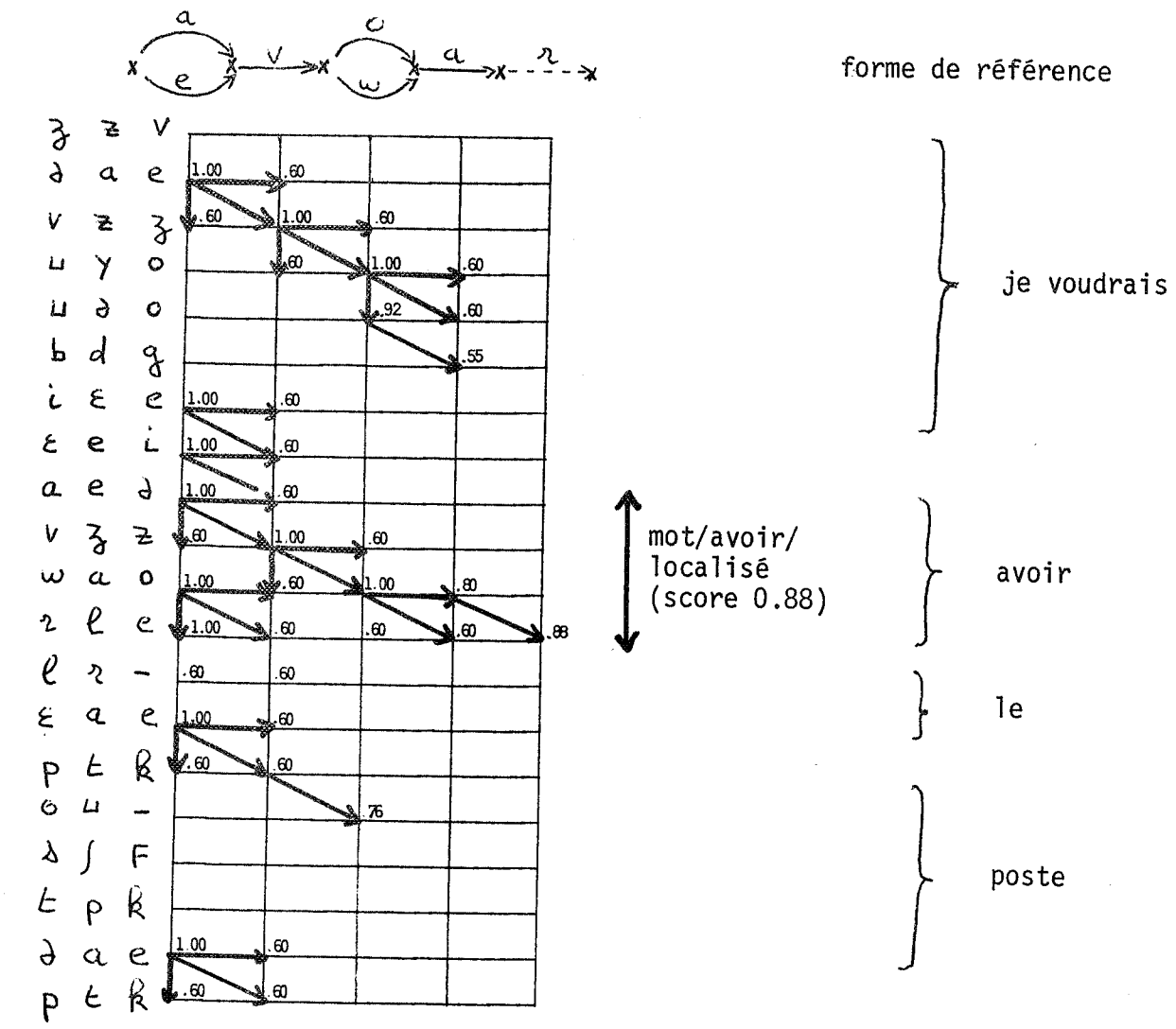
Remarques :

L'obtention du treillis d'entrée a été simulée à l'aide des travaux de Jean-Paul HATON et Jean-Marie PIERREL (1,2).

Nous avions donné aux différents paramètres les valeurs suivantes :

$$K(1,0) = 0.5,\ K(0,1) = 0.8,\ \text{seuil d'abandon} = 0.5.$$

Exemple :



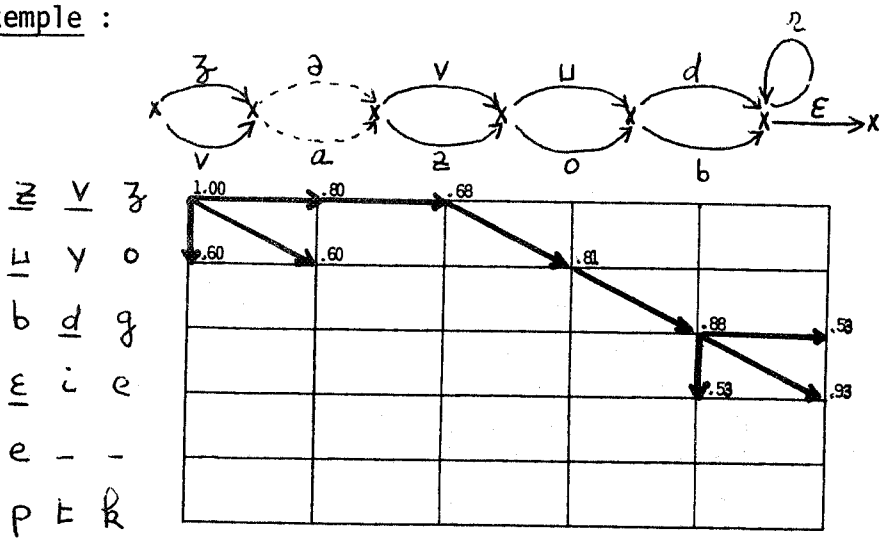
Recherche du mot /avoir/ dans la phrase :
je voudrais avoir le poste 340

Figure II

3.3. Vérification de mots

Si on restreint la zone du treillis d'entrée dans laquelle s'effectue la recherche, l'algorithme peut servir de vérificateur d'hypothèses émises par les niveaux syntaxico-sémantiques en un point donné du treillis. La longueur de la zone de recherche est légèrement supérieure à la longueur maximale du mot recherché. Dans ce cas, la similitude instantanée entre phonèmes n'est plus booléenne mais est calculée à l'aide d'une matrice de substitution interphonémique [3]. Il est possible, mais nous ne l'avons pas encore essayé, d'implanter une rétro-action sur le niveau acoustique quand les phonèmes mis en correspondance sont différents, par exemple, pour calculer un indice de ressemblance à partir de spectres.

Exemple :



élision du [ə] et fusion du groupe [ʒ], [v] et du groupe [ɛ][v]

Reconnaissance de je voudrais dans j'voudrais poste ...

4. CONCLUSION

Nous avons présenté dans cet article divers aspects du niveau lexical de traitement dans un système de compréhension automatique de la parole. Dans un tel système, le lexique doit contenir toutes les informations de diverses natures relatives aux mots, et son rôle est primordial dans l'émission et la vérification d'hypothèses, c'est-à-dire dans le fonctionnement du système. L'organisation lexicale présentée ici est en début de réalisation, ce qui n'a pas permis de varier les exemples pratiques. Enfin, certains points n'ont pas été abordés, faute de place, par exemple les problèmes très importants d'apprentissage en rapport avec le lexique, liés par exemple à l'inférence automatique de lexiques.

5. BIBLIOGRAPHIE

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

[2] J.M. PIERREL, "Contribution à la compréhension automatique du discours continu" Thèse de spécialité, Université de Nancy I, nov 1975.

[3] J.F. MARI, "Thèse de 3ème cycle" (à paraître), Université de Nancy I.

[4] J.S. BRIDDLE, "An efficient elactic-template method for detecting given words in running speech" Brit. acoust. soc. Proc. Tome 2, 1973.

[5] R.W. CHRISTIANSEN and C.K. RUSHFORD, "Detecting and locating key words in contineous speech using linear predictive coding" IEEE Trans on acoustics, Speech and Signal Processing, Vol. ASSP-25, n° 5, oct. 1977.

[6] R. VIVES, "L'analyse lexicale dans le système KEAL pour la reconnaissance de la parole continue" 7e JEP, Nancy, 1976.

[7] J.Y. GRESSER et R. VIVES, "A similary index between strings of symbols, Application to automatic word and language recognition". Proc of the first Inter. Joint conf. on Pattern Recognition, oct. 30 - nov. 1, 1973, Washington.

[8] B. CAUSSE, D. DOURS, R. FACCA, G. PERENNOU, "Evaluation d'une méthode ascendante d'analyse lexicale dans le discours continu", 7e JEP, Nancy, 1976.

[9] V.M. VELICHDO and N.G. ZAGORIKO, "Automatic Recognition of 200 words" Int Journal of Man-Machine Studies, 2, 223, 235, 1970.

[10] H. SAKOE and S. CHIBA, "A dynamic-programming approach to continuous speech recognition", Poc. IEEE Symp on Speech Rec., pp 101-104 april 1974.

[11] J.M. PIERREL, "MYRTILLE II un système de compréhension du discours continu", rapport CRIN 78-R-019, juillet 1978.

[12] J.P. HATON, J.M. PIERREL, "Data Structures and Architecture of MYRTILLE II Speech Understanding System", 4th I.J.C.P.R., nov. 7-10, 1978, Kyoto, Japan.

10èmes JOURNÉES D'ÉTUDE SUR LA PAROLE

GRENOBLE - 30 MAI - 1er JUIN 1979

ELEMENTS D'UN LEXIQUE POUR LA RECONNAISSANCE AUTOMATIQUE DES COMPTINES ENFANTINES PARLEES

N.Tirandaz et C.Berger-Vachon

Laboratoire de Physique Electronique (Prof.Mesnard),
Université Claude Bernard
Lyon

RESUME

Dans cet article, les auteurs envisagent les éléments à introduire dans un dictionnaire utilisable dans une méthode de reconnaissance automatique des comptines enfantines parlées.

Une première étude sur les comptines conduit à la distribution syllabique des phonèmes. Puis, à chacun des mots, on associe un code syntaxico-sémantique construit à partir de la psychologie de l'enfant ; et on étudie la succession des différents codes dans les comptines. On envisage ensuite l'organisation sémantique des phrases pour renforcer la prédiction et pour construire une grammaire déduite des comptines.

La prise en compte de ces données est indiquée sur un exemple.

SOME LEXICAL ELEMENTS NEEDED FOR SPOKEN CHILD DITTIES AUTOMATIC RECOGNITION

N.Tirandaz and C.Berger-Vachon

SUMMARY

In this paper, the authors consider elements to be introduced in a dictionary used in an automatic recognition system of spoken child ditties.

Firstly ditties analysis leads to a syllabic distribution of phonems ; then a syntactic-semantic code is given with each word. This code is built according to the child psychology. The codes succession in ditties is established. Then, the semantic organization of ditties is studied to strengthen the word prediction and to produce a grammar network.

A way to process these datas is shown on an example.

10èmes JOURNÉES D'ÉTUDE SUR LA PAROLE
GRENOBLE - 30 MAI - 1er JUIN 1979

ELEMENTS D'UN LEXIQUE POUR LA RECONNAISSANCE AUTOMATIQUE DES COMPTINES ENFANTINES PARLEES

N.Tirandaz et C.Berger-Vachon

INTRODUCTION

On se propose d'étudier la composition d'un dictionnaire utilisable pour la reconnaissance des comptines enfantines par une méthode prédictive tenant compte de contraintes syntaxiques et sémantiques limitées. A cette fin, on utilise les sorties du vocodeur du CNET à Lannion (G.FERRIEU et col.,1968).

L'analyse de l'onde sonore donne, à la suite d'un prétraitement (G.MERCIER 1974), d'une part une suite de phonèmes probables affectés chacun d'un certain poids, et d'autre part une première segmentation en syllabes. La suite du traitement peut relever de plusieurs politiques selon le but recherché.

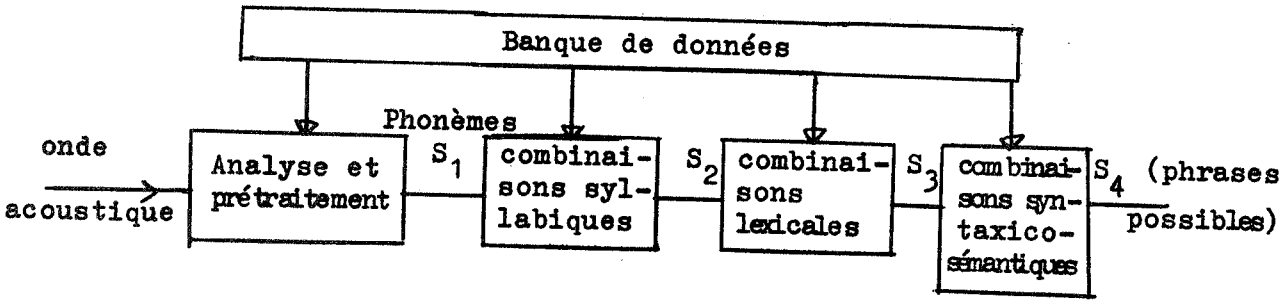
Lorsqu'on cherche à identifier un mot isolé, on peut rechercher une plus grande similarité entre l'archétype sonore déjà en mémoire et les différentes combinaisons des phonèmes. Cette méthode donne d'assez bons résultats dans le cas d'un nombre limité de mots, et elle est déjà assez ancienne (N.LINGREN, 1965). On peut améliorer les résultats lorsque le vocabulaire devient plus important en introduisant les contraintes entre les mots inhérentes aux langages, ou encore en prenant des mots isolés et indépendants construits en fonction des performances de l'étage acoustique (J.A.DREYFUS-GRAF et coll., 1974 ; C.BERGER-VACHON, 1975).

Le problème devient plus délicat lorsqu'on traite la parole en continue. Les méthodes globales semblent totalement inadaptées pour des raisons à la fois épistémologiques (P.M.POSTAL, 1973) et pratiques : temps de machine, volume du vocabulaire, etc. L'introduction de la syntaxe, de la sémantique et de la pragmatique devrait permettre de lever des ambiguités de l'étage acoustique. C'est la syntaxe, toutefois qui, le plus souvent, joue un rôle prépondérant dans ce traitement (B.CAUSSE, D.DOURS, R.FACCA, G.PERENNOU, 1976 ; J.P.HATON, J.M.PIERREL, 1976 ; S.E.LEVINSON, 1978 ; S.E.LEVINSON, A.E.ROSENBERG, J.L.FLANAGAN, 1978). L'apport de la sémantique et de la pragmatique reste limité à une restriction de l'environnement : l'automatisation d'un standard téléphonique par exemple et le système KEAL de Lannion (R.VIVES, 1976 ; P.QUINTON, 1976), l'interrogation d'une banque de données (L.HARRIS, 1977, où l'interrogation ne se fait pas oralement), ou encore l'automatisation des réservations dans une agence de voyage (L.SAITTA, 1978).... Cette étude est une tentative d'évaluation de la contribution d'une sémantique plus large dans la chaîne de reconnaissance.

L'introduction de la sémantique semble à la fois évident et délicat dans un processus de reconnaissance automatique, évident puisque c'est l'objet même de la sémiotique du langage, délicat car il faut préciser le contenu même du message à reconnaître. Il n'est pas nécessaire d'entendre et d'identifier tous les phonèmes d'un mot pour le reconnaître. L'écoute

n'est pas seulement une reconnaissance de forme passive, mais c'est surtout une activité psychologique et une élaboration du discours. D'autre part, on sait depuis longtemps que la compréhension est plus précoce chez l'enfant que la locution : la compétence précède la performance (D.MC NEIL, 1970, a et b). Le langage enfantin semble des plus intéressants pour l'introduction d'une sémantique élémentaire. Ce langage n'est pas simple, comme pourrait l'être un langage artificiel (V.BELLUGI et R.BROWN, 1964), mais il permet l'introduction d'une grammaire déduite et d'une sémantique rudimentaire en tenant compte de la psychologie de l'enfant (ou supposée telle).

On propose la stratégie de reconnaissance suivante (fig.1).



S_1 : ensemble ouvert comprenant la liste des phonèmes les plus probables

S_2 : ensemble ordonné de combinaisons possibles affectées chacune d'une mesure de vraisemblance

S_3 : ensemble des mots possibles

S_4 : la phrase possible.

Figure 1 - Stratégie de reconnaissance

LE CORPUS

Le corpus du vocabulaire à reconnaître (50 comptines enfantines) est composé de 1140 mots répartis en 500 mots distincts. Le dictionnaire, servant de banque de données à la reconnaissance contient d'une part l'ensemble des données nécessaire à la reconnaissance des phonèmes et à une première segmentation en syllable, et d'autre part les informations suivantes quant aux 500 mots du vocabulaire de base :

1º - une décomposition syllabique des mots

2º - la fréquence de chaque syllabe dans le corpus. L'introduction de cette fréquence mérite une petite parenthèse : chez l'enfant et même chez l'adulte, un mot phonétique n'est reconnu que s'il est déjà connu, le régime articulatoire de chacun est grandement conditionné par la langue maternelle qu'il pratique. Il est donc tout-à-fait normal de fournir à la machine ce que l'enfant a besoin pour saisir et distinguer un mot porteur de signification.

3º - la fréquence de chaque phonème

4º - un code caractérisant la catégorie sémantico-syntaxique de chaque mot

5º - un ensemble de relations sémantiques permettant de construire le

réseau sémantique d'une phrase.

6ᵉ - les relations sémantiques implicites entre chaque mot du vocabulaire.

7ᵉ - les automates prédictifs appropriés à l'utilisation de chacune de ces informations.

La composition syllabique du corpus

Les 1140 mots de notre corpus comprennent : 160 syllabes monophonémiques, 1041 syllabes diphonémiques directes (ta, pa, la, etc...), 47 syllabes diphonémiques inversées (el, al ...), 256 syllabes triphonémiques (fla, pat, ...) et 24 syllabes quadriphonémiques (trwa, frwa, etc...).

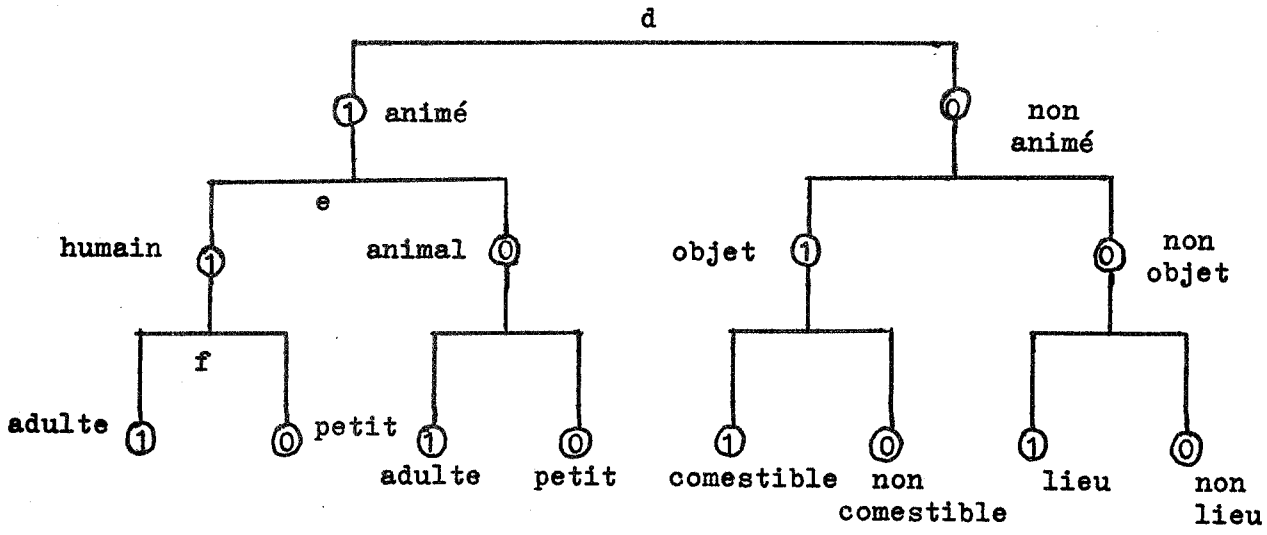
Ces fréquences, liées avec les fréquences des phonèmes, permettent un premier tri entre les diverses syllabes possibles, en appliquant aux phonèmes candidats la formule des probabilités composées de Bayes (cf.tableau 1)

Le code sémantico-syntaxique

Ce code doit permettre de condenser une certaine quantité d'information sur chaque mot, permettant à la fois de comprendre ce qui lui est propre et sa relation avec d'autres mots éventuels. Les informations sont codées en binaire et le code comprend 6 bits : abcdef.

Les bits a et b permettent la distinction entre trois grandes catégories syntaxiques : les noms communs (00), les verbes (01 ou 11) et tous les autres mots (10).

Pour les noms communs, le bit "c" concerne le genre (1, masculin et 0, féminin); les bits "d", "e", "f" portent sur le signifié :



Quelques exemples de codes

0 : nom commun féminin abstrait (exemple bonté, etc)
1000 : nom commun masculin abstrait (exemple voyage, temps)
100111 : nom propre
010011 : verbe infinitif transitif direct et indirect
100000 : article masculin.

Par adulte, nous entendons grand par rapport à l'enfant (aussi bien pour les animaux que pour les humains), une fois que la distinction entre le moi et le non moi est faite ; en effet les catégories grandes personnes et enfants jouent un rôle certain. On peut aussi constater que pour un enfant, un animal, quel qu'il soit, n'est pas comestible ; en effet ce qu'il mange, ce n'est pas un animal mais une viande, qui est un objet inanimé. En ce qui concerne les non animés, l'oralité reprend sa place de premier choix, ce qui amène à distinguer, comestible et non comestible. Ce choix tient à l'importante place de l'oralité dans la psychologie de l'enfant et de l'adulte.

Pour les verbes, nous distinguons d'une part, infinitif, participe passé et les autres éléments de la conjugaison, et d'autre part la nature des verbes : verbes auxiliaires (être, avoir et aller), verbes de mouvements, verbes intransitifs et transitifs.

Pour les autres catégories de mots, on utilise l'ensemble des six bits pour distinguer les noms propres, les articles, les adjectifs, etc.

Dans l'optique d'une grammaire déduite, nous introduirons dans le dictionnaire la fréquence des codes qui précèdent et succèdent à un code donné, les effectifs totaux et la fréquence de chaque code pour commencer ou pour terminer une comptine. Cette démarche semble a priori limiter la généralité de ce travail et trop particulariser notre méthode aux 50 comptines. Néanmoins, nous avons adopté cette approche en vue de l'élaboration d'une grammaire déduite. Cette démarche semble en outre compatible avec le mécanisme d'apprentissage de l'enfant qui se fait par mimétisme. Le tableau numéro 2 indique les informations relatives à la succession des codes dans le dictionnaire.

L'organisation sémantique de la phrase

Une phrase véhicule plusieurs idées. Certaines de ces idées sont implicites, et d'autres sont exprimées par la combinaison des mots de la phrase. Ces combinaisons peuvent être directement perceptibles (tous les éléments de la combinaison existent dans la phrase) ou en partie elliptiques (certains éléments sont manquants). Ainsi dans la comptine : Tara, le petit rat s'en va au Canada, l'idée implicite : Tara (nom propre), le rat (animal doué de mouvement), s'en va (verbe aller) et Canada (nom propre) seront contenus dans les codes. Tandis que l'idée elliptique :"Tara est un petit rat" et celle de"Canada est un lieu"ne sont pas totalement exprimées. L'information contenue dans une phrase sera représentée par un ensemble de réseaux orientés dont il faut préciser les noeuds et les branches. Ceci revient à dire que l'ensemble des codes devra être divisé en 3 catégories : noeuds, branches (ou relations) et qualificatifs (pour les noeuds et pour les branches).

Peuvent être des noeuds : les noms (communs ou propres), les pronoms, les adjectifs (lorsque la relation est un état)
Peuvent être des branches : les verbes et certains mots grammaticaux : et, ou, avec, de, les pronoms possessifs et les adjectifs possessifs ...

Les qualifiants pour les noeuds sont essentiellement des adjectifs. Cependant certains mots tels que "à", "dans", "sur", "àgauche", "àdroite", "sous", etc, peuvent servir, en plus des adverbes, de qualifiants pour les verbes. Ce sont en fait des mots servant à la localisation des choses ou des situations.

La compréhension d'une phrase passe donc par l'établissement d'un réseau

sémantique au fur et à mesure que les mots sont reconnus. Cette élaboration continue doit donc permettre de prédire la catégorie de mots attendue de la même façon qu'un auditeur arrive à anticiper sur le discours d'un locuteur ; d'autre part, on peut enlever les incertitudes sur les mots déjà reconnus par des tests de compatibilité. Cette prise en compte de la compréhension (donc de la sémantique) dans la reconnaissance exige l'établissement d'une série de relations "dites de base", supposées comprises et connues par la machine(et par l'enfant!).

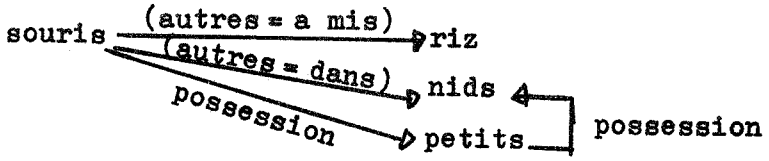
Dans une première étape, nous référant toujours à l'enfant, on peut définir les relations de bases suivantes :

- l'Etat ou l'Existence : l'importance sémantique de cette relation semble évidente.
- Possession et Appartenance : ce sont deux notions réciproques que nous ne distinguerons pas au départ et que nous définirons dans le sens de la possession ; ainsi nous dirons : Jean a des dents, à la place des dents de Jean.
- Accompagnement
- le Mouvement
- Incorporation, dont l'importance psychologique a déjà été soulignée
- Faire
- autres (notions nécessaires pour pouvoir traiter les cas résiduels).

On indique ci-dessous un exemple de liaison sémantique d'une comptine :

"la souris a mis du riz dans le nid de ses petits"

noeuds : souris, riz, nid, petits (les noms communs, code 00----)



On voit donc que la formalisation de cette notion de réseau imposera des simplifications qui ne sont pas encore toutes résolues. Parmi les relations, trois verbes ont retenu plus particulièrement notre attention compte tenu du rôle particulier qu'ils jouent dans ce réseau : ce sont les verbes être, avoir et aller.

Verbe être

Il peut apparaître dans deux configurations possibles : soit en tant qu'auxiliaire, soit en tant que liaison sémantique de type (1) c'est-à-dire exprimant l'état ou l'existence.

Dans les cinquante comptines étudiées, le verbe être apparaît comme auxiliaire trois fois. Chaque fois, il est suivi du participe passé d'un autre verbe ; il faut donc reconnaître le code du mot suivant le verbe être pour pouvoir établir la liaison sémantique entre deux noeuds.

En tant qu'expression de la liaison sémantique "existence", le verbe être apparaît dans notre corpus des comptines quatorze fois. Il semble que nous ne puissions nous attendre à des restrictions dans cette liaison sémantique, car onze fois sur quatorze on a des expressions telles que "c'est" ou "c'était".

Verbe avoir

De même que le verbe être, le verbe avoir peut paraître dans deux confi-

gurations possibles : en tant qu'auxiliaire et dans ce cas le mot qui le suit est un participe passé. Dans nos comptines, ce cas se présente 22 fois et dans ce cas (comme dans le cas du verbe être) il faut trouver la relation exprimée par le verbe conjugué.

La deuxième configuration sémantique possible pour le verbe avoir est la possession (ou l'appartenance). On rencontre 14 fois le verbe avoir dans cette configuration.

Verbe aller

En tant qu'auxiliaire, le verbe aller est toujours suivi d'un verbe à l'infinitif (code 010...). Dans nos comptines, nous le rencontrons une fois. L'auxiliaire a le même rôle que pour les verbes être et avoir. En tant que déplacement et mouvement, nous rencontrons le verbe aller 12 fois. En principe, le noeud actif d'un verbe de mouvement ne peut être qu'un être animé. Cependant, il faut introduire quelques exceptions ; par exemple, nous rencontrons dans nos comptines la phrase "Arnaud a un grand bateau, un bateau qui va sur l'eau" ; le noeud actif du verbe aller est "bateau", qui par définition n'appartient pas à la catégorie des êtres animés (homme et animal). Cela nous amène à souligner la nécessité d'une sémantique implicite qui permettrait une telle reconnaissance. Les liaisons sémantiques de ce type, mouvement, acceptent comme noeud passif : soit des noms communs de lieu, soit des noms propres de lieu, plus rarement des noms abstraits (voyage) ou des objets (l'eau).

Une étude extensive de l'ensemble des liaisons sémantiques nous permettra de mieux évaluer la contribution de la sémantique à la reconnaissance de la parole en continue. Cependant, le contenu du dictionnaire, tel qu'il est à ce jour, permet de lever certaines ambiguités. Pour mieux l'illustrer nous allons prendre en exemple la comptine suivante :

"Tara, le petit rat s'en va au Canada avec Sacha le petit chat".

L'étage acoustique nous donne les éléments suivants, en appliquant la méthode et le vocodeur de Lannion, sans apprentissage :

N° de syllabe	1	1	2	2	3	3	4	4
phonèmes probables	k t g p b d ʃ	ã õ a o	r w	ã õ o r	p r j n w m z l v	o e oe	w k n v	o e oe ɛ̃

La fréquence des phonèmes nous permet de commencer les combinaisons avec le phonème t, ensuite la probabilité des syllabes tã, ta et to étant sensiblement équivalente , nous constatons que dans notre lexique nous avons les mots suivants : tant, temps, ta. Or aucun n'appartient à un code permettant le début d'une comptine. On doit alors chercher un mot disyllabique (les plus nombreux dans les comptines) ; la deuxième syllabe peut être indifféremment rã, rõ et ro.

Nous constatons (dans notre lexique, cf.fig.2) qu'aucun mot disyllabique ne peut être formé. Mais la modification de la dernière syllabe par substitution de |a| à |ã| donne une solution valable : TARA....

mot	TARA	NID	SOURIS	VA	LOUP
division syllabique	2	1	2	1	1
code	100111	1001	100	111101	1101
combinaison syl.					
1er choix	TA-RA	NI	SU-RI	VA	LU
2ème choix	TAR-A	-	SUR-I	-	-

Figure 2 - Extrait du lexique concernant les mots.

CONCLUSION

La prise en compte de la sémantique ne permet pas de résoudre l'ensemble des difficultés rencontrées dans la reconnaissance de la parole en continue. En fait, il est difficile de rattraper de trop grosses défectuosités de l'étage acoustique, mais on peut lever certaines ambiguités lorsqu'une ou plusieurs solutions sont phonétiquement possibles.

Une méthode prédictive, bouclée, permet la recherche d'une solution en proposant à chaque instant, le choix de solutions secondaires satisfaisant aux tests de compatibilités.

Les essais de reconnaissances portant sur l'ensemble des comptines permettront de préciser la portée de cet apport sémantique.

	a	e	o	u	i	ɛ	y	ã	ɛ̃	õ	ə	ø	tri	tot
p	.059	.035	.059	.059	.059			.035	.118	.094	.118		.365	.027
t	.057	.100	.057	.029	.107	.007	.114	.057	.036	.029	.121		.221	.044
k	.250	.014	.194	.028	.208			.028	.014	.028	.042		.194	.023
b	.093	.019	.074	.074	.130		.037	.037	.037	.019	.056		.333	.017
d	.021	.186	.036		.093	.007	.114	.114	.036	.029	.257	.064	.043	.044
g	.172	.034	.069	.069				.034			.034		.552	.009

Tableau 1 : Extrait de la distribution syllabique des phonèmes.
Ex. la probabilité d'avoir "PO" compte tenu du fait que p est connu, est de .059.

La colonne "tri" donne la fréquence des syllabes triphonémiques qui commencent par le phonème de la ligne, tandis que la colonne "tot" donne la fréquence d'occurence du phonème dans l'ensemble des comptines.

code	fréquence du code	fréquence en début	fréquence en fin
000000	.012	.000	.143

(nom féminin abstrait) est suivi de :
(10011,.071),(100010,.071),(100110,.143),(100111,.071)
(101000,.071),(101100,.071),(101110,.214),(101111,.143)
suit les codes suivants :
(100010,.143),(100000,.143),(101110,.071),(101001,.071)
(101111,.143)

100000	.096	.073	.000

(article masculin) est suivi de :
(1000,.136),(1111,.018),(1110,.009),(1101,.055),(1100,.191)
(1001,.064),(1011,.091),(1010,.227),(100001,.009),(100110,.009)
(110111,.018),(101000,.109),(101111,.073)
suit :
(1000,.018),(1101,.018),(1100,.009),(1010,.045),(11,.018),(10,.009)
(10001,.018),(10011,.036),(10100,.009),(11011,.045),(110110,.009)
(111010,.018),(111011,.145),(111101,.018),(111110,.009),(111111,.045)
(100011,.027),(100100,.009),(100110,.009),(100111,.136),(101000,.018)
(101001,.018),(101000,.018),(101101,.018),(101110,.155)
(101111,.009),(1010,.009)

Tableau 2 - Organisation du dictionnaire contenant les codes. Dans chaque parenthèse, on indique un code et sa fréquence.

Tableau 3 : Quelques exemples de comptines :

Amédée
le jardinier
m'a donné
une clef
toute dorée

Mon ami Arnaud
a un grand bateau
un bateau très beau
avec un drapeau
quand il va sur l'eau
sur l'eau du ruisseau
tous les animaux
applaudissent Arnaud

Le geai
a vendu
de la glu
à l'araignée
pour coller
ses filets

Dans la rue
j'ai vu
bubu la tortue
qui m'a vendu
de la laitue

BIBLIOGRAPHIE

1 - V.BELLUGI, R.BROWN - "The acquisition of language" Mon.Soc.Child development, 1964, 29, n°1.

2 - C.BERGER-VACHON, G.MESNARD - "Etude théorique et expérimentale des confusions données par un vocodeur. Application à la reconnaissance de la parole." Annales des Télécommunications, 30, 5-6, mai-juin 1975, pp 139-149.

3 - B.CAUSSE,D.DOURS,R.FACCA,G.PERENNOU - "Evaluation d'une méthode ascendante d'analyse lexicale dans le discours continu". 7e J.E.P. Proc. pp 37-55, 1976.

4 - J.A.DREYFUS-GRAF et coll. - "Reconnaissance objective et subjective de la parole (phonocode) JEP 1974, proc. pp 132-136.

5 - G.FERRIEU,J.PONCIN,G.ROUX,J.VINCENT-CARREFOUR - "Synthèse et reconnaissance de la parole par calculateurs". Echo des Recherches, juin 1968, pp 30-42.

6 - J.P.HATON,J.M.PIERREL - "Interaction entre le niveau lexical, syntaxique et sémantique en reconnaissance de la parole en continue". 7e JEP proc. pp 73-89, 1976.

7 - C.R.HARRIS - "A high performance natural language interface for Data Base Query". Technical report - TR77-1 Dartmouth College, USA, 1977.

8 - S.E.LEVINSON - "The effect of syntactic analysis on word recognition accuracy". Bell System, Techn.Journal (USA) 57, 5, 1627-1644, 1978.

9 - S.E.LEVINSON,A.E.ROSENBERG,J.L.FLANAGAN - "Evaluation of a word recognition system using syntax analysis" Bell System, Tech Journal (USA) T.57, N°5, pp 1619-1626, 1978.

10 - N.LINGREN - "Automatic recognition of human language". IEEE Spectrum, T2, mars p.114, avril p.44, mai p.104, 1965.

11 - D.MCNEILL (a) - "The development of language". P.E.MUSSEN(ed) Carmichael's manual of child psychology, 3e ed. New York, 1970.

12 - D.MCNEILL (b) - "The acquisition of language. The study of developpemental linguistics - New York - Harper Row - 1970.

13 - G.MERCIER - "Segmentation automatique de la parole continue en syllabes et suites phonétiques et identification des phonèmes, Programme KEAL III - Rapport CEI/CSI n°2, février 1974, pp 27-28.

14 - P.M.POSTAL - "La sphère de la syntaxe en communication, langage et pensée de G.A.Moller pp 19-26, SIMEP Villeurbanne, 1979.

15 - P.QUINTON - "Un analyseur syntaxique adapté à la reconnaissance de la parole - 7e JEP, 1976 - proc. pp 89-103.

16 - R.VIVES - "L'analyse lexicale dans le système KEAL pour la reconnaissance de la parole en continue". 7e JEP, 1976, proc. pp 115-129.

= =

10èmes JOURNÉES D'ÉTUDE SUR LA PAROLE

GRENOBLE - 30 MAI - 1ER JUIN 1979

UTILISATION DE L'INFORMATION PHONEMIQUE ET SYLLABIQUE POUR LA RECONNAISSANCE DE MOTS PRONONCES ISOLEMENT OU DANS DES PHRASES.

R. VIVES

CENTRE NATIONAL D'ETUDES DES TELECOMMUNICATIONS
DEPARTEMENT "SERVICES SPECIAUX DU TELEPHONE"

RESUME

Les progrès réalisés depuis quatre ans dans le domaine de la reconnaissance de la parole au niveau de la segmentation syllabique et de la segmentation phonémique du spectre acoustique d'un signal de parole, ainsi qu'au niveau de l'étiquetage des segments phonémiques ont induit des modifications intéressantes du traitement au niveau de la détection lexicale.

Dans cet article on trouve une mise à jour des techniques de reconnaissance de mots isolés et de détection de mots dans des phrases, utilisées dans le système KEAL. Les méthodes présentées peuvent s'appliquer dans tous les systèmes de reconnaissance de la parole travaillant par niveaux successifs d'analyse.

DISCRETE WORD RECOGNITION AND CONNECTED SPEECH RECOGNITION USING PHONEMIC AND SYLLABIC INFORMATION

R. VIVES

CENTRE NATIONAL D'ETUDES DES TELECOMMUNICATIONS
DEPARTEMENT "SERVICES SPECIAUX DU TELEPHONE".

SUMMARY

This paper deals with the lexical level of a speech understanding system. The design and the use of an index operating between phonetic and lexical level have been already described (VIVES, R., & al., 1973).

The last two years, the results obtained in discrete word-and in connected speech recognition, have led to a new definition of the algorithm, first in order to take into account two special phenomenons which often occurs in the phonetic analyzer output (merging of two or more segments and spreading of a phonetic unit), secondly in order to use syllabic informations. Fig. 1 gives an example of the phonetic analyzer output for the French utterance /sɛ̃kɑ̃(t)tRwa/ : The phonetic unit /s/ is spread on the first two segments ; /R/ and /w/ merges in the 8th segment ; detection of the word /ɛ̃/ is performed on the 3rd segment with a rate of 1000/1000 with the old index. Improvement on boundaries and detection rates for /sɛ̃kɑ̃t/ (resp. /tRwa/) is achieved by inflicting a spreading penalization (resp. merging penalization) instead of an insertion one (resp. omission one). The new boundaries of /ɛ̃/ are found by taking into account the syllabic information : /ɛ̃/ belongs to the first syllable /sjɛ̃/ and accordingly, its detection rate is revalued.

The first part of this paper describes the algorithm used to compute a similarity index for discrete word recognition. The second part refers to word detection in connected speech. Then in the third part a discussion on the behaviour of the index is developped. The effect of 6 mis-recognized forms of the word /glwaR/ are shown in fig. 4. The curve 1 represents the index (IR) as a function of the weight x assigned to the spread segment /a/ in /glwa$^{(x)}$R/. The curve 2 represents IR as a function of the weight x assigned to the insert segment ɑ : /glwaɑR). The curve 3 represents IR as a function of the confidence degree assigned to /a/. The effect of the weight x assigned to the segment /a/ is shown in curve 4. The curve 5 (resp. 6) represents IR as a function of the weight x assigned to the merged segment /a/ (resp. /w/) in /glw$\overset{R}{a}$/ (resp. /gl$\overset{a}{w}$R/). The effect of the weight x assigned to the segment /R/of/paRte/ when a match is performed with/pate/(1) and /paRte/(2) is shown in fig. 5. For x = 0.5, advantage is given to/paRte/.

10èmes JOURNÉES D'ÉTUDE SUR LA PAROLE
GRENOBLE - 30 MAI - 1er JUIN 1979

UTILISATION DE L'INFORMATION PHONEMIQUE ET SYLLABIQUE POUR LA RECONNAISSANCE DE MOTS PRONONCES ISOLEMENT OU DANS DES PHRASES.

R. VIVES

CENTRE NATIONAL D'ETUDES
DES TELECOMMUNICATIONS
DEPARTEMENT "SERVICES SPECIAUX
DU TELEPHONE" - LANNION

INTRODUCTION

KEAL est un système de Communication Homme/Machine utilisant le dialogue oral, qui comprend un système de reconnaissance de la parole par niveaux successifs d'analyse (MERCIER, G., & al., 1978 a).

Dans ce système de reconnaissance, le niveau lexical apparaît comme essentiel : il reçoit des informations de l'analyse phonémique (MERCIER, G., 1978 b) et de l'analyse prosodique (VIVES, R.,& al., 1977) et peut fournir des données aux modules syntaxique et sémantique (QUINTON, P., 1977).

Les expériences que nous avons menées depuis l'élaboration des algorithmes de calcul du niveau lexical (VIVES, R., & al., 1973) nous ont conduit à utiliser de plus en plus d'informations élaborées au niveau de l'analyse phonémique et prosodique, et à nous appuyer sur des connaissances linguistiques de moins en moins ténues.

C'est ainsi que nous avons utilisé :

- La segmentation phonémique et les degrés de confiance accordés à ces segments,
- L'étiquetage phonémique multiple des segments et les taux de détection de chaque phonème,
- Des marqueurs prosodiques de début et fin de mot,
- Le codage phonémique "idéal" pour les mots de référence,
- Une mesure de ressemblance entre phonèmes (VIVES, R., 1976).

Les résultats obtenus durant ces deux dernières années en reconnaissance de mots isolés et en reconnaissance de la parole continue nous ont obligés à redéfinir avec précision les algorithmes de calcul du niveau lexical afin qu'ils tiennent compte, de façon cohérente, de phénomènes phonémiques particuliers comme la fusion ou la dispersion de segments phonémiques et d'information nouvelle comme la segmentation syllabique.

La figure 1 donne un exemple de spectre phonémique obtenu après prononciation du nombre "cinquante trois". Dans cet exemple le phonème /s/ apparaît dans les deux premiers segments du spectre phonémique : il s'agit d'un cas de dispersion qui pouvait être aussi pénalisant qu'une insertion dans le calcul de notre indice de ressemblance. Le 8ième segment du spectre phonémique illustre un cas de fusion du /R/ et du /w/ du mot /tRwa/.

.../...

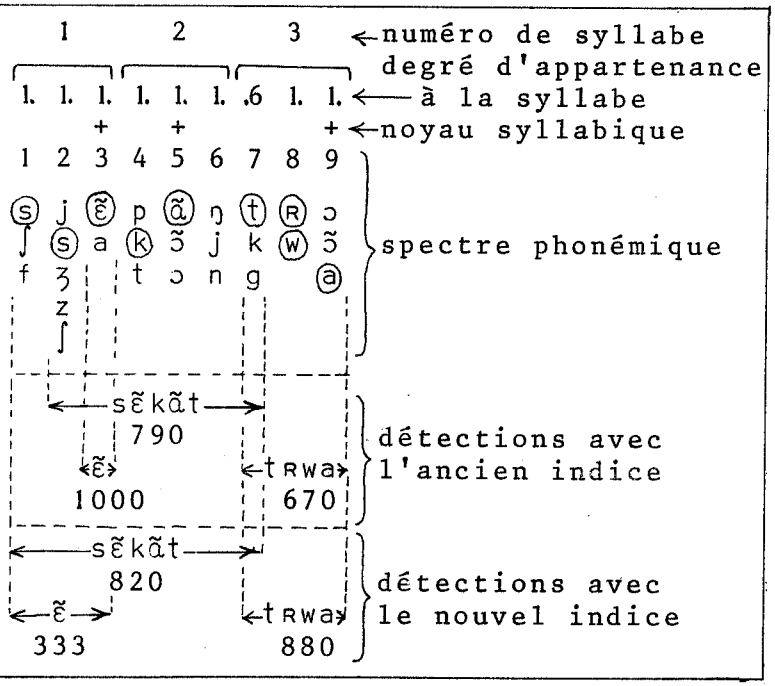


Figure 1: Information syllabique et spectre phonémique obtenu pour la phrase:"cinquante-trois".Visualisation des améliorations sur les limites et les scores de détection calculés pour 3 mots.

Syllabic and phonetic output for the french utterance:"cinquante-trois".Improvement on boundaries and detection rates for 3 words are shown.

Ce cas, relativement fréquent, était traité comme si l'analyse phonémique avait omis le phonème /w/. Enfin le mot "un", codé / ɛ̃ / pouvait être détecté au milieu de la première syllabe /sjɛ̃/ avec un score de 1000/1000 : nous avons utilisé l'information syllabique pour réajuster le score et les limites de détection des mots au milieu des phrases.

Cet article décrit dans une première partie la méthode de base du calcul d'un indice de ressemblance pour effectuer la reconnaissance de mots isolés.

La seconde partie traite du cas de la détection de mots dans la parole continue.

La troisième partie est une discussion sur les effets des différents paramètres intervenant dans le calcul de l'indice de ressemblance.

On évoque les limites de la méthode et l'on propose des axes de recherche qui nous paraissent indispensables à son enrichissement.

DESCRIPTION DE LA METHODE DE CALCUL D'UN INDICE DE RESSEMBLANCE POUR LA RECONNAISSANCE DE MOTS ISOLES

En reconnaissance de mots isolés par la méthode analytique, le programme de segmentation et d'analyse phonémique transforme le signal de parole en un spectre phonémique (SP) constitué d'une suite de syllabes et de segments étiquetés par des phonèmes.

L'algorithme se propose d'établir une relation directe entre le SP d'un mot prononcé et la transcription phonémique idéale (T P I) de chaque mot à rechercher.

.../...

On peut distinguer trois étapes :
- La construction d'une matrice de coïncidence,
- La recherche d'un chemin,
- Le calcul d'un indice de ressemblance R.

1. Matrice de coïncidence

Nous utilisons une mesure de ressemblance entre phonèmes, qui tient compte des deux premiers formants pour les voyelles et de la classification hiérarchisée établie par DELATTRE, pour les consonnes. Entre consonnes et voyelles la ressemblance est considérée comme nulle.

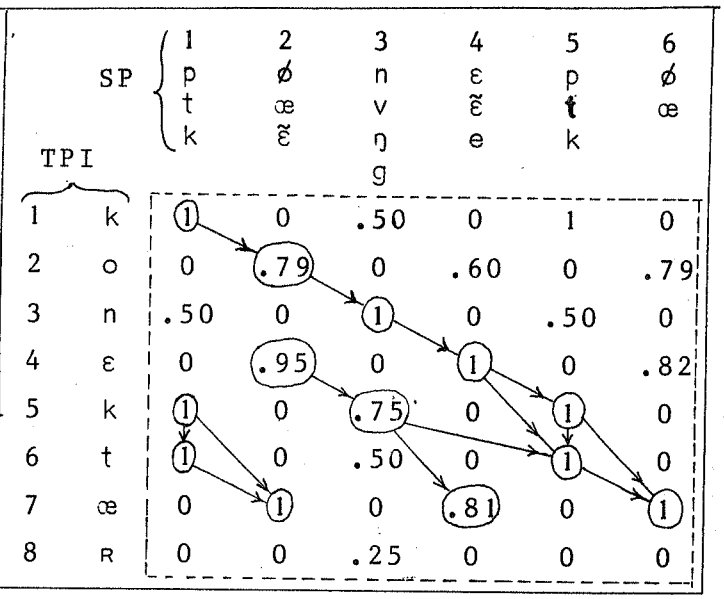


Figure 2: Exemple de matrice de coïncidence; le mot prononcé est: "connecteur". Les chemins explorés sont matérialisés par des nombres encerclés reliés entre eux.

Coincidence matrix for the French uttered word: "connecteur". The different paths appear like arrow connected encircled digits.

SP: spectre phonémique (phonetic analyser output)

TPI: transcription phonémique idéale (phonetic form).

La figure 2 donne un exemple de matrice de coïncidence : chaque case M(i, j) de la matrice de coïncidence contient une valuation, comprise entre 0 et 1, de la ressemblance existant entre le ième élément de la transcription phonémique et le Jième segment du spectre phonémique.

2. Construction d'un chemin

On appelle chemin, une suite d'éléments M (i, j) de la matrice, vérifiant les conditions suivantes :

2.1. M (i, j) $\geqslant$ 0,75 signifie que l'on ne veut s'intéresser qu'aux couples ayant une ressemblance phonémique suffisante.

2.2. Le couple (i_1, j_1), représentant la "tête" du chemin doit satisfaire à la double condition :

$$\begin{cases} 1 \leqslant i_1 \leqslant i_{max}/2 + 1 \\ 1 \leqslant j_1 \leqslant j_{max}/2 + 1 \end{cases}$$

avec i_{max} et j_{max} dimensions de la matrice M.

2.3. Un couple ayant appartenu à un chemin ne peut constituer la tête d'un autre chemin.

.../...

2.4. Deux éléments successifs d'un chemin, $M(i_n, j_n)$ et $M(i_{n+1}, j_{n+1})$ doivent vérifier :

→ soit : $\begin{cases} i_n < i_{n+1} \leqslant i_n + 2 \\ j_n < j_{n+1} \leqslant j_n + m \end{cases}$

(voir zone en pointillé sur la figure 3)

où m est le plus petit entier tel que :

$\sum_{S=j_n+1}^{j_n+m} \text{Poids}(S) \geqslant 2$, avec $\text{poids}(S) \in [0,1]$ symbolisant le degré de confiance accordé au segment S,

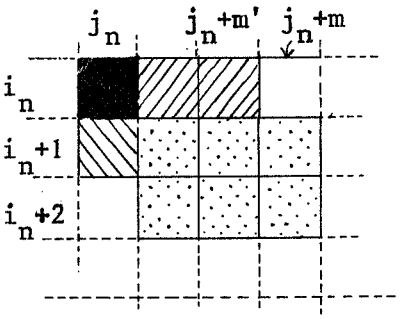


Figure 3 : Le successeur de l'élément $M(i_n,j_n)$ ■, peut se trouver dans la zone hachurée ////// en cas de dispersion, dans la zone hachurée \\\\\\ en cas de fusion ou dans la zone en pointillé.

$M(i_{n+1},j_{n+1})$, the element next to $M(i_n,j_n)$ may lie on the striped area ////// in case of spreading, or on the striped area \\\\\\ in case of merging, or in the stippled area.

→ soit : $i_{n+1} = i_n$ avec $j_{n+1} = j_n + m'$,

déplacement horizontal pour tenir compte d'une éventuelle dispersion d'un phonème de la TPI sur plusieurs segments du SP (voir zone hachurée /// sur la figure 3), où m' est le plus petit entier tel que :

$$\sum_{s=j_n+1}^{j_n+m'} \text{poids}(s) \geqslant 1 ,$$

→ soit : $j_{n+1} = j_n$ avec $i_{n+1} = i_n + 1$, déplacement vertical pour tenir compte de la possibilité de fusion de plusieurs phonèmes consécutifs de la TPI sur un seul segment du SP (voir zone hachurée \\\\ sur la figure 3).

On notera dans la suite n_h et n_v, respectivement, une valuation des déplacements horizontaux et verticaux effectués dans un chemin.

n_h (resp. n_v) sera défini par la somme des degrés de confiance accordés aux segments correspondant aux éléments $M(i_{n+1}, j_{n+1})$ provoquant un déplacement horizontal (resp. vertical).

3. Calcul de l'indice de ressemblance

Un indice de ressemblance est calculé pour chaque chemin défini par une suite de couples (i_1, j_1), (i_2, j_2), ...(i_n , j_n), associés à une séquence de cases $M(i_1, j_1)$, $M(i_2, j_2)$, ...$M(i_n, j_n)$ dans la matrice de coïncidence.

.../...

Plus les deux formes que l'on compare sont ressemblantes, plus les cases M(i, j) qui participent au cadrage sont proches d'une même diagonale dans la matrice M.

L'équation i - j = n ($n \in \mathbb{Z}$), définit, dans cette matrice, un ensemble de cases se trouvant sur une même diagonale. De même l'équation i-j = α, avec

$$\alpha = \frac{\sum_{p=1}^{n} A(i_p,j_p) \times (i_p - \mathcal{J}_p)}{\sum_{p=1}^{n} A(i_p,j_p)} ,$$

peut représenter la "diagonale" la plus proche de toutes les cases $M(i_p, j_p)$ d'un chemin, pondérées par $A(i_p, j_p)$, coefficient global de ressemblance entre le $i_p$ième phonème de la TPI et le $j_p$ième segment du SP. $A(i_p,j_p)$ est défini par :

$$A(i_p, j_p) = M(i_p, j_p) \times \text{poids}(j_p) \times \grave{C}(i_p, j_p) ,$$

avec

$M(i_p, j_p)$ = ressemblance phonémique maximum entre le $i_p$ième phonème de la TPI et le $j_p$ième segment du SP,

poids (j_p)= Degré de confiance ou poids, accordé au $j_p$ième segment du SP,

$\grave{C}(i_p,j_p)$ = Taux de détection du $i_p$ième phonème de la TPI, dans le $j_p$ième segment du SP,

et $\mathcal{J}_p = \sum_{s=1}^{j_p} \text{poids}(s)$.

Pour toutes les cases $M(i_q, j_q)$ du chemin, on pose :

$$e_q = \left| (i_q - \mathcal{J}_q) - \alpha \right|$$

qui représente l'écart de la case $M(i_q, j_q)$ à la diagonale moyenne. L'indice de ressemblance du chemin $M(i_1, j_1)$, $M(i_2, j_2)$,....., $M(i_n, j_n)$ s'obtient par la formule suivante :

$$R = \frac{1}{S^2} \sum_{q=1}^{n} A(i_q,j_q) \times (S - e_q)$$

avec $S = \text{Sup}\left[(i_{max} + n_h), \left(n_v + \sum_{\ell = j_1}^{j_n} \text{poids}(\ell)\right)\right]$

.../...

On évalue ainsi chaque chemin dans la matrice de coïncidence.

La ressemblance entre le SP et la TPI est définie par l'indice de ressemblance maximum trouvé.

DETECTION DE MOTS DANS LA PAROLE CONTINUE

Nous ne reviendrons pas sur les problèmes particuliers de la reconnaissance de la parole continue que nous avons développés dans un précédent article (VIVES, R., 1976). Nous allons simplement indiquer comment nous utilisons l'information syllabique pour parfaire la détection de mots au milieu des phrases.

La figure 1 donne une idée de l'éventail des informations syllabiques fournies par le module de segmentation et d'analyse phonémique : à chaque segment phonémique correspond un numéro de syllabe, un degré d'appartenance à la syllabe et une marque pour le noyau de la syllabe.

Le début et la fin de chaque syllabe, comme le début et la fin de chaque détection lexicale sont caractérisés par un numéro de segment phonémique.

Le nouvel algorithme va procéder à un recadrage du début et de la fin de chaque détection sur respectivement le début et la fin des syllabes correspondantes, seulement dans le cas où le noyau de la syllabe concernée est couvert par le cadrage.

En cas de recadrage, l'indice de ressemblance du chemin $M(i_1,j_1), M(i_2,j_2), \ldots M(i_n, j_n)$ va être modifié par une nouvelle évaluation de S.

Si j_{deb} et j_{fin} sont les nouvelles limites de la détection on aura :

$$S = \mathrm{Sup}\left[(i_{max} + n_h),\ \left(n_v + \sum_{\ell = j_{dep}}^{j_{fin}} \mathrm{poids}\ (\ell)\right)\right]$$

Dans l'exemple de la figure 1, le mot /ɛ̃/, détecté au milieu de la syllabe /s j ɛ̃/, obtenait un score de 1000/1000. Avec la modification, /ɛ̃/ sera cadré sur les limites de la première syllabe et n'obtiendra plus qu'un score de 333. Sur ce même exemple, avec l'ancien indice, on trouvait une détection de /sɛ̃ k ɑ̃ t/ entre les segments 2 et 7.

Le nouvel algorithme en donne une détection entre les segments 1 et 7.

Le recadrage sur le segment 1 est dû à la prise en compte du phénomène de dispersion du /s/ sur les deux premiers segments. Il faut surtout remarquer qu'il n'y a pas de recadrage pénalisant de la fin de la détection sur la fin de la syllabe 3 : le noyau de cette syllabe, /ɔ/, est extérieur au cadrage. Cette clause nous permet de tenir compte des nombreux phénomènes de liaison entre mots, se produisant dans la parole continue.

DISCUSSION

L'indice de ressemblance (IR) que nous proposons intègre un certain nombre de dégradations pouvant apparaître dans le SP, qui peut différer de la TPI par des confusions d'un phonème avec un autre, par des insertions d'un ou de plusieurs phonèmes, par des omissions, par des fusions de plusieurs phonèmes en un seul et par des dispersions d'un phonème sur plusieurs.

.../...

Mis à part le cas de la confusion qui produit une baisse de l'IR uniforme quelque soit son emplacement, on constate que l'IR pénalise plus fortement les dégradations se produisant au milieu des mots, que celles qui apparaissent à leurs extrémités.

La figure 4 illustre les variations de l'IR sur 6 exemples de dégradation du mot /glwaR/. La courbe 1 représente la variation de l'IR en fonction du poids x du segment /a/ qui est dispersé : /glwaăaR/. La courbe 2 montre l'effet de l'insertion, après le /a/ d'un segment de poids x différent de /a/ ou /R/ : /glwaăxR). La courbe 3 donne la variation de l'IR en fonction de la ressemblance du /a/ avec un autre phonème (il y a confusion totale pour une ressemblance nulle). La courbe 4 représente la variation de l'IR en fonction du poids x attribué au segment /a/ (Il y a omission pour x = 0). Les courbes 5 et 6 montrent les variations de l'IR en fonction du poids x attribué au segment sur lequel se réalise une fusion respectivement du /a/ et du /R/, et du /w/ et du /a/. La fusion /glw$\overset{R}{a}$/ est moins pénalisante, car plus externe que la fusion /gl$\overset{a}{w}$R/.

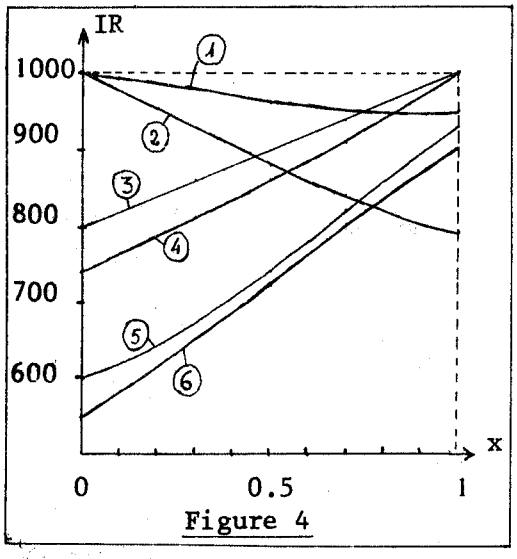


Figure 4

La figure 5 illustre les variations de l'IR entre le SP/paRte/ et les TPI/pate/ (courbe 1) et /paRte/ (courbe 2), en fonction du poids x attribué au segment /R/ du SP. Pour x = 0.5, le calcul favorise /paRte/ (860) par rapport à /pate/ (850) qui a un segment en moins.

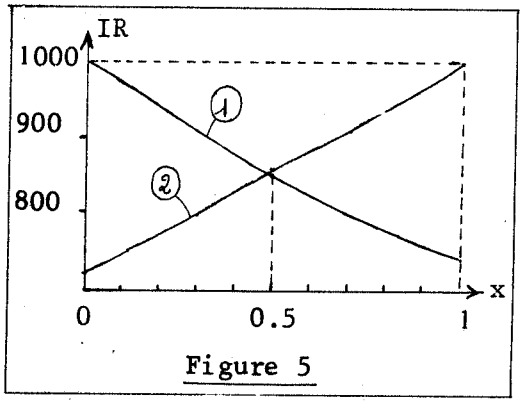


Figure 5

L'IR apparaît comme une approche géométrique de l'idée de similitude pouvant exister entre deux séquences de phonèmes : plus les deux formes que l'on compare sont ressemblantes, plus la séquence de points $M(i_p, j_p)$ qui participe au cadrage est proche d'une même diagonale. Ce concept est malheureusement insuffisant : les mots et les phrases que l'on prononce ne se réduisent pas à des séquences de phonèmes et de syllabes ; il semble nécessaire d'y ajouter des considérations morphologique, phonologique et prosodique.

Le niveau morphologique devrait permettre de rendre l'IR fonction de la qualité des dégradations : la confusion du 2e phonème dans /glwaR/ et dans /paRti/ est pénalisé de la même façon alors qu'elle est loin de provoquer les mêmes effets sur les deux chaînes. Il semble que c'est à ce niveau que les études risquent d'être les plus longues et les plus délicates.

Les phénomènes phonologiques connus comme les réductions de voyelles dans les mots, les liaisons et les enchaînements, peuvent être pris en compte dans le codage des mots de référence mais il sera plus élégant de les contrôler à l'aide de règles au niveau d'un module phonologique.

Enfin au niveau prosodique, des études sont en cours pour utiliser les résultats des travaux de VAISSIERE, J., (1977) sur la durée des syllabes et l'é-

longation des consonnes initiales des mots.

REFERENCES

MERCIER, G., QUINTON, P., VIVES, R., 1978 a, KEAL : un système pour un dialogue avec une machine ; théorie et technique de l'informatique - Actes du congrès de l'AFCET 13-15 nov., pp. 304-314.

MERCIER, G., 1978 b, Evaluation des indices acoustiques utilisés dans l'analyseur phonétique du système KEAL - 9ème J.E.P. LANNION, pp. 321-342.

QUINTON, P., 1977, Utilisation d'un analyseur syntaxique pour la reconnaissance de la parole continue - Annales des Télécommunications Tome 32, n° 9-10, pp. 323-336.

VAISSIERE, J., 1977, Premiers essais d'utilisation de la durée, pour la segmentation en mots, dans un système de reconnaissance - 8ème J.E.P. AIX-en-PROVENCE, pp. 345-352.

VIVES, R., GRESSER, J.Y., 1973, A similarity index between strings of symboles. Application to automatic word and language recognition Proceeding of the 1st International Joint Conference on Pattern Recognition - Washington DC, pp. 308-317.

VIVES, R., 1976, L'analyse lexicale dans le système KEAL pour la reconnaissance de la parole continue - 7ème J.E.P. NANCY , pp. 115-128.

VIVES, R., LE CORRE, C., MERCIER, G., VAISSIERE, J., 1977, Utilisation pour la reconnaissance de la parole continue de marqueurs prosodiques extraits de la fréquence du fondamental - 8ème J.E.P. AIX-en-PROVENCE, pp. 353-363.