73/1

COLLOQUE NATIONAL SUR LE TRAITEMENT DU SIGNAL ET SES APPLICATIONS

NICE du 26 au 30 AVRIL 1977

RECONNAISSANCE ANALYTIQUE DE MOTS PRONONCES ISOLEMENT *

AN ADAPTIVE ISOLATED-WORD SPEECH RECOGNITION SYSTEM

Pierre ALINAT

THOMSON-CSF, Division A.S.M. 06802 Cagnes-sur-Mer (France)

THOMSON-CSF, A.S.M. Division, 06802 CAGNES-sur-MER (France)

RESUME

La reconnaissance de mots prononcés isolément (et appartenant à un vocabulaire limité) est réalisée analytiquement au moyen d'un système composé de trois parties :

- Dispositif d'observation (cochlée artificielle)
- Réduction du débit d'information
- Reconnaissance analytique proprement dite.

Les principes généraux utilisés pour cette dernière étape sont exposés, notamment :

- Localisation des niveaux en descendant la hiérarchie, c'est-à-dire en localisant les mots, puis les syllabes, puis les phonèmes.
- Existence d'unités de décision, c'est-à-dire des niveaux à partir desquels on descend la hiérarchie.
- Attribution d'une note de fiabilité (sûr - probable possible) à la reconnaissance de chaque élément.
- Construction d'hypothèses à partir des éléments les plus fiables et test de la validité de ces hypothèses.

Une réalisation et les résultats obtenus sont brièvement décrits.

SUMMARY

Isolated-word recognition is realized analytically. The system is composed of 3 parts :

a) Analysis device (Analog cochlea)

b) Reduction of the information rate by feature extraction

c) Analytic classification.

For the last part, used principles are described :

- Existence of subunits : features, phonemes, syllabes and words hierarchically organized
- Existence of rules connecting the subunits
- Subunit localization begining with words and ending with features
- Use of level of confidence.

Realized system is described and obtained results are given.

* Cette étude a été partiellement financée par L'institut de Recherche d'Informatique et d'Automatique (Contrat n° 74.132).

RECONNAISSANCE ANALYTIQUE DE MOTS PRONONCES ISOLEMENT
AN ADAPTIVE ISOLATED-WORD SPEECH RECOGNITION SYSTEM

1. - INTRODUCTION.

Dans le domaine de la communication homme-machine, la reconnaissance de la parole est appelée à jouer un rôle prépondérant. Dans une étude expérimentale récente, Chapanis [1] a montré que pour communiquer, la voix était généralement très nettement supérieure aux autres moyens : écriture manuscrite, machine à écrire, vision : les temps d'exécution de travaux complexes (montages d'objets, recherche de renseignements...) varient presque dans un rapport 2 selon que les communications se font avec ou sans voix.

De façon générale, les avantages des entrées vocales sont la rapidité, la facilité d'usage car il n'y a normalement rien à apprendre à l'utilisateur, la liberté laissée aux mains et au regard, et enfin la possibilité de donner des ordres complexes sans utiliser une machine à écrire. Les entrées vocales sont donc particulièrement utiles en tant qu'interface entre un homme (ou un groupe d'hommes) d'une part, et une machine élaborée capable de tâches compliquées d'autre part.

Tout ceci explique que les études sur la reconnaissance vocale soient très nombreuses de part le monde. Elles portent essentiellement sur les points suivants :

- Reconnaissance de mots prononcés isolément et appartenant à un vocabulaire limité. Les principales équipes et réalisations sont citées dans la référence [2]. Il s'agit généralement de systèmes de reconnaissance très sommaire du type extracteur + classificateur (reconnaissance globale) qui nécessitent une phase d'apprentissage pour chaque nouveau locuteur et pour chaque nouveau vocabulaire.

- Reconnaissance de parole continue. Les principales équipes et réalisations sont citées dans la référence [3]. Pour ce genre de système, le mot compréhension semble plus adapté que le mot reconnaissance qui sous-entend normalement choix parmi un nombre fini d'objets. Les réalisations sont bien moins avancées que pour les mots isolés.

L'étude ici décrite porte sur la reconnaissance analytique de mots isolés, c'est-à-dire une reconnaissance utilisant plusieurs niveaux : critères, phonèmes syllabes, mots, propositions, phrases, liés entre eux par des règles.

Avant de décrire le système étudié et les résultats obtenus, nous allons exposer les principes généraux qui ont été utilisés.

2. - PRINCIPES GENERAUX UTILISES.

Les principales étapes du traitement sont :

- Le système d'observation qui convertit le signal acoustique en un certain nombre de mesures.
- La réduction du débit d'information qui permet de ne conserver que les informations nécessaires à la reconnaissance.
- La reconnaissance analytique proprement dite qui consiste à explorer les différents niveaux en se servant des connaissances que l'on possède sur la langue (règles).

Nous allons détailler successivement ces trois étapes.

2.1. Le système d'observation.

On ne saurait trop insister sur l'importance du système d'observation, c'est-à-dire du capteur qui convertit l'onde sonore en signaux électriques. En effet, on oublie trop souvent que les possibilités et la complexité des systèmes de reconnaissance sont conditionnées par les étages d'entrée. Ces étages sont normalement constitués d'un microphone (supposé de bande passante suffisante pour prendre en compte toute la partie intéressante du spectre) suivi d'un dispositif qui réalise plus ou moins une "analyse spectrale". Toute la difficulté réside dans le choix de ce dispositif. Les systèmes généralement utilisés sont fixés plus ou moins empiriquement par les habitudes ou par les possibilités des outils mathématiques :

- Analyse par banc de filtres passe-bande suivis de détection-intégration, les différents paramètres (répartitions des fréquences centrales, fonctions de transfert, nombre de filtres), étant fixés d'après l'expérience acquise empiriquement au cours des années.

- Statistiques des passages à zéro en sortie d'un ou plusieurs filtres passe-bande. Cette méthode est inspirée par le fait que de la parole écrêtée reste compréhensible. Son intérêt est de nécessiter peu de matériel.

- Il existe trois méthodes mathématiques connues pour estimer une densité spectrale :

  . La méthode classique Transformée de Fourier (FFT)
  . Analyse spectrale à maximum d'entropie (codage prédictif linéaire)
  . Analyse spectrale à maximum de vraisemblance.

Les deux premières méthodes sont employées couramment pour l'analyse de la parole.

a) FFT. Cette façon de faire est équivalente à un banc de filtres passe-bande identiques, dont les fréquences centrales seraient linéairement réparties. Le choix des paramètres est donc fort restreint.

b) Codage prédictif linéaire. Là encore, bien que la méthode soit différente, l'échelle en fréquence est toujours linéaire. L'intérêt de cette méthode réside surtout dans sa facilité de mise en oeuvre : en particulier les taux de reconnaissance obtenus en utilisant ce codage ne sont pas meilleurs que ceux obtenus avec un banc de filtres 1/3 d'octave [4].

L'emploi de tous les systèmes d'observation qui viennent d'être décrits, permet d'obtenir des résultats très convenables. Cependant, dans notre cas, pour améliorer les résultats, c'est une cochlée artificielle qui est utilisée. Elle a été décrite lors d'un colloque précédent [9]. Il faut signaler que, bien que moins perfectionnée que celles qui sont étudiées à l'heure actuelle [10,11], elle a déjà été suffisante pour permettre de mettre en évidence le critère relatif aux zones formantiques et celui relatif à la distinction entre [P] [T] et [K].

2.2. La réduction du débit d'information.

Cette opération est effectuée après le traitement par la cochlée artificielle en extrayant toutes les 8 ms les informations suivantes :

- Amplitude
- Présence et valeur du fondamental FO
- Spectre à dominante HF pour les consonnes fricatives
- Position des 2 premiers formants F1 et F2 utilisés pour les voyelles
- Position du formant F2F utilisé pour les consonnes fricatives
- Présence de nasalisation pour les voyelles
- Présence des phases soutenues pour les consonnes explosives sourdes (SP), sonores (SV), nasales (SN) et pour le [L].

Ces informations sont extraites en permanence sans faire appel aux étages supérieurs. Elles ont été choisies à partir des études faites précédemment sur les critères caractérisant les phonèmes [5]. Il faudrait leur rajouter un complément relatif à la phase transitoire des consonnes explosives. Ce n'est que dans la suite du traitement que l'on décide lesquelles on utilise : par exemple pour les voyelles, on ne tient pas compte de F2F et pour les consonnes fricatives on ne tient pas compte de F1 et F2. L'ensemble de ces informations représente environ 5000 bits/s.

2.3. La reconnaissance analytique.

Dans une méthode de reconnaissance analytique (ou encore structurelle, ou syntaxique) il faut distinguer deux points :

2.3.1. Les connaissances a priori sur l'objet à reconnaître, c'est-à-dire :

a) L'existence de niveaux hiérarchisés :

  . critères (plus bas niveau)
  . phonèmes
  . syllabes
  . mots - groupe de mots
  . proposition - phrases.

b) Les règles qui lient tous ces niveaux. Elles peuvent être classées en :

- Règles de localisation qui servent à désigner l'instant ou l'intervalle de temps approximatif où les éléments à reconnaître se situent. Par exemple, ces règles servent à localiser les syllabes, ou les groupes de mots etc... Les règles de prosodie font partie de cette catégorie. Ces règles sont utilisées en passant des hauts niveaux vers les bas niveaux. Elles sont généralement simples.
- Règles de classification qui servent à faire la reconnaissance proprement dite, c'est-à-dire à identifier les éléments précédemment localisés. Ces règles sont utilisées des bas niveaux vers les hauts niveaux. De façon générale, l'ensemble de ces règles de classification est simple et fréquemment employé pour les bas niveaux, plus complexe et moins souvent employé pour les hauts niveaux. Les règles de déduction de la nature des phonèmes à partir des critères et les règles de syntaxe (grammaire) font partie de cette catégorie.

Cette séparation des règles en deux catégories est à rapprocher des notions d'opérateur de commande d'une part, et des opérateurs arithmétiques et syntaxiques d'autre part, explicités dans la référence [6]. Notons qu'il est absolument nécessaire de connaître au moins approximativement l'ensemble des règles ci-dessus pour pouvoir réaliser un système valable.

2.3.2. La façon d'utiliser toutes ces règles.

Il faut prendre en compte toutes les déformations et les bruits qui perturbent le signal :

- Si la parole était parfaitement prononcée, c'est-à-dire en respectant absolument toutes les règles (et avec un grand rapport signal à bruit), la redondance serait considérable. Le locuteur profite de cette redondance pour prononcer de façon plus relâchée, c'est-à-dire en ne respectant pas certaines règles. La prononciation demande alors moins d'efforts et la compréhension est tout de même possible. Pour profiter lui aussi de cette redondance, l'auditeur n'exécute que le minimum d'opérations (en faisant appel au moins de règles possible).

- Pour une langue donnée, le système de règles n'est pas très rigoureux : il peut varier légèrement d'un locuteur à l'autre, d'un auditeur à l'autre. Il varie également selon la région (accent régional) et dans le temps (les langues évoluent lentement).

- Il se rajoute fréquemment à la parole des bruits extérieurs d'amplitude non négligeable.

En tenant compte de tous ces"bruits" et en cherchant à minimiser le nombre et l'importance des opérations à effectuer, on est conduit à utiliser les principes suivants :

a) Il faut descendre les niveaux (de la phrase vers les critères) grâce aux règles de localisation [7] [8]. Par exemple, pour la reconnaissance de mots isolés, il faut localiser le mot, puis les syllabes, puis les phonèmes. En particulier, un mot est localisé au moyen de règles spécifiques à cette tâche et non pas comme une suite de syllabes ou de phonèmes. Cette façon de faire est un cas particulier du principe de traitement de signal : il faut détecter la présence d'un signal avant d'estimer ses paramètres.

Si ce principe n'est pas respecté, c'est-à-dire si on remonte les niveaux (des critères vers la phrase), de très nombreux éléments sont détectés à tort. Par exemple, le système reconnaîtra alors des phonèmes qui n'existent pas en plus de ceux qui ont été réellement prononcés (ainsi CHA È N au lieu de CH É N [5]).

En agisssant selon ce principe, on est moins sensible aux bruits et malformations car les règles de localisation sont d'autant plus "robustes" (c'est-à-dire résistantes au bruit) que le niveau auquel elles s'appliquent est élevé : par exemple, la localisation des syllabes (et donc le compte de leur nombre) est généralement moins sensible aux bruits que la localisation des phonèmes constitutifs de ces syllabes.

b) A toute information est associée une note de fiabilité. Cette note est quantifiée grossièrement par exemple "sûr - probable - possible".

c) A partir des informations les plus fiables, on fait des hypothèses. Ces hypothèses sont testées en fonction de toutes les autres informations (et du contexte). Elles sont acceptées si aucune impossibilité n'apparaît. Il est plus économique de deviner ainsi à partir de ce qui est sûr (quitte à se tromper et recommencer avec une autre hypothèse) que de tester successivement toutes les hypothèses possibles sans tenir compte de ce qui est sûr.

d) Le principe de la descente des niveaux est incompatible avec le fait que la parole arrive sous forme d'un flot continu : il faudrait stocker toutes les informations jusqu'à ce que la parole cesse ! En réalité, il faut fixer un ou plusieurs niveaux supérieurs en-dessous duquel on descend les niveaux : c'est "l'unité de décision" dont il est question dans la référence [7]. Le choix de ces niveaux est un compromis : plus ils sont hauts, moins on fera d'erreur et de calcul (grâce aux principes de descente des niveaux) mais plus on aura d'informations à stocker.

Pour la reconnaissance de parole continue, il y a trois unités de décision qui semblent pratiques : la proposition, le groupe de mots et la syllabe. L'organisation générale du système peut alors être à 3 étages.

- A partir des informations extraites toutes les 8 ms sur une durée de l'ordre de 2 à 3 syllabes :

  . localisation des syllabes
  . localisation des phonèmes
  . reconnaissance des phonèmes

- A partir des phonèmes reconnus (groupés en syllabes) sur une durée de l'ordre de 1 à 2 groupes de mots :

  . localisation des groupes de mots
  . localisation et reconnaissance des mots élémentaires.

- A partir des mots reconnus (et regroupés en groupe de mots) sur une durée de l'ordre d'une proposition :

  . localisation de la proposition
  . compréhension de la proposition.

Bien entendu, chaque fois qu'on passe à une unité de décision supérieure, on est conduit à corriger certaines erreurs de niveaux inférieurs.

Au-dessus du groupe de mots, il faut tenir compte de ce que la langue parlée est assez différente de la langue écrite. Dans notre cas, c'est-à-dire pour la reconnaissance de mots isolés (de longueur inférieure ou égale à 3 syllabes) on a choisi la syllabe et le mot comme unité de décision.

3. - SYSTEME CONSTRUIT ET RESULTATS.

Un système de reconnaissance de mots isolés a été construit en utilisant les principes exposés ci-avant. Au niveau de la reconnaissance des phonèmes, la distinction entre les trois classes de consonnes explosives [P B M] [T D N] [K G], n'est pas prise en compte. De plus, le système n'est pas prévu pour reconnaître les semi-voyelles.

La figure 1 donne un exemple de sortie de l'étage réducteur d'information pour le mot [Mélodie]. Toutes les informations extraites chaque 8 ms sont inscrites sur une même ligne. Ce sont ces informations qui sont envoyées au système de reconnaissance proprement dit dont la première tâche est de localiser les syllabes. Sur la figure 1 les zones encadrées à droite représentent les voyelles (c'est-à-dire les noyaux des syllabes). La syllabe la plus accentuée est la dernière : [I] et la moins accentuée, celle du milieu : [O] qui, prononcée de façon relâchée, devient [EU]

Les différentes consonnes sont recherchées entre les voyelles et la suite de syllabes ainsi obtenue est comparée au vocabulaire.

Le vocabulaire utilisé pour tester le système comportait 100 mots (de 1 à 3 syllabes) ; 10 locuteurs ont prononcé 20 mots chacun (au total 200 mots). Le taux de succès a été de 76 %. Il y a eu 6 % d'erreurs causés par un nombre inexact de syllabes, 8 % d'erreurs dues à de mauvaises reconnaissances de voyelles et de consonnes fricatives et 10 % d'erreurs produites au niveau de la comparaison avec le vocabulaire.

4. - CONCLUSION.

La reconnaissance de parole par la méthode analytique est plus complexe à mettre au point et plus volumineuse à réaliser que celle par méthode globale. Cependant elle a sur cette dernière de grands avantages :

- Pas d'apprentissage pour chaque nouveau locuteur
- La taille du vocabulaire est moins limitée
- La reconnaissance analytique jusqu'au niveau syllabe inclus sera utilisée pour la reconnaissance de parole continue tandis qu'il n'en est pas question pour la reconnaissance globale.

Il reste toutefois à résoudre le problème de la reconnaissance des consonnes explosives et des semi-voyelles. Par ailleurs, il faudra perfectionner la détermination de la position des formants, la recherche de la nasalisation des voyelles et introduire la notion d'accentuation des syllabes.

Lorsque toutes ces questions seront résolues, et à ce moment-là seulement, on pourra envisager une lente adaptation des règles utilisées en fonction des défauts de prononciation (accents régionaux et défauts proprement dits) du locuteur. Une telle adaptation n'a rien à voir avec l'apprentissage servant à obtenir la "copie" nécessaire aux systèmes de reconnaissance globaux : en effet, dans le cas d'une adaptation, les résultats sont corrects dès le début. Ils s'améliorent par la suite, mais il n'y a jamais distinction entre phase d'apprentissage et phase de reconnaissance.

RECONNAISSANCE ANALYTIQUE DE MOTS PRONONCES ISOLEMENT

AN ADAPTIVE ISOLATED-WORD SPEECH RECOGNITION SYSTEM

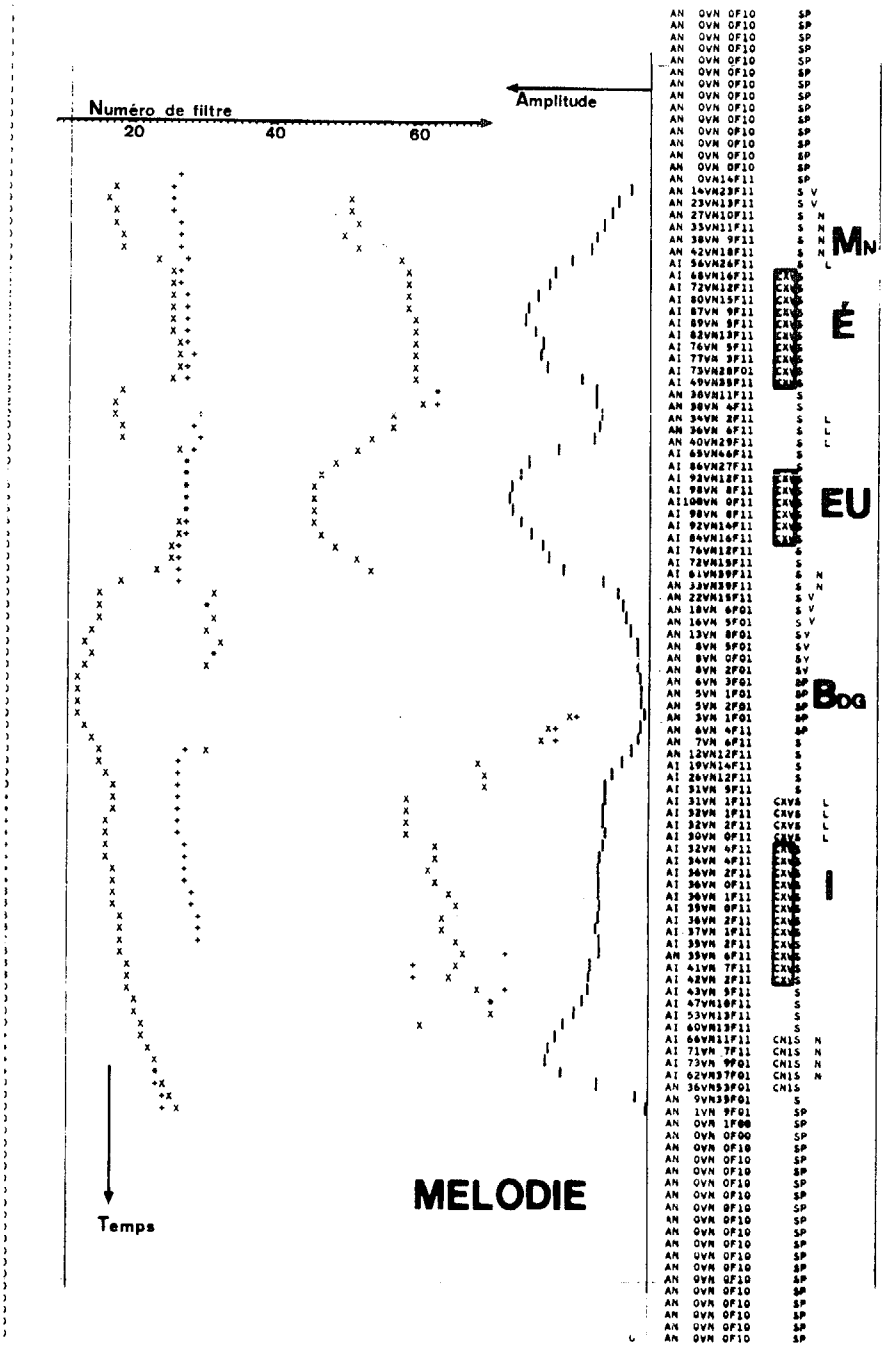

Figure 1 - Sortie de l'étage réducteur d'information pour le mot [Mélodie] . Les deux premiers formants sont représentés à gauche (x) ; de même, le formant utilisé pour les consonnes fricatives (+). Au centre l'amplitude et sur la partie droite des informations relatives au fondamental et aux phases soutenues des consonnes explosives (S P V N L).

RECONNAISSANCE ANALYTIQUE DE MOTS PRONONCES ISOLEMENT
AN ADAPTIVE ISOLATED-WORD SPEECH RECOGNITION SYSTEM

BIBLIOGRAPHIE :

[1] A. CHAPANIS "Interactive Human Communication" Scientific American vol 232 (3) p. 36-42, 1975

[2] T.B. MARTIN "Practical Applications of Voice Input to Machine" Proc. IEEE vol 64, p. 481-501 Avril 1976

[3] G.M. WHITE "Speech Recognition : A Tutorial Overview" IEEE Computer p. 40-53, Mai 1976

[4] G.M. WHITE et R.B. NEELY "Speech Recognition Experiments with Linear Prediction. Band-Pass Filtering and Dynamic Programming" IEEE Trans. ASSP 24 (2), Avril 1976

[5] P. ALINAT "Etude des Phonèmes de la Langue Française au Moyen d'une Cochlée Artificielle. Application à la Reconnaissance de la Parole" Revue Technique THOMSON-CSF vol 7 n° 1, p. 91-123, Mars 1975

[6] C. ROCHE "Information utile en Reconnaissance des Formes et en Compression de Données. Application à la Génération Automatique de Système de Reconnaissance Optique et Acoustique" Thèse Doctorat d'Etat 1972.

[7] G.A. MILLER "Decision Units in the Perception of Speech" IRE Trans. on Information Theory p. 81-83, Février 1962

[8] W.A. LEA "An Approach to Syntactic Recognition Without Phonemics" IEEE vol AU-21 n° 3, p. 249-258, Juin 1973

[9] P. ALINAT "Reconnaissance des Phonèmes de la Langue Française" Colloque sur le traitement du signal et ses applications Nice 7-12 mai 1973

[10] J. CAELEN et G. PERENNON "Un modèle d'Oreille Appliqué à l'Analyse de la Parole" 7ème Journées d'Etudes sur la Parole, Nancy Mai 1976

[11] D. KIM, C.E. MOLNAR, R.R. PFEIFFER "A System of Non-Linear Differential Equations Modeling Basilear Membrane Motion" JASA vol 54 p. 1517-1529, 1973
