Décodage acoustico-phonétique ascendant

A bottom-up acoustic-phonetic decoding system

Henri MELONI

Laboratoire d'Informatique, Faculté des Sciences, 33, rue Louis Pasteur, 84000 Avignon

Docteur ès Sciences (1982), Professeur à l'université d'Avignon, responsable du Laboratoire d'Informatique, Président du Groupe de la Communication Parlée de la Société Française d'Acoustique, Directeur du Groupe d'Intelligence Artificielle de Luminy (URA 816 du CNRS) de 1985 à 1989. Domaines de recherche : Reconnaissance Automatique de la Parole, Programmation en Logique, Apprentissage Automatique.

Philippe GILLES

Laboratoire d'Informatique, Faculté des Sciences, 33, rue Louis Pasteur, 84000 Avignon

Titulaire du DEA d'Informatique et Mathématiques (option Intelligence Artificielle) de l'université d'Aix-Marseille II (1988). Boursier MRT, Moniteur à l'université d'Avignon. Prépare actuellement un Doctorat sur le thème du Décodage Acoustico-Phonétique multilocuteur de la parole continue.

RÉSUMÉ

Nous décrivons un système de décodage acoustico-phonétique qui produit des treillis phonétiques en localisant et identifiant simultanément les unités au moyen de divers types de distances spectrales ajustées en fonction des phonèmes, du contexte et de certaines caractéristiques du locuteur. Les résultats obtenus — pour des mots isolés ou pour des énoncés continus — font apparaître tous les phonèmes effectivement énoncés avec un score d'identification particulièrement intéressant pour la sélection ascendante de cohortes d'items dans un lexique étendu.

MOTS CLÉS

Reconnaissance de la parole, décodage acoustico-phonétique.

ABSTRACT

We describe a bottom-up acoustic and phonetic decoding system which produces phonetic lattices by simultaneously locating and identifying the units by means of various types of spectral distances adjusted according to the phonemes, the context, and the speaker's characteristics. The results — both for isolated words and for continuous speech — give all the phonemes that have been pronounced, with an efficiency particularly interesting for a bottom-up selection of restricted sets of elements in a large vocabulary.

KEY WORDS

Speech recognition, acoustic and phonetic decoding.

1. Introduction

Le problème du décodage acoustico-phonétique de la parole continue n'a toujours pas reçu de solution satisfaisante malgré les travaux considérables qui lui ont été consacrés depuis de nombreuses années [8], [15]. Dans une phase strictement « ascendante » du processus de DAP — du signal vers les symboles phonétiques (syllabes, diphones, phonèmes, phones, traits, indices, etc.) — les difficultés sont d'autant plus grandes que les informations sur le contexte ne peuvent être qu'hypothétiques et généralement peu fiables.

Les diverses techniques employées dans le cadre du DAP « ascendant » (Représentation des Connaissances, Quantification Vectorielle, Modèles de Markov, Réseaux Neuro-Mimétiques, etc.) se révèlent impuissantes à prendre en compte simultanément les informations relatives aux variations aléatoires, contextuelles, inter-individuelles et linguistiques des unités phonétiques.

Une solution fréquemment utilisée consiste à ne retenir lors d'un premier traitement que les informations « sûres » identifiables sans ambiguïté dans le signal vocal [4], [7], [14], [17], [23]. Cependant la notion de fiabilité d'une information est très délicate à utiliser car certains types de

déformations contextuelles des phonèmes rendent vraisemblables des hypothèses erronées en l'absence d'une connaissance précise de l'environnement phonétique et/ou des caractéristiques individuelles du locuteur. Ces erreurs sont d'autant plus dommageables pour la suite du processus de reconnaissance que le coefficient de fiabilité d'une unité symbolique localisée et identifiée est plus grand.

Dans le cadre d'un DAP strictement « ascendant » de la parole continue, les systèmes fondés sur la représentation explicite des connaissances sont peu opérationnels à cause du manque d'informations relatives au contexte, tandis que les systèmes auto-organisateurs réclament de très grandes quantités de données pour encoder les connaissances et pour les adapter aux différents locuteurs.

Une alternative à ces difficultés consiste à proposer, en utilisant une caractérisation très limitée du locuteur (spectres des phases essentielles des phonèmes dans des contextes peu déformants), l'ensemble des unités valuées qui ont une probabilité d'occurrence non négligeable sur une portion de signal. La segmentation en unités symboliques n'est qu'une conséquence de la mise en relief de certaines zones dont les limites ne coïncident pas toujours avec les phonèmes [12]. Le treillis phonétique produit avec cette méthode contient généralement la séquence correcte des phonèmes principaux des noyaux vocaliques et consonantiques. Ces unités permettent notamment de sélectionner des suites valuées correspondant aux items d'un lexique. Une phase de vérification « descendante » est alors nécessaire pour sélectionner — indépendamment du locuteur — les séquences les plus probables dans les cohortes de mots concurrents.

Nous proposons un système qui, à partir d'un apprentissage semi-automatique très limité (énoncé de quelques phrases par chaque locuteur), produit des treillis phonétiques dans lesquels les unités sont localisées et identifiées simultanément au moyen de divers types de distances spectrales ajustées en fonction des phonèmes, du contexte et de certaines caractéristiques du locuteur. Seul un ensemble très restreint de règles générales assure la supervision du processus ; la plupart des traitements « intelligents » ayant été intégrés dans les algorithmes de comparaison.

2. Représentation paramétrique et outils

Le signal de parole est numérisé sur 16 bits à une fréquence de 12,8 kHz puis préaccentué et caractérisé chaque 10 ms par son énergie globale, la densité des passages par zéro et les énergies spectrales dans 24 canaux répartis suivant une échelle de Mel (*fig.* 1). Les spectres peuvent être obtenus par différentes méthodes (FFT, LPC, Cepstre, Vocodeur, modèle d'oreille), mais nous utilisons le plus souvent une modélisation par LPC (14 coefficients). Le choix de cette représentation est dicté par l'impératif de suivre au plus près la formalisation des connaissances acoustiques et phonétiques proposée par les experts de ce domaine.

Un ensemble d'outils permet de définir et de calculer dynamiquement de nombreux paramètres auxiliaires obtenus par combinaison des 26 attributs initiaux [4], [18]. Dans le cadre du système de DAP « ascendant » nous utilisons essentiellement comme paramètres secondaires les valeurs de différentes distances entre les spectres de référence d'un locuteur et ceux calculés sur le signal, diverses fonctions d'instabilité spectrale ainsi que quelques indices évaluant contextuellement les caractères vocalique et consonantique.

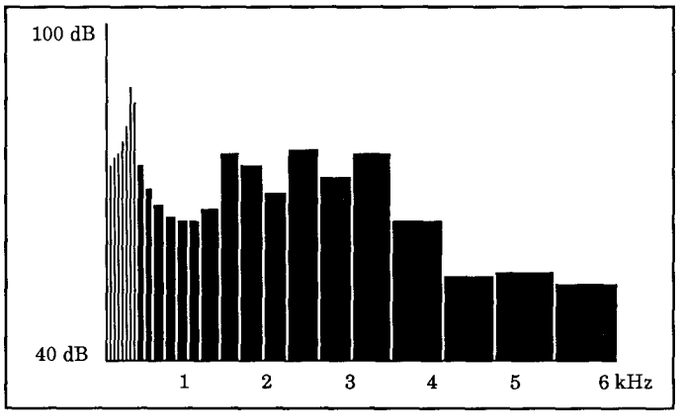

Figure 1. — Répartition de l'énergie spectrale d'un son vocalique dans les 24 canaux distribués suivant une échelle de Mel.

La localisation de certaines zones intéressantes sur les courbes caractérisant l'évolution temporelle des paramètres est effectuée au moyen de prédicats évaluables qui permettent de définir des schémas de formes (pics, vallées, zones monotones, etc.) et d'identifier les événements correspondant à certaines de ces descriptions. De plus, ces outils assurent le passage de la représentation numérique — au moyen des paramètres — vers divers types de symboles plus ou moins complexes constitués d'associations de formes simples [4], [18].

Cet environnement de travail pour la représentation et le traitement de connaissances acoustiques, phonétiques et linguistiques a été utilisé pour réaliser divers systèmes de DAP et de reconnaissance de la parole [5], [17], [19]. Les performances des techniques utilisant des connaissances explicites sont liées à la quantité d'informations symboliques « sûres » disponibles lors de l'activation des règles. Cette méthode convient donc mieux à une phase « descendante » du processus de reconnaissance. Ces difficultés nous ont conduit à la réalisation du système actuel de DAP « ascendant » auquel sera associé un processus de vérification descendante parmi les cohortes de mots déduites du treillis phonétique.

3. Structure générale du système

3.1. RÉFÉRENCES SPECTRALES

Le système de décodage acoustico-phonétique « ascendant » (*fig.* 2) utilise un ensemble de références spectrales correspondant aux phases stables des phonèmes (tenue des voyelles et des constrictives, occlusion des plosives) et

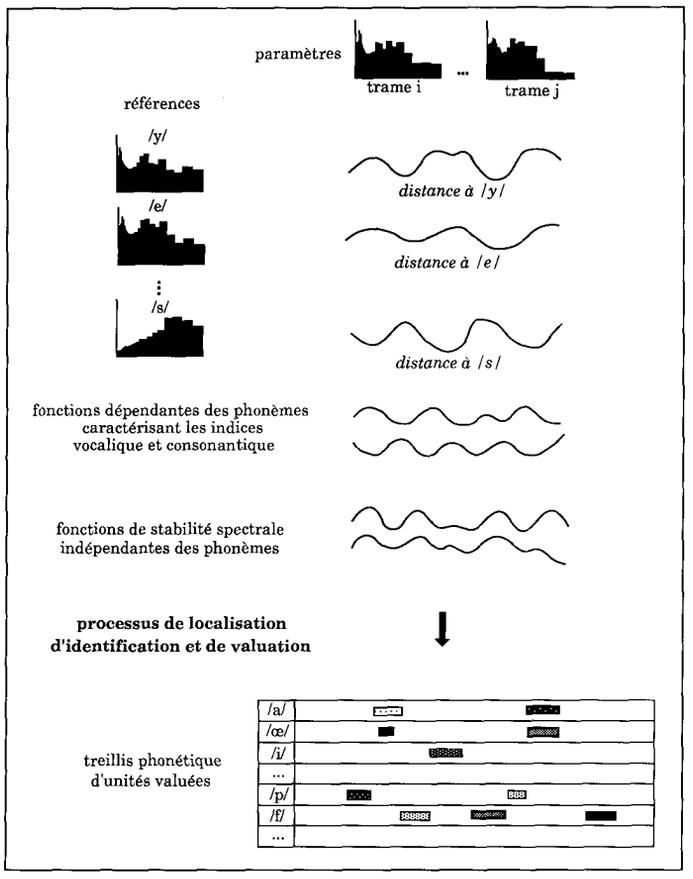

Figure 2. — Schéma du système de Décodage Acoustico-Phonétique « ascendant ».

de certaines zones transitoires intéressantes (burst des occlusives sourdes). Le but ici n'est pas de décrire précisément les sons — qui de toute manière sont plus ou moins fortement modifiés par le contexte — mais de proposer une forme moyenne proche de la réalisation idéale du phonème pour un locuteur.

L'acquisition des références est effectuée pour chaque locuteur à partir d'un ensemble très limité de phrases dans lesquelles les phonèmes apparaissent dans des contextes peu déformants. Cette phase d'adaptation est donc très rapide et peu contraignante. Toutefois, des modèles de références indépendants du locuteur ont été utilisés dans certaines applications (localisation de phonèmes dans des phrases connues) avec des résultats intéressants du point de vue de la localisation mais dégradés pour la valuation des unités identifiées.

3.2. STRATÉGIE GÉNÉRALE

Différents types de distances aux références sont systématiquement calculés et constituent les paramètres essentiels pour localiser et identifier les phonèmes. Les intervalles du signal dont l'une des distances est inférieure à un seuil très large sont potentiellement considérés comme candidats probables. Nous verrons que les décisions sont d'autant plus faciles à prendre que les algorithmes de calcul des distances intègrent une certaine « intelligence » résultant de la prise en compte de connaissances diverses.

Un ensemble très restreint de règles simples utilise les hypothèses déduites à partir des distances ainsi que certains paramètres supplémentaires (instabilités spectrales, caractère vocalique ou consonantique, etc.) pour proposer des unités phonétiques valuées au moyen d'une distance indépendante du phonème en un point particulièrement stable du signal.

4. Distances

Pour le calcul de distances à des vecteurs de référence représentant des unités phonétiques de nombreuses techniques ont été utilisées dans de multiples contextes [2], [11], [13], [20]. Le choix d'une méthode optimale dépend de nombreux facteurs tels que la qualité du signal, le type de représentation paramétrique (coefficients MFCC, spectres, etc.), le nombre de vecteurs employés dans le calcul, la prise en compte d'informations concernant l'unité phonétique à identifier et/ou son contexte, etc.

Malgré l'adaptation des références spectrales au locuteur, la variabilité acoustique des phonèmes — résultant essentiellement des phénomènes de coarticulation — demeure importante et interdit une identification très précise des unités phonétiques. Cependant, les distances doivent mettre en valeur les informations les plus pertinentes dans des contextes divers. Compte tenu du système de représentation paramétrique que nous avons choisi, notre travail a consisté à résoudre les problèmes suivants :

— adaptation au locuteur rapide et peu contraignante,

— ajustement des niveaux d'énergie spectrale,

— prise en compte optimale des variations de position et d'amplitude des maxima spectraux (formants),

— intégration d'informations contextuelles disponibles dans le signal.

4.1. COMPARAISON DIRECTE OU DIFFÉRENTIELLE

Une manière simple de mesurer la distance entre deux spectres est de calculer la somme de la valeur absolue ou du carré des écarts entre chacun de leurs 24 canaux. Exemple :

$$d_t = \frac{1}{24} \sum_{i=1}^{24} |Cr_i - Cs_i|$$

où Cr_i correspond au canal i de la référence et Cs_i au canal i du signal pour la trame t). Cette technique a plusieurs inconvénients dont les deux principaux sont le masquage de particularités spectrales significatives et la difficulté d'un ajustement optimal des énergies.

Une alternative à ces difficultés consiste à effectuer une mesure différentielle (comparaison de la dérivée des spectres) qui met en évidence les mouvements des formants. Dans ce cas l'ajustement des niveaux d'énergie est automatique. De plus, si l'on borne supérieurement les valeurs des écarts, on favorise la position fréquentielle des

formants par rapport à leur amplitude dont on minimise les variations prises en compte. Une des distances différentielles utilisées correspond à la formule :

$$d_t = \frac{1}{24-k} \sum_{i=k+1}^{24} \min \times \times [\,|(Cr_i - Cr_{i-k}) - (Cs_i - Cs_{i-k})|, Sr_i\,] .$$

Le seuil Sr_i peut dépendre du phonème de référence et de la position du canal dans le spectre. La valeur de k (nombre de canaux de décalage) est généralement égale à 1.

Ce type de distance est particulièrement intéressant pour les spectres dont les formants sont nettement marqués (la plupart des voyelles à l'exception des nasales et de /u/ (*fig.* 3)).

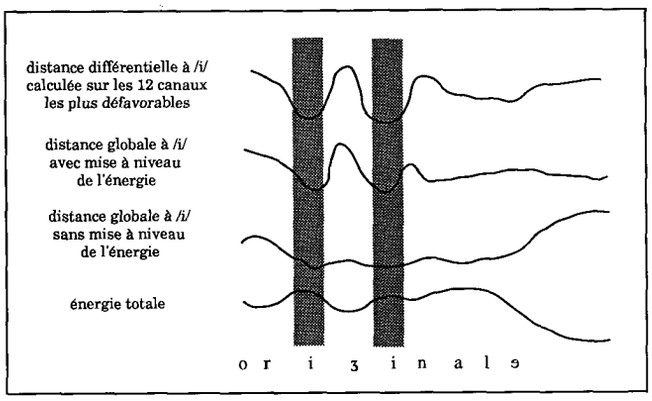

Figure 3. — Valeurs de quelques distances à la voyelle /i/ sur une portion de phrase. Sur cet exemple, la localisation et l'identification sont plus précises si l'on prend en compte les 12 canaux maximisant les écarts sur la dérivée.

4.2. PRISE EN COMPTE DES CARACTÉRISTIQUES DES PHONÈMES

Dans certains cas l'utilisation des distances différentielles donne des résultats médiocres et nous calculons directement les valeurs sur les canaux. Pour échapper aux difficultés signalées dans le paragraphe précédent, il est indispensable de sélectionner de manière optimale le sous-ensemble des canaux utiles et de réaliser l'ajustement des énergies.

La connaissance des caractéristiques spectrales des phonèmes permet de résoudre en partie ces problèmes. La liste des canaux sur lesquels sera effectuée la comparaison est spécifique à chaque phonème de même que la liste des canaux utilisés pour mettre les énergies à niveau. Dans ce cas nous obtenons une distance calculée suivant la formule :

$$d_t = \frac{1}{m} \sum_{j=1}^{m} |Cr_{i_j} - Cs_{i_j} + d\mathrm{E}_t|$$

avec $i_j \in [1, 24]$ et $j = 1, ..., m$

$$d\mathrm{E}_t = \frac{1}{n} \sum_{k=1}^{n} Cr_{i_k} - Cs_{i_k}$$

avec $i_k \in [1, 24]$ et $k = 1, ..., n$.

Ce type de distances localise et identifie correctement les phonèmes existants effectivement dans le signal mais propose quelquefois des hypothèses recevables dans des zones où le phonème n'apparaît pas. Pour certaines unités telles que les silences et la phase d'occlusion des plosives sourdes, nous n'effectuons pas de réajustement des paramètres de manière à mettre en évidence le niveau particulièrement bas de l'énergie.

Une autre façon de procéder au choix des canaux de comparaison indépendamment du phonème consiste à ne retenir pour le calcul de la distance que les n éléments dont la différence est la plus grande. L'ajustement des énergies peut être effectué soit sur un ensemble donné de canaux dépendant du phonème, soit sur l'énergie totale des 2 spectres. Les distances correspondent donc, dans ce cas, à des fonctions du type :

$$d_t = \frac{1}{n} \sum_{i=1}^{n} d\mathrm{C}_i \quad \text{avec} \quad i \in [i_1, i_n]$$

$$d\mathrm{C}_{i_1} \geqslant d\mathrm{C}_{i_2} \geqslant \cdots \geqslant d\mathrm{C}_{i_n} \geqslant \cdots \geqslant d\mathrm{C}_{i_{24}}$$

$$d\mathrm{C}_j = |Cr_j - Cs_j + d\mathrm{E}_t| .$$

Suivant la valeur donnée à n, les distances correspondantes sont plus ou moins efficaces pour localiser et identifier les phonèmes dans des contextes particuliers (*fig.* 3).

4.3. PRISE EN COMPTE DU CONTEXTE

L'ajustement des énergies est effectué soit à partir des canaux significatifs d'un phonème, soit indépendamment des unités au moyen de l'énergie globale. Nous pouvons améliorer la sélectivité des distances en réalisant la mise à niveau de spectres par rapport à celui qui, dans un contexte proche, peut être le plus représentatif de caractéristiques particulières du phonème concerné. Nous distinguons trois situations correspondant aux voyelles, aux consonnes fricatives sourdes et aux autres consonnes. Dans chacun de ces cas, la mise à niveau de l'énergie ne prend pas en compte la trame traitée mais celle de son environnement immédiat (dans l'intervalle $[t-3, t+3]$ par exemple) qui atteste le mieux le caractère du phonème recherché.

Pour les voyelles, la trame utilisée est celle qui comporte le plus d'énergie dans les canaux correspondant au phonème traité, pour les fricatives sourdes le spectre pris en compte est celui associé au maximum de la densité des passages par zéro, et enfin pour les autres consonnes le calcul est effectué sur la trame d'énergie minimum. Cette dernière particularité a des conséquences sur certains choix stratégiques et favorise la localisation correcte des consonnes les plus fermées dans des séquences consonantiques. Dans de tels contextes, la consonne d'aperture minimale est assurée d'être localisée convenablement, tandis que les autres consonnes plus ouvertes ne seront pas toujours identifiées lors de la phase ascendante du DAP. Une simple vérifica-

tion descendante, fondée sur les distances au phonème permet de repérer l'unité dans une zone temporelle restreinte.

Les résultats obtenus avec cette technique sont très sensiblement meilleurs surtout pour les phonèmes importants des noyaux vocaliques et consonantiques pour lesquels la localisation est beaucoup plus précise (*fig.* 4).

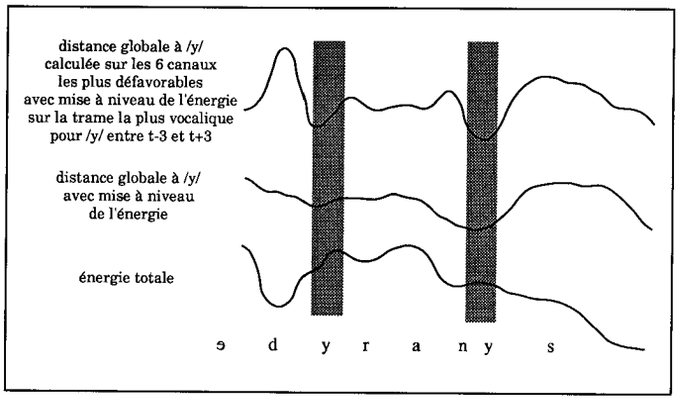

Figure 4. — Illustration des effets du recentrage de la distance autour des trames les plus vocaliques pour la voyelle /y/. La première occurrence de ce phonème, dans le contexte /r/ est décalée par rapport au maximum d'énergie totale alors que pour la seconde le maximum d'énergie totale coïncide avec l'indice pour /y/.

Les distances retenues pour la localisation et l'identification des phonèmes dépend de chacun d'eux. Pour les voyelles ouvertes (/a/, /ɛ/, /ɔ/, /œ/, /ã/, /ɛ̃/, /ɛ̃/, /œ̃/), le maximum d'énergie coïncide avec la zone la plus vocalique qui dépend peu de l'énergie dans les canaux correspondant aux formants. Dans le cas des voyelles très fermées (/i/, /y/, /u/) il est indispensable de prendre en compte de façon précise les zones spectrales des formants et d'effectuer la mise à niveau des spectres par rapport à la trame la plus vocalique pour le phonème traité. Les distances différentielles sont utilisées pour les phonèmes dont les formants sont très bien marqués. Certaines ambiguïtés peuvent être levées par l'emploi conjoint de plusieurs types de distances dont les comportements peuvent être différents suivant le contexte (ex. : l'utilisation des canaux les plus défavorables est indispensable pour la localisation correcte d'un /i/ adjacent à une consonne nasale).

5. Construction du treillis de phonèmes et résultats

La figure 2 illustre la stratégie générale du système pour construire le treillis de phonèmes. Les formes particulières apparaissant sur les courbes qui mesurent la distance aux phonèmes (vallées, zones dont les valeurs sont inférieures à un seuil) constituent les événements de départ qui conduiront à la sélection des unités.

Dans la plupart des systèmes de DAP « ascendant », une phase de segmentation en macro-classes pseudo-phonétiques est mise en œuvre préalablement à l'identification. Les limites de cette technique nous ont conduit à associer les processus de localisation et d'identification ; la segmentation intervient exclusivement pour valider et préciser les choix proposés au moyen des distances.

5.1. INDICES UTILISÉS POUR LA SÉLECTION DES UNITÉS

Les phonèmes qui ont une probabilité d'occurrence non négligeable sur une portion du signal (en fonction des valeurs des distances) n'ont été repérés qu'en fonction de leur ressemblance spectrale et indépendamment des niveaux de l'énergie. Dans certaines situations, une voyelle peut apparaître dans une zone consonantique et réciproquement. Afin de réduire ces anomalies mais sans rejeter des hypothèses correctes, nous avons défini des fonctions dépendantes des phonèmes qui mesurent le caractère vocalique et le caractère consonantique des unités phonétiques.

Dans certains contextes l'examen des courbes d'énergie (énergie moyenne, énergie des basses ou hautes fréquences, etc.) ne suffit pas à localiser les voyelles ou les consonnes. Cependant, si l'on considère, pour chaque phonème, un ensemble de canaux particuliers (correspondant aux formants des voyelles par exemple), l'évolution temporelle de l'énergie située dans ces zones spectrales signale plus précisément les unités (*fig.* 5). Les indices « vocalique » et « consonantique » définis suivant ce principe nous permettent d'une part d'éliminer les portions de signal pour lesquelles l'occurrence d'un phonème est peu vraisemblable et, d'autre part, de préciser l'intervalle sur lequel est centrée l'unité en réalisant une intersection des 2 zones (distances et indices).

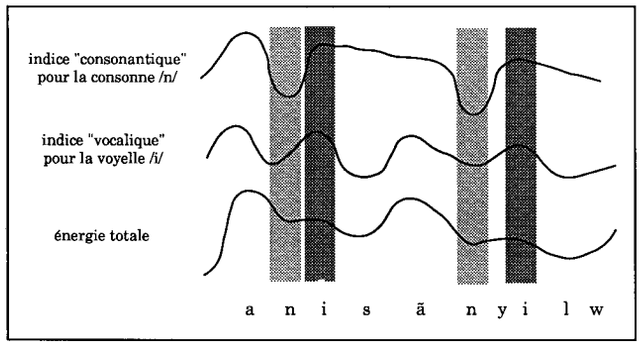

Figure 5. — Les indices « vocalique » et « consonantique » dépendant des unités mettent en évidence et précisent les limites des phonèmes difficiles à localiser.

Dans une séquence de phonèmes semblables du point de vue des indices et malgré la prise en compte des caractéristiques précises du phonème traité, l'unité la plus marquée est mise en évidence. C'est le cas notamment dans les suites de consonnes où la plus fermée est plus facile à repérer que les autres. Aussi, dans le processus d'accès lexical nous devons tenir compte de cette particularité du système et ne considérer que les unités indiscutablement localisables pour effectuer la mise en correspondance d'un item avec une portion du treillis.

5.2. INSTABILITÉ SPECTRALE ET LOCALISATION DE CIBLES PHONÉTIQUES

Le calcul de la variation temporelle des spectres permet de localiser les zones stables et les zones instables du signal qui coïncident généralement avec certaines phases des phonèmes (tenue, burst, transition, etc.) [1], [3], [7], [9], [10], [16], [17], [21]. Toutefois, malgré la simplicité et l'efficacité de cette technique de détermination d'intervalles acoustiquement cohérents, il demeure impossible, dans des conditions naturelles d'élocution, d'effectuer de manière ascendante une mise en correspondance exacte entre les événements acoustiques et les unités phonétiques.

Les différentes distances que nous utilisons localisent des zones de forte ressemblance à un phonème donné. Toutefois, les contraintes contextuelles rendent probables de nombreuses solutions dont la position optimale est décalée par rapport à la partie stable du noyau phonémique ; c'est le cas notamment dans une séquence telle que /k a/ où, sur la première partie du noyau vocalique, la probabilité d'occurrence du phonème /ɛ/ est plus importante que celle de la voyelle /a/ (*fig.* 6). Afin de calculer un score pour chaque hypothèse dans des conditions analogues, la valeur retenue est évaluée au minimum d'instabilité de l'intervalle défini par les distances et les indices. Cette technique n'assure pas — surtout pour les phonèmes transitoires ou dont la tenue est de faible durée — que l'hypothèse phonétique est valuée de façon optimale, mais elle permet de comparer les solutions en un point du signal indépendant du phonème.

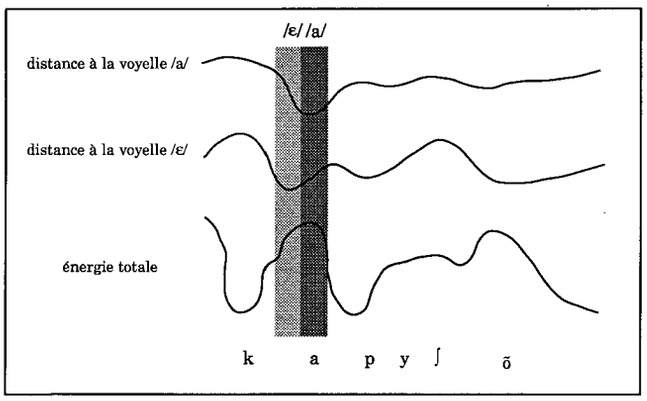

Figure 6. — Les distances aux voyelles /a/ et /ɛ/ donnent des résultats semblables pour la voyelle /a/ précédée de la consonne /k/, mais les minima sont décalés par rapport à la position stable de la voyelle (/a/ coïncide mieux avec le noyau).

Comme pour les distances, le calcul de l'instabilité du signal est effectué soit en comparant directement les canaux de deux trames temporellement décalées, soit en mesurant l'écart de la dérivée des deux spectres. Le choix des canaux utilisés peut dépendre du phonème, mais nous avons retenu une solution qui consiste à ne prendre en compte que les n canaux qui maximisent la distance. La formule utilisée :

$$I_{t,d} = \frac{1}{n} \sum_{j=1}^{n} dX_{t,d,i} \quad \text{avec} \quad i \in [i_1, i_n]$$

$$dX_{td,i_1} \geqslant dX_{t,di_2} \geqslant \cdots \geqslant dX_{t,d,i_{n+1}} \geqslant dX_{t,d,i_{24}}$$

$$dX_{t,d,k} = |Cs_{t+[d|2]+1,k} - Cs_{t-[d|2],k}|$$

mesure la distance sur les n canaux les plus dissemblables entre les spectres situés de part et d'autre de la trame t et décalés de $d+1$ positions. La valeur $d = 2$ donne généralement les meilleurs résultats dans de nombreux contextes ; il est nécessaire d'augmenter le décalage temporel si l'évolution spectrale est particulièrement lente entre deux unités acoustiquement proches.

5.3. TREILLIS PHONÉTIQUE ET RÉSULTATS

Le treillis phonétique est obtenu en mémorisant les unités évaluées de la manière suivante :

— repérage des intervalles dont une distance au phonème est inférieure à un seuil choisi en fonction de l'unité phonétique et des résultats souhaités (les zones sont calculées au moyen des outils de reconnaissance de formes),

— sélection des intervalles cohérents avec le phonème traité du point de vue des indices « vocalique » et « consonantique » (la zone initiale est limitée à son intersection avec celle définie par l'indice),

— calcul du score affecté au phonème par évaluation de la distance globale sur la trame de plus grande stabilité de l'intervalle associé (*fig.* 7).

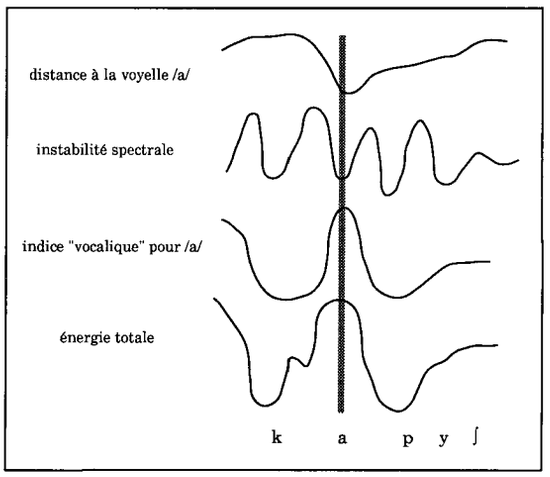

Figure 7. — Choix de la zone sur laquelle est évalué le score d'un phonème.

Les treillis obtenus contiennent généralement tous les phonèmes énoncés dans les phrases si l'on accepte des distances élevées par rapport aux références (*fig.* 8). Les seules unités qui peuvent faire défaut correspondent aux situations où la qualité « consonantique » est difficile à mettre en évidence (ce peut être le cas notamment dans des séquences de plusieurs consonnes où seule la plus fermée est assurée d'être localisée ; ex. : /t/ dans /t r w/).

Compte tenu de la simplicité du système (apprentissage très réduit, pas de prise en compte du contexte), les scores

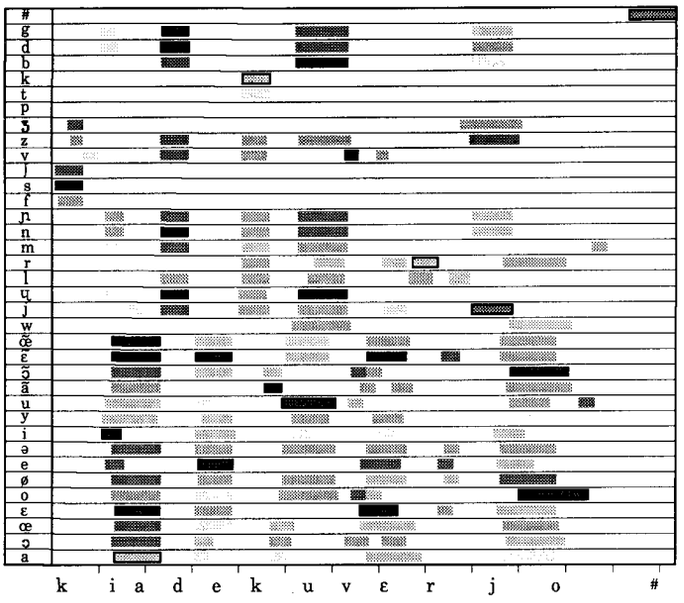

Figure 8. — Treillis des phonèmes correspondant à la phrase : « Qui a découvert Io ? ». Le score des phonèmes est matérialisé par le caractère sombre de leur représentation.

obtenus pour les phonèmes effectivement énoncés sont généralement bons (table 1) ; l'unité attendue apparaît fréquemment en tête surtout si le contexte est peu déformant ou si la syllabe est accentuée.

Les treillis produits ont été utilisés d'une part pour la reconnaissance de la parole continue dans un univers restreint (interrogation d'une base de données sur les planètes) et d'autre part, pour la sélection de mots d'un grand vocabulaire énoncés isolément. Dans le premier cas les tests ont été effectués pour 2 locuteurs qui ont prononcé une centaine de phrases interrogatives construites au moyen d'une grammaire simple et d'un lexique limité à 400 mots. En un point particulier de l'analyse d'un énoncé, le nombre de mots probables peut être important, surtout lorsqu'il s'agit de noms propres (ex. : « Qu'a découvert X ? », ou X est le nom d'un des savants répertoriés dans la base de données). Dans le cas de notre test, l'algorithme de parcours de l'arbre des solutions — qui est fondé sur le principe de « shortfall and density scoring » [22] — propose la solution correcte en première position pour plus de 90 % des phrases. Ces résultats intéressants sont en partie la conséquence des fortes contraintes linguistiques et du nombre limité de mots acoustiquement proches à chaque branchement. Une plus grande complexité de la grammaire et un lexique à la fois plus riche et plus difficile nécessiteront une phase descendante de DAP afin de discriminer plus précisément les cohortes de mots candidats.

Table 1

Résultats de l'identification ascendante des voyelles. Les tests sont effectués à partir d'une base de sons étiquetés dans des phrases ou des mots. Les pourcentages de reconnaissance correspondent à la valeur du score du phonème par rapport aux autres candidats pour les 5 premiers choix.

voyelle	% 1ère position	% 2ème position	% 3ème position	% 4ème position	% 5ème position	% trouvé
a	41	50	55	61	71	100
ɔ	40	62	72	100	100	100
ɛ	67	87	92	94	96	100
œ	52	68	78	86	97	100
o	76	92	100	100	100	100
e	77	87	89	91	94	100
ø	75	97	100	100	100	100
i	74	89	90	92	96	100
y	94	94	94	97	97	100
u	62	92	97	97	97	100
ɑ̃	62	62	72	86	86	100
ɔ̃	46	84	93	95	97	100
ɛ̃	51	82	84	92	92	100
œ̃	39	68	82	92	92	100

Pour le second système, le lexique est constitué d'un millier de mots d'un dictionnaire de l'informatique dont chacun est décrit symboliquement au moyen de la séquence des phonèmes qui le composent. Les tests ont porté sur des énonciations isolées (d'éléments pris au hasard dans ce vocabulaire) effectuées par 4 locuteurs pour lesquels nous avons mémorisé les références spectrales des unités phonétiques. Le mot prononcé apparaît toujours en bonne place dans une cohorte restreinte à 10 éléments du lexique. Le processus d'identification est fondé sur un algorithme qui prend en compte les scores des phonèmes candidats ainsi que le taux de recouvrement du signal de parole. Malgré un taux de reconnaissance important en première position, une étape descendante du DAP est nécessaire pour classer plus finement les hypothèses surtout dans le cas de grands vocabulaires.

6. Conclusion

Le système de DAP « ascendant » que nous avons développé a permis de résoudre simplement et efficacement le problème de la construction d'un treillis phonétique contenant toutes les unités énoncées. L'adaptation au locuteur est effectuée très rapidement car les techniques employées ne nécessitent pas une grande quantité de données d'apprentissage. Toutefois, les méthodes plus sophistiquées de classification (Quantification Vectorielle, Réseaux Neuro-Mimétiques, etc.) permettraient d'affiner les résultats au détriment d'une plus grande lourdeur de mise en œuvre.

Les résultats obtenus sont intéressants et parfaitement utilisables dans des applications telles que l'identification de mots isolés ou la reconnaissance de la parole continue. Pour compléter le système de DAP, il convient cependant de réaliser une phase « descendante » du processus qui devra limiter et classer précisément les solutions concurrentes dans un contexte connu. Quelle que soit la méthode employée pour cette opération (connaissances explicites, HMM, RNM, etc.) les processus retenus devront être indépendants du locuteur ou très simplement adaptables.

Manuscrit reçu le 2 octobre 1990.

BIBLIOGRAPHIE

[1] C. Abry, D. Autesserre, C. Barrera, C. Benoit, L.-J. Boe, J. Caelen, G. Caelen-Haumont, M. Rossi, R. Sock, N. Vigouroux (1985), « Propositions pour la segmentation et l'étiquetage de la base de données des sons du français » ; *XIV^e JEP,* GALF, Paris, pp. 156-153.

[2] T. H. Applebaum, A. H. Hanson, H. Wakita (1987), « Weighted distance measures in vector quantization based speech recognizers » ; *Proceedings ICASSP,* pp. 1155-1158.

[3] J. S. Bridle, R. M. Chamberlain (1983), « Automatic labelling of speech using synthesis-by-rule and non-linear time-alignment » ; *Speech Com.,* Vol. 2, n° 2-3, pp. 187-189.

[4] R. Bulot (1987), « Techniques d'Intelligence Artificielle pour la reconnaissance de la parole, application au décodage acousticophonétique » ; *Thèse de l'université d'Aix-Marseille II.*

[5] R. Bulot, P. Nocera (1989), « A system for speech recognition using both connexionist methods and Knowledge » ; *Proceedings European Conference on Speech Recognition,* Paris, pp. 533-536.

[6] J. Caelen, G. Caelen-Haumont (1981), « Indices et propriétés dans le projet ARIAL II » ; *Processus d'Encodage et de Décodage Phonétique,* GALF, Toulouse, pp. 128-139.

[7] J. Caelen (1985), « Introduction à une segmentation cinématique » ; *XIV^e JEP,* GALF, Paris, pp. 1129-132.

[8] Calliope (1989), *La parole et son traitement automatique,* Masson.

[9] D. Fohr, J.-P. Haton, F. Lonchamp, L. Sauter (1985), « Méthodes de segmentation syllabique en reconnaissance de la parole » : *XIV^e JEP,* GALF, Paris, pp. 164-167.

[10] O. Fujimura (1981), « Temporal organization of articulatory movements as a multidimentional phasal structure » ; *Phonetica,* Vol. 38, pp. 66-83.

[11] A. H. Jr. Gray, J. D. Markel (1976), « Distance Measures for Speech Processing » ; *IEEE Trans. Acoust. Speech and Signal Proc.,* Vol. *ASSP 24,* n° 5.

[12] Y. Gong, F. Mouria, J. P. Haton (1989), « Un système de reconnaissance de la parole continue sans segmentation » ; *IEEE Trans. Acoust. Speech and Signal Proc.,* Vol. *ASSP 24,* n° 5.

[13] F. Itakura, T. Umesaki (1987), « Distance measure for speech recognition based on the smoothed group delay spectrum » ; Actes du 7^e Congrès AFCET RFIA, Paris, novembre 1989.

[14] D.-H. Klatt (1979), « Speech perception : a model of acoustic-phonetic access » ; *J. Phonetics 7,* pp. 279-312.

[15] W.-A. Lea, J. E. Shoup (1980), « Contributions of the ARPA-SUR project » ; in : *Trends in Speech Recognition,* Prentice Hall.

[16] J. S. Lienard, M. Mlouka (1972), « Segmentation automatique de la parole en phonatomes » ; *III^e JEP,* GALF, Lannion, pp. 347-355.

[17] H. Meloni (1982), « Étude et réalisation d'un système de reconnaissance automatique de la parole continue » ; *Thèse de doctorat d'État,* Aix-Marseille II.

[18] H. Meloni, R. Bulot (1986), « Un système de traitement de connaissances pour le décodage acoustico-phonétique » ; *ICA Symposium on speech recognition,* Mc Gill University.

[19] H. Meloni, P. Gilles, A. Betari (1989), « A Knowledge-Based System for Speaker-Independent Recognition of Letters » ; *Proceedings European Conference on Speech Recognition,* Paris, pp. 625-628.

[20] N. Nocerino, F. K. Soong, L. R. Rabiner, D. H. Klatt (1985), « Comparative study of several distorsion measures for speech recognition » ; *Proceedings ICASSP,* pp. 25-28.

[21] G. Perennou, De Calmes M. (1985), « Segmentation en événements phonétiques et en unités syllabiques » ; *XIV^e JEP,* GALF, Paris, pp. 142-146.

[22] W. A. Woods (1982), « Optimal search strategies for speech understanding control » ; *Artificial Intelligence 18,* pp. 295-326.

[23] V. W. Zue (1983), « The use of phonetic rules in automatic speech recognition » ; *Speech Com.,* vol. 2, n° 2-3, pp. 181-186.