Utilisation d'un modèle d'audition et de connaissances phonétiques en reconnaissance automatique de la parole

On the use of an auditory model and phonetic knowledge for automatic speech recognition

Jean-Claude JUNQUA
CRIN/INRIA
Campus scientifique
BP 239, 54506 Vandœuvre Lès Nancy, France
et Speech Technology Laboratory
3888 State Street, Santa Barbara, California 93105

Jean-Claude JUNQUA est ingénieur diplômé de l'École Nationale Supérieure d'Électricité et de Mécanique de Nancy (ENSEM) en électronique et automatique et de l'Université de Nancy I où il a obtenu le titre de docteur d'Université en informatique. De 1981 à 1986 il a assuré la responsabilité technique du Centre de Recherche en Informatique de Nancy (CRIN) et a participé aux travaux du groupe *Reconnaissance des Formes et Intelligence Artificielle* (RFIA). De 1987 à 1988 il a effectué des travaux de recherche sur la reconnaissance automatique de la parole dans le laboratoire Speech Technology Laboratory situé à Santa Barbara, Californie. Présentement en détachement du CNRS et Ingénieur de Recherche à Speech Technology Laboratory, il assure la responsabilité de plusieurs projets liés à l'amélioration de la robustesse des systèmes de reconnaissance automatique de la parole, l'étude de l'effet Lombard, et la réalisation d'environnements logiciels destinés à faciliter le traitement automatique de la parole.

RÉSUMÉ

L'introduction de connaissances dans les systèmes de reconnaissance de parole (RAP) est un bon moyen d'améliorer les performances des systèmes actuels. Dans cet article nous proposons le système ORION dans le cadre d'une application de reconnaissance multilocuteur de mots isolés. ORION est un système hybride à deux passes intégrant plusieurs sources de connaissances : psychoacoustiques, physiologiques et phonétiques. Pendant la première passe un modèle d'analyse acoustique perceptivement fondé (PLP), combinant des caractéristiques instantanées et des caractéristiques spectrales dynamiques, est utilisé pour fournir des vecteurs de paramètres à un algorithme de programmation dynamique. A l'issue de cette première passe plus de 98 % de mots ont été correctement reconnus pour un vocabulaire de chiffres et 12 références par mot. L'introduction de connaissances phonétiques durant la deuxième passe diminue l'erreur de reconnaissance de plus de 60 % (par rapport aux résultats obtenus lors de la première passe) pour un vocabulaire de mots acoustiquement similaires (E-SET).

MOTS CLÉS

Reconnaissance automatique, multilocuteur, mots isolés, discrimination, physiologie, psychoacoustique, phonétique, système hybride, mesure de distance, caractéristiques spectrales dynamiques, modèle d'analyse acoustique.

SUMMARY

Including speech knowledge in automatic speech recognition (ASR) systems is a good way to improve the performance of recognizers. In this paper, we propose the ORION system which deals with speaker-independent ASR for isolated-words. ORION is a two-pass hybrid system which uses several types of knowledge. This knowledge applies to psychoacoustics, physiology and phonetics. During the first pass an auditory model, PLP (perceptually-based linear prediction analysis) combines static and dynamic features to provide a set of parameters to the dynamic programming algorithm. After this stage 98 % recognition accuracy was obtained for a digit vocabulary and 12 templates per word. The introduction of phonetic knowledge in the second pass decreases the error rate by more than 60 % (compared to the results of the first pass) for a confusable vocabulary (E-SET).

KEY WORDS

Automatic speech recognition, speaker-independent, isolated-words, discrimination, physiology, psychoacoustics, phonetics, hybrid system, distance measure, dynamic features, auditory model.

1. Introduction

Deux grandes approches ont été proposées pour construire des systèmes de reconnaissance automatique de la parole. La première est fondée sur une représentation paramétrique de la parole qui utilise des mesures faites sur le signal de parole. Cette phase est en général suivie d'un algorithme de reconnaissance des formes (du type de la programmation dynamique) afin d'effectuer la classification. Pour déterminer la représentation paramétrique, des modèles auditifs ont récemment reçu beaucoup d'attention. Bien que notre connaissance des phénomènes liés à la perception de la parole soit très limitée, il a été montré qu'un système qui modélise des propriétés du système auditif humain pouvait générer une meilleure représentation de la parole que les systèmes traditionnels [19, 22, 26, 13, 1, 5, 11, 12, 35]. La deuxième approche utilise un langage de représentation pour décrire les unités acoustiques [6, 37]. Une telle approche utilise intensivement les techniques d'intelligence artificielle. Dans ce cas, le principal problème est l'acquisition et la représentation de la connaissance. Les deux approches mentionnées diffèrent par le niveau des connaissances prises en considération. Dans la première approche (utilisant une méthode de reconnaissance globale), les connaissances sont essentiellement incluses dans la modélisation acoustique de la parole alors que dans la deuxième (reconnaissance à l'aide d'indices acoustiques) les connaissances sont exprimées par l'intermédiaire du langage de représentation et des unités acoustiques utilisées.

La reconnaissance automatique fondée sur l'extraction d'indices acoustiques a été proposée comme une alternative à la reconnaissance de formes [9]. Un système de reconnaissance à base d'indices est généralement constitué de deux étapes :

- identification des indices distinctifs à partir d'une représentation acoustique, généralement un spectrogramme,
- identification des unités lexicales à partir des indices acoustiques.

La deuxième étape est généralement réalisée à partir d'un système à base de règles de production.

Les systèmes de reconnaissance automatique actuels ont beaucoup de mal à tenir compte des caractéristiques *spécifiques* du signal de parole. Il faut cependant noter que, comme notre connaissance des mécanismes liés à la parole est très limitée, les systèmes de reconnaissance ne doivent aussi pas « oublier » de prendre en compte notre *ignorance* [28]. Un exemple de modèle « d'ignorance » est fourni par l'algorithme de programmation dynamique qui modélise notre ignorance à propos des variations de la parole dans l'espace temps. Un tel algorithme tient compte de tous les indices acoustiques même ceux qui ne sont pas pertinents. Il donne aussi la même importance à tous les détails acoustiques même si les transitions sont perceptuellement plus importantes que les parties stables du signal [17]. L'introduction de connaissances sur la parole dans les systèmes de reconnaissance automatique de la parole (*RAP*) est un bon moyen d'améliorer les performances des systèmes actuels. Toutefois, ces connaissances doivent être introduites avec précaution. Dans le même ordre d'idées, il a été montré que des systèmes hybrides, utilisant des modèles mathématiques et des connaissances sur la parole, permettaient d'améliorer les scores de reconnaissance et aidaient à triompher des limitations associées aux systèmes traditionnels [3].

L'approche proposée ici est fondée sur l'utilisation de plusieurs sources de connaissances dans un système hybride à deux passes, appelé *ORION*, qui utilise des connaissances phonétiques pendant la deuxième passe. Les connaissances phonétiques facilitent la *discrimination* des mots appartenant à des classes du vocabulaire constituées de mots acoustiquement similaires. Avec les méthodes traditionnelles, des différences mineures mais de longue durée peuvent l'emporter sur des différences majeures et ainsi contribuer à de mauvaises décisions. De façon générale les méthodes discriminantes orientent la reconnaissance sur les parties des mots qui sont différentes acoustiquement. Par définition, les informations discriminantes sont présentes uniquement dans une partie du mot. De telles méthodes ont déjà permis d'améliorer les performances obtenues. Citons en particulier le système à deux passes de Rabiner et Wilpon [33] qui utilise une fonction de pondération statistique donnant plus d'importance à la partie du mot facilitant la discrimination. Une variante de cette méthode a aussi été proposée par Casacuberta *et al.* [7] qui n'ont utilisé aucune hypothèse sur la distribution statistique sous-jacente des données manipulées. Lamel et Zue [25] proposèrent une méthode donnant plus d'importance aux transitions consonne-voyelle après avoir séparé le vocabulaire en sous-classes grâce au contour de l'énergie calculée pour les basses fréquences. Enfin, Cole *et al.* [8], et Bradshaw *et al.* [3] proposèrent des méthodes utilisant des indices phonétiques. Les résultats obtenus par ces différentes méthodes ne sont toutefois pas encore satisfaisants. Un des buts du système proposé dans cet article est, comme tous les systèmes cités précédemment, d'améliorer les performances obtenues pour des vocabulaires de mots acoustiquement similaires. Du fait de sa structure hybride il se rapproche davantage du système proposé par Bradshaw *et al.* même si les techniques utilisées sont très différentes. *ORION* utilise un modèle d'analyse acoustique qui simule

des propriétés du système auditif humain et accentue l'importance donnée aux transitions. De plus, il bénéficie de modèles perceptuels à la fois dans la partie analyse acoustique et dans l'extraction d'indices.

Le développement de ce système est né de deux idées. La première est venue de notre incapacité, avec les systèmes de comparaison de formes, à améliorer les scores de reconnaissance pour des vocabulaires difficiles. Même en donnant plus d'importance aux transitions, pour un vocabulaire de mots anglais comme le « *E-SET* = {B, C, D, E, G, P, T, V, Z, FEED} » (extrait de la base de données *D*1 définie à la section 3.2), notre meilleur système (utilisant un algorithme de programmation dynamique) a obtenu 68 % de mots correctement reconnus (en reconnaissance multilocuteur avec neuf références par mot). Ceci dirigea nos recherches vers le développement d'un système pouvant traiter de façon particulière les vocabulaires difficiles sans perdre les avantages d'un algorithme de comparaison de formes pour les autres mots du vocabulaire. La deuxième idée est liée à l'évolution rapide des systèmes de reconnaissance. De nouvelles méthodes d'analyse, mesures de distance, indices, sont souvent testés et comparés. Aussi, un outil de recherche, permettant une grande souplesse sans perdre les avantages d'un système de reconnaissance intégré, nous apparut nécessaire. Ces considérations ont conduit au développement du système ORION qui est décrit dans la section suivante.

2. Vue générale du système

La figure 1 montre un diagramme du système réalisé.

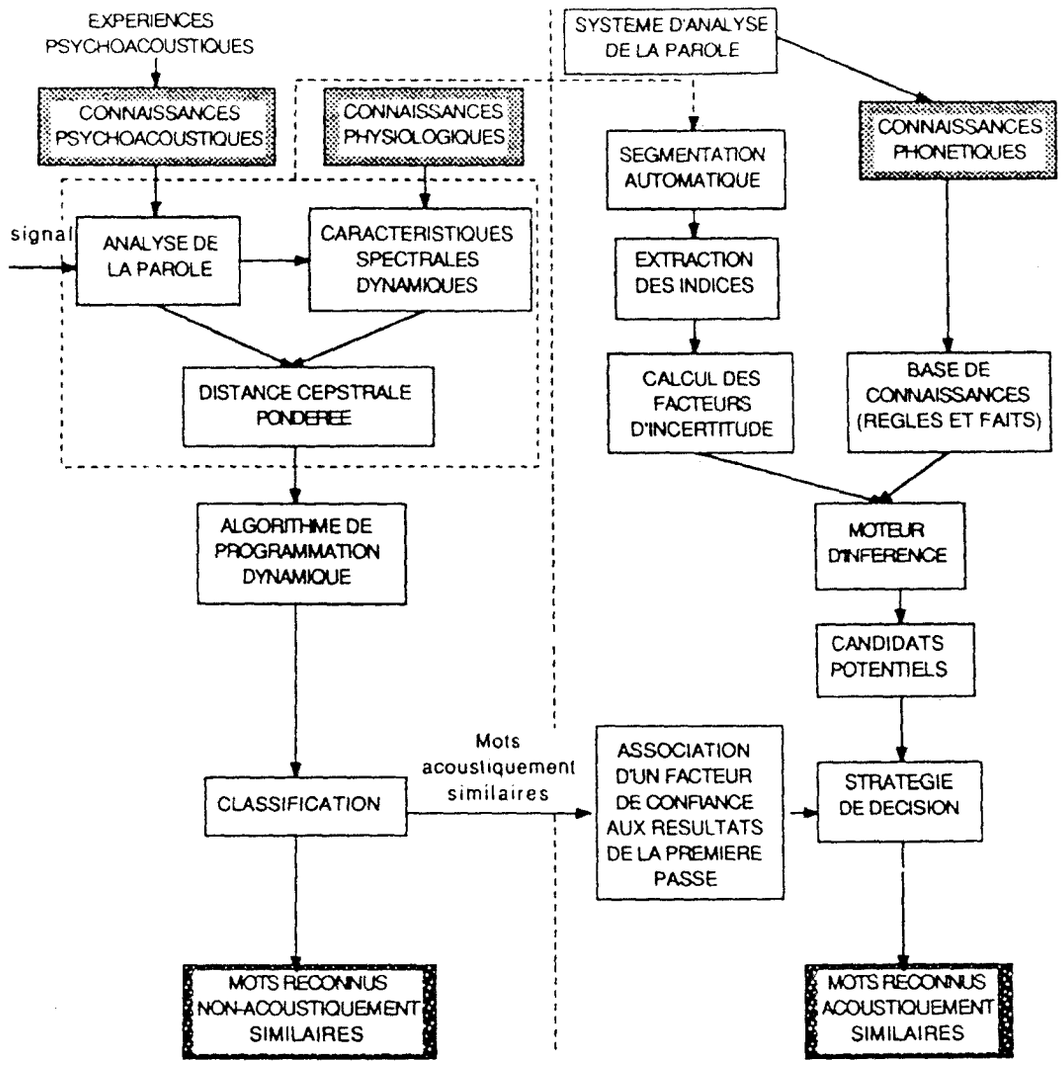

Figure 1. — Diagramme fonctionnel du système hybride ORION (l'entrée du signal de parole se fait au niveau de la chaîne de gauche).

Pendant la première passe, un algorithme de programmation dynamique utilise le modèle d'analyse acoustique perceptivement fondé *PLP* (en abréviation de « perceptually-based linear prediction analysis ») [20] pour générer des coefficients cepstraux du spectre auditif. Ces coefficients cepstraux sont obtenus grâce à une approximation du spectre de fréquence fourni par l'analyse *PLP* (leur calcul sera détaillé dans la prochaine section). Ces coefficients sont ensuite combinés avec d'autres coefficients, modélisant des caractéristiques spectrales dynamiques du signal d'entrée, pour former un vecteur représentant un segment de parole. L'ensemble des vecteurs, représentant un mot, est ensuite traité par un algorithme de programmation dynamique qui utilise une distance cepstrale pondérée. La première passe fournit le mot reconnu pour les mots du vocabulaire qui n'appartiennent pas à une classe de mots confondus facilement. Les classes de mots confondus facilement (ou difficiles à reconnaître) ont été identifiées à l'aide de matrices de confusion fournies par des tests préliminaires. Ces matrices de confusion ont aussi été utilisées pour pondérer par un facteur de confiance les candidats générés par l'algorithme de programmation dynamique. Une deuxième passe est invoquée si les premiers candidats fournis par l'algorithme de programmation dynamique appartiennent à une classe de mots du vocabulaire étudié définie comme difficile à reconnaître. Dans cette étude, les efforts ne se sont pas portés sur l'algorithme de classification. Cependant, des tests préliminaires ont montré que, pour le vocabulaire étudié, la méthode choisie suffisait à résoudre le problème posé par la classification. Des indices temporels et fréquentiels sont alors extraits automatiquement avant d'être fournis à un moteur d'inférence. Grâce à une base de connaissances construite au préalable, le moteur d'inférence génère des candidats pondérés par un facteur de confiance. Une stratégie de décision, prenant en compte les résultats donnés par le moteur d'inférence et ceux générés par l'algorithme de programmation dynamique, est alors chargée de fournir la décision finale.

Avant d'extraire les indices temporels et fréquentiels, un algorithme de segmentation automatique est appliqué afin de déterminer les différents phonèmes du mot analysé. Ceci permet d'extraire la plupart des indices dans la partie du mot qui permet de le discriminer avec les autres mots de la même classe.

Le système a été développé de façon modulaire afin de permettre une souplesse importante pour tester de nouveaux algorithmes. Notre connaissance (phonétique, du système auditif, etc.) n'est pas statique, pas plus que ne l'est notre compréhension des techniques à introduire dans les systèmes de reconnaissance. Le système développé est organisé autour d'un ensemble de modules. Ceci permet de se concentrer sur une partie du système sans avoir à redévelopper un nouveau système chaque fois qu'une modification doit être faite. De plus, lors des tests, le système complet ou une partie du système seulement peut être évalué. Un tel système est un outil de recherche qui permet une importante souplesse. Il s'est avéré très utile dans les études réalisées.

3. Le modèle PLP optimisé

3.1. PRÉLIMINAIRES

Tout système de *RAP* utilisant des formes de référence emploie un module qui compare des segments de parole en terme de similitude ou dissimilitude. Ce module est divisé en deux parties : le *module d'analyse,* qui est chargé d'extraire un ensemble de paramètres à partir du segment de parole, et le *module qui contient la mesure de distance.* Ces deux modules interagissent. Ils ne doivent donc pas être étudiés séparément.

La technique d'analyse *PLP* modélise un spectre auditif par un modèle tout pôle d'ordre réduit en utilisant la technique d'autocorrélation de la prédiction linéaire. Comme l'indique la figure 2, *PLP* diffère de l'analyse standard *LPC* par une intégration en bandes critiques du spectre de puissance [15], une pré-accentuation par des courbes d'isosonie [34], et une conversion d'intensité en

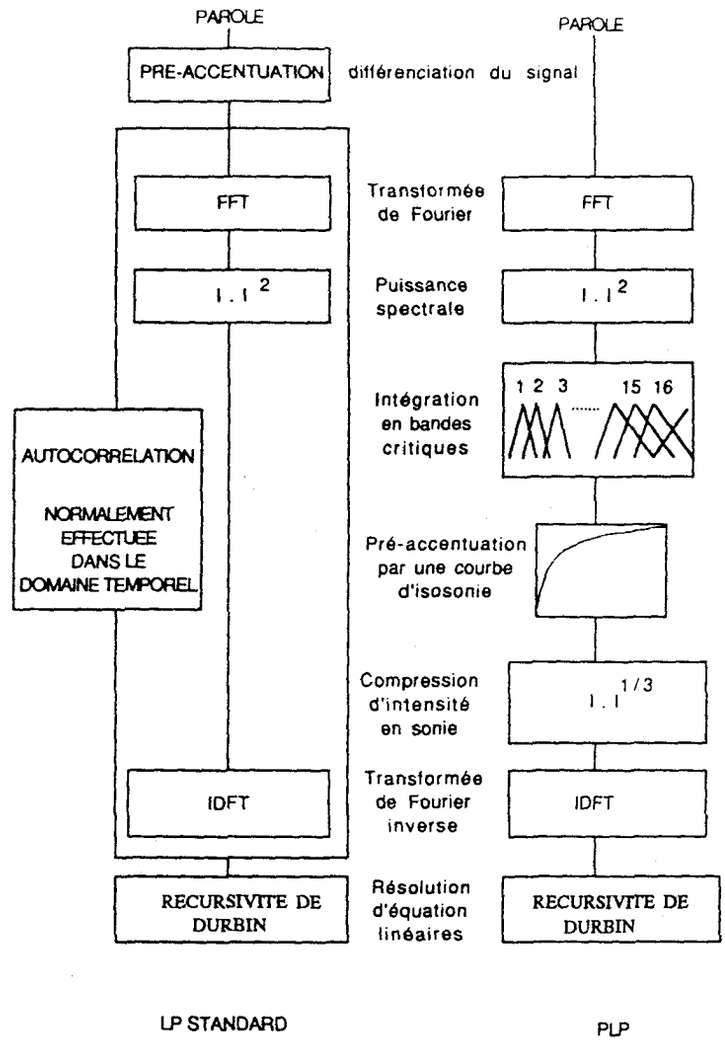

Figure 2. — Diagramme fonctionnel de la technique d'analyse PLP.

sonie [36]. Ces différentes étapes, qui simulent des concepts psychoacoustiques bien établis, sont suivies d'une modélisation par la fonction d'un modèle tout pôle d'ordre réduit qui fournit une représentation compacte de la forme du spectre auditif en terme de pôles. L'avantage de cette fonction est de mettre l'accent, lors de la modélisation, sur les pics du spectre de fréquence.

La compatibilité des paramètres obtenus avec cette méthode d'analyse et la technique de prédiction linéaire est très utile au niveau des applications pratiques.

Dans ce travail, 17 filtres à bandes critiques ont été utilisés sur une gamme de fréquence couvrant 17 Barks. Ces filtres ont été simulés en intégrant le spectre de fréquence produit par une *FFT* du signal de parole multiplié par une fenêtre de Hamming. La taille des trames de parole a été fixée à 10 ms. Le spectre obtenu à l'aide de l'analyse *PLP* a été approché par M coefficients cepstraux déduits d'un modèle tout pôle d'ordre M. Dans tous les tests, un algorithme de programmation dynamique, utilisant une contrainte locale symétrique et une contrainte de pente d'ordre 2 [30], a permis d'effectuer les comparaisons entre les mots de test et les mots de référence.

3.2. CONDITIONS EXPÉRIMENTALES

Dans le cadre de ce travail les bases de données suivantes ont été utilisées :

— *base de données D1* : 104 mots habituellement employés pour définir les touches d'un clavier. Les mots ont été produits dans un environnement non bruité. 10 locuteurs (6 hommes et 4 femmes) et 3 répétitions ont été considérés. Cette base de données est appelée « keyboard » ;

— *base de données D2* : un sous-ensemble de la base de donnée *D1* constitué de classes de mots acoustiquement similaires {B, D, G, P, T, V, C, E, FEED, F, S, X, LINE, NINE, ONE} ;

— *base de données D3* : vocabulaire des chiffres produits en environnement non bruité par des locuteurs de dialectes différents. 96 locuteurs et une répétition (48 hommes et 48 femmes) ont été considérés.

Tous les mots ont été enregistrés isolément (à une fréquence d'acquisition de 10 kHz) et les frontières de mots ont été déterminées manuellement. Aucune technique de classification ou procédure spéciale de sélection des mots de référence n'a été utilisée.

3.3. ÉVALUATION ET OPTIMISATION DU MODÈLE PLP

Des études préliminaires [19] ont suggéré, à partir d'une étude comparative, que les meilleures performances (pour les modèles étudiés), en *RAP* multilocuteur de mots isolés, étaient obtenues à l'aide d'un modèle d'ordre réduit utilisant l'analyse *PLP*. Dans le but de valider cette hypothèse (avec une base de données plus importante), une série de tests a été menée en reconnaissance multilocuteur. Dans ces tests l'ordre du modèle d'analyse a varié de cinq à quatorze. Cinq modèles d'analyse acoustique ont été comparés et la base de données *D1* a été utilisée. Trois types d'analyse : *LP, PLP* et banc de filtres critiques (extrait de l'analyse *PLP*), et deux types de distance : Euclidienne et *RPS* (distance Euclidienne sur des coefficients cepstraux pondérés par leur index) [31, 18] ont été considérées. Enfin, la paramétrisation cepstrale, qui a déjà donné de bons résultats en reconnaissance de la parole comparée à d'autres types de représentation [2, 22], a été retenue (sauf pour l'analyse par banc de filtres qui utilise une représentation de la pente spectrale [24]).

Chaque test de reconnaissance a utilisé les vocabulaires de deux locuteurs (un homme et une femme) comme ensemble de référence et les vocabulaires des autres locuteurs comme ensemble de test. Différentes combinaisons ont été envisagées ce qui a conduit a plus de 15 000 comparaisons pour chaque test. La figure 3 présente les résultats obtenus (le modèle CB_SM représente le modèle utilisant l'analyse par banc de filtres critiques et est similaire au « critical-band slope metric » défini par Klatt [24]).

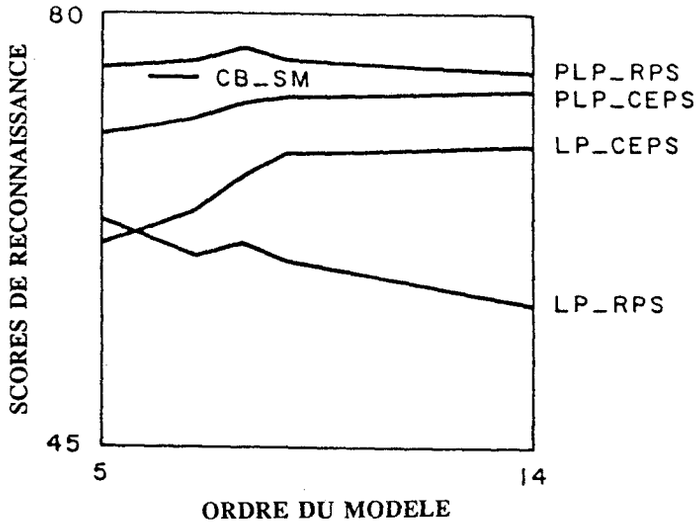

Figure 3. — Comparaison de plusieurs modèles d'analyse acoustique en reconnaissance multilocuteur.

Une augmentation de l'ordre du modèle *PLP_RPS* au-delà de l'ordre huit diminue les scores de reconnaissance. De manière plus générale, la combinaison d'un modèle d'ordre élevé et de la distance *RPS* diminue les performances. Dans notre étude, il a été remarqué que la distance *RPS* est particulièrement sensible aux variations de pentes survenant autour des pics du spectre de fréquence. Ainsi, un ordre du modèle trop élevé entraîne une accentuation des détails du spectre de fréquence, en particulier ceux qui fournissent des informations parasites. Le fait que l'analyse *PLP* fournisse un spectre de fréquence plus lisse que celui rendu par l'analyse *LP* semble être une des causes associées aux meilleures performances obtenues avec le modèle *PLP_RPS*.

Observant que la plupart des erreurs obtenues avec le modèle *PLP* sont au niveau des consonnes, des caractéristiques spectrales dynamiques ont été introduites afin de donner plus d'importance aux transitions. En effet, il a été montré, lors de tests physiologiques, que les transitions jouaient un rôle important dans la perception de la parole [10]. Ces caractéristiques spectrales dynamiques ont été représentées par l'intermédiaire de coefficients de régression [17] obtenus grâce à une combinaison linéaire des coefficients *PLP_RPS*. Les mêmes tests que ceux rapportés précédemment ont été effectués. Le nouveau modèle d'analyse acoustique, utilisant à la fois des caractéristiques instantanées et dynamiques (8 coefficients instantanés et 4 coefficients de régression donnent les meilleurs résultats), a diminué l'erreur de reconnaissance de 6,6 %.

Enfin, la mesure de distance a été optimisée [21]. Une nouvelle mesure de distance pondérant chaque coefficient cepstral par une puissance de son index a permis de s'affranchir de certaines des limitations de la distance *RPS*. Cette nouvelle mesure de distance est définie par :

$$E_n = n^S S \geq 0 . \quad (1)$$

où n représente l'index du coefficient cepstral et S une variable permettant d'accentuer l'influence des pics spectraux et de la pente du spectre de fréquence dans le calcul de la mesure de distance. Ce modèle optimisé combinant les caractéristiques spectrales dynamiques et la nouvelle mesure de distance (avec S = 0,6) a diminué l'erreur de reconnaissance de 8,5 % par comparaison à l'erreur de reconnaissance obtenue avec le modèle *PLP_RPS*. Enfin, des tests complémentaires sur la base de données *D3* utilisant ce modèle optimisé (d'ordre 5), 12 références par mots et 48 locuteurs de test ont permis d'obtenir des scores de reconnaissance supérieurs à 98 % (moyenne pour les différents locuteurs de test), soit 10 % de diminution du taux d'erreur de reconnaissance par rapport au modèle *PLP_RPS*.

C'est ce modèle optimisé qui est utilisé dans le système *ORION*.

4. Acquisition et représentation des connaissances phonétiques

4.1. UTILISATION DE DISTINCTIONS PHONÉTIQUES

En regardant de plus près les erreurs de reconnaissance des précédentes évaluations utilisant le modèle d'analyse acoustique *PLP*, quatre sous-groupes ou classes (appartenant à la base de données *D1*) de mots confondus facilement ont été identifiés (base de données *D2*) : *E-SET* = {B, C, D, E, G, P, T, V, Z, FEED}, {M, N}, {F, S, X} et {LINE, NINE}. Lorsqu'un mot appartenant à une de ces classes est mal reconnu, il a été observé que les premiers mots reconnus appartiennent à la même classe. L'étude présentée dans ce chapitre s'est intéressée à la classe *E-SET* mais les conclusions peuvent être généralisées aux autres classes définies précédemment.

Le modèle *PLP* optimisé fournit un score de reconnaissance de 68 % en reconnaissance multilocuteur (avec 9 références par mot) sur la classe *E-SET*. Le problème rencontré avec l'algorithme de programmation dynamique et plus généralement dans les algorithmes de comparaison de formes est que toutes les parties du mot testé ont le même poids pendant la reconnaissance. Le mot « FEED » est quelquefois confondu avec les autres mots de la même classe alors qu'il n'y a pas de confusion possible en regardant un spectrogramme de ce mot (un indice distinctif est, par exemple, la présence très fréquente d'une période de silence et d'une barre d'explosion à cause du phonème /*d*/).

Des techniques de discrimination ont déjà été proposées pour améliorer la reconnaissance de mots acoustiquement similaires [33, 7, 8, 3, 29]. Dans l'approche proposée, afin d'effectuer la discrimination, le système de reconnaissance invoque une deuxième passe dont le but est d'extraire des indices acoustiques distinctifs menant à une identification correcte. Les mots sont décrits en terme de propriétés acoustiques. Ces propriétés ont été déterminées par l'étude visuelle de spectrogrammes traditionnels et de spectrogrammes auditifs (ou pseudo-spectrogrammes) obtenus à partir de l'analyse *PLP*. Ces spectrogrammes ont été visualisés et plus généralement « édités » à l'aide du système d'analyse de spectrogrammes *STAR* [23] qui a été développé dans le cadre du système *ORION*.

Dans cette étude, les avantages d'un système utilisant des indices acoustiques comme unités principales de reconnaissance, pour distinguer des mots acoustiquement similaires, sont présentés. Le principal niveau de variabilité pris en compte est celui qui est dû à la différence de morphologie entre les locuteurs. En effet, le but de cette étude est la reconnaissance de mots isolés et de ce fait les différences de style et de vitesse d'élocution ne sont pas très importantes. De plus, le vocabulaire choisi (*E-SET* dans un premier temps) limite beaucoup l'influence du contexte.

4.2. DÉFINITION ET EXTRACTION DES INDICES DISCRIMINANTS

Afin de ne pas engendrer des calculs trop importants, des indices grossiers ont été sélectionnés. Grâce à l'observation de spectrogrammes et pseudo-spectrogrammes neuf indices de base ont été considérés :

- présence d'une barre d'explosion accompagnée de ses caractéristiques (force, position, etc.),
- durée de la partie non voisée au début de chaque mot,
- présence d'une barre de voisement,
- présence d'une période de silence (à l'intérieur du mot) accompagnée de ses caractéristiques (durée, position),
- mouvements du deuxième et troisième formant calculés au début de la partie voisée de chaque mot à l'aide de l'analyse *LPC* d'ordre quatorze,
- énergie dans certaines bandes de fréquence avant le début de la voyelle,
- durée de la partie consonne et de l'autre partie du mot,

• taux de passages par zéro,
• mouvements des deux pics spectraux obtenus par le modèle *PLP* d'ordre cinq.

Les indices concernant le voisement et l'extraction des formants et des pics spectraux du modèle *PLP* (après calcul des coefficients *PARCOR*) ont été calculés à l'aide des fonctions se trouvant dans le logiciel de traitement de signal *ILS*. Enfin l'extraction du burst (détecté, lorsqu'il est effectivement présent dans nos expériences, à 85 %) et de l'énergie dans diverses bandes de fréquence a été réalisé à partir de spectrogrammes digitaux du signal de parole analysé.

En ce qui concerne la classe *E-SET*, l'information importante se trouve au début des mots. Afin d'être pertinents les indices qui ont été décrits nécessitent une segmentation, au préalable, des mots en consonnes et voyelles. Par conséquent, dans le but de déterminer les frontières associées aux différents segments, un système de segmentation automatique (*SAIPH*), utilisant la technique d'analyse *PLP* et des caractéristiques spectrales dynamiques, a été développé.

Afin de segmenter un mot en unités élémentaires, une mesure de transition, pouvant préserver les changements intervenant dans le signal de parole, doit être définie. Une façon de représenter les transitions est de modéliser les caractéristiques dynamiques du spectre de fréquence. Signalons toutefois d'autres méthodes de segmentation automatique fondées sur l'utilisation de classes grossières [16] ou d'événements phonétiques [32]. *SAIPH* utilise des coefficients de régression [17] dérivés des coefficients cepstraux fournis par l'analyse *PLP*. Une mesure de transition (pour la trame k) a été définie par :

$$(2) \qquad \mathrm{TM}(k) = \sum_{i=1}^{M} \frac{(i \times r_i^k)^2}{\mathrm{M}}$$

où M est l'ordre du modèle d'analyse et r_i le i-ième coefficient de régression.

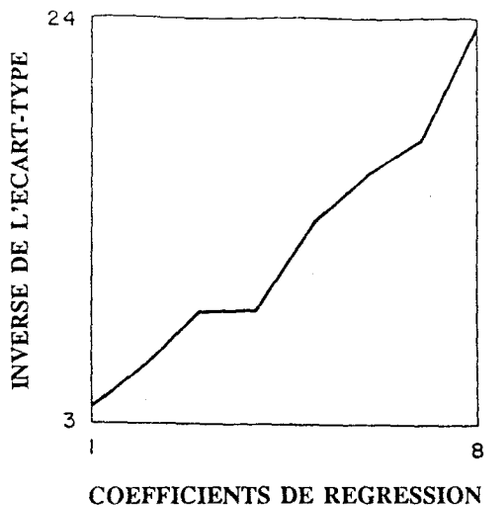

Figure 4. — Inverse de l'écart type des coefficients de régression obtenus à partir de l'analyse PLP pour un vocabulaire alphanumérique.

Dans l'équation (2), chaque coefficient de régression est multiplié par son index avant de l'élever au carré. Comme le montre la figure 4, où l'inverse de l'écart-type des coefficients de régression est représenté, cette multiplication est une bonne approximation de la normalisation des variations de chaque coefficient de régression.

L'équation (2) peut aussi être interprétée comme le calcul d'une mesure de transition sur les coefficients cepstraux pondérés *RPS*. Une caractéristique de ces coefficients est d'être très sensible aux brusques variations de pente survenant autour des pics du spectre de fréquence.

La fenêtre optimale pour calculer les coefficients de régression a été trouvée expérimentalement égale à 70 ms (avec une taille de trame de 10 ms). Enfin, cette mesure de transition a été lissée par une moyenne flottante pour éliminer les pics parasites.

Comme le montre la figure 5, les pics de la courbe de transition du mot « V » (choisi comme exemple) indiquent la frontière entre les parties consonne et voyelle ainsi que le début et la fin du mot.

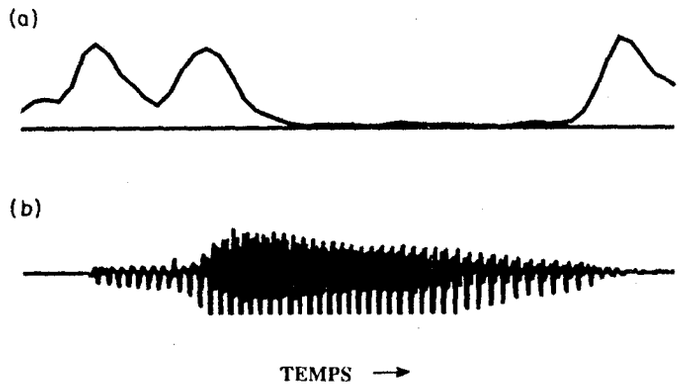

Figure 5. — Courbe de transition (*a*) et signal temporel (*b*) associés au mot « V ».

Cette mesure de transition ne dépend pas du locuteur et peut être calculée en même temps que le modèle d'analyse acoustique utilisé lors de la première passe.

Le modèle PLP d'ordre cinq a aussi été choisi pour aider à la discrimination. Comme le montre la figure 6 pour les

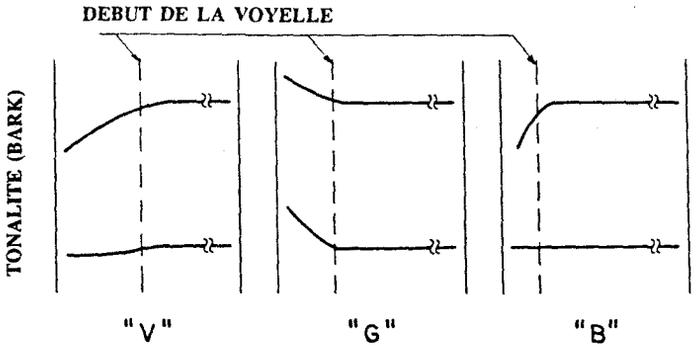

Figure 6. — Mouvements des pics spectraux du modèle PLP d'ordre cinq pour les mots « V », « G », « B ».

mots (« V », « G », « B »), une représentation grossière du spectre de fréquence, à l'aide d'un modèle d'ordre réduit de l'analyse *PLP*, constitue un indice particulièrement intéressant dans le cadre du vocabulaire étudié.

De telles caractéristiques spectrales grossières ont déjà été employées en reconnaissance de parole. Par exemple, dans les travaux de Makhoul [27] un modèle LPC à deux pôles était utilisé.

4.3. REPRÉSENTATION DES CONNAISSANCES ET RAISONNEMENT INCERTAIN

Les connaissances phonétiques ont été encodées à l'aide d'un langage de représentation qui décrit la connaissance en terme de règles et de faits. La représentation utilisée par ce langage est exprimée par une syntaxe à base de « frames » (qui dans ce cas désigne un ensemble d'informations contenues dans une structure commune). Deux sortes de connaissances sont distinguées :

- *connaissances schématiques,*
- *connaissances applicatives,*

où les *connaissances schématiques* décrivent quels sont les attributs qui peuvent être associés à un indice (exemple : faible, moyen, fort), et les *connaissances applicatives* représentent un ensemble de faits et de règles qui manipulent ces indices. Les connaissances sont encodées à l'aide d'une structure à base d'arbre de décision pour chaque règle traitant d'un mot particulier.

Cette étude a été orientée vers l'acquisition et la représentation des connaissances plutôt que vers le développement d'un moteur d'inférence. Pour ce dernier, un produit commercial a été sélectionné. Il s'agit de *KWB* (en abréviation de « knowledge workbench ») qui est en fait un environnement de programmation pour développer des systèmes experts. Le langage de représentation des connaissances utilisé fait aussi partie de cet ensemble logiciel. Le moteur d'inférence (écrit en prolog) permet, en particulier, de communiquer avec des fonctions écrites dans des langages usuels (par exemple C) afin de récupérer les indices extraits automatiquement. De plus, il fournit des outils intéressants de mise au point et de manipulation de *connaissances incertaines.*

Les indices ont été décrits de façon qualitative (variables linguistiques) en terme de faible, moyen, fort, etc. afin de construire une base de connaissances facilement modifiable par des chercheurs qui ne seraient pas familiés avec le système. C'est aussi la façon la plus naturelle de décrire des spectrogrammes.

Le système manipule des connaissances incertaines en utilisant un mécanisme défini dans MYCIN [4] qui associe à chaque indice extrait un nombre compris dans un intervalle donné. Ainsi, chaque indice est caractérisé par une variable linguistique et une valeur d'incertitude comprise entre 0 et 1. Ceci permet de fournir au moteur d'inférence des informations continues et non discrètes. La valeur d'incertitude correspond au degré de confiance que l'on attribue à la présence d'un indice donné dans le signal manipulé.

Enfin, une valeur d'incertitude (entre 0 et 1) est aussi associée à chaque branche de l'arbre de décision représentant une règle. Cette valeur d'incertitude a été ajustée en fonction des résultats de reconnaissance obtenus et de la confiance que l'on attribue à chaque règle. Par exemple, une des branches de l'arbre de décision constituant une règle et définissant le mot « B » est :

if burst_strength='weak' and voice_bar='yes'
and plp_peak_1='flat' and plp_peak 2='increase_fast'
and begin_no_voiced='very_small'
then word='b' cf(0.7).

où *burst_strength* indique l'énergie avec laquelle une barre d'explosion est présente, *voice_bar* indique l'existence (ou l'absence) d'une barre de voisement, *begin_no_voiced* représente la durée de la partie non voisée en début de mot, et *plp_peak_1* et *plp_peak_2* représentent les mouvements des pics spectraux du modèle *PLP* d'ordre cinq dans la transition consonne-voyelle.

Si l'on compare ces règles à celles utilisées par un système de décodage acoustico-phonétique tel qu'*APHODEX* (réalisé au CRIN) [6] nos règles sont beaucoup plus simples. En effet, les règles développées utilisent uniquement des informations extraites sur le signal manipulé alors que les règles d'*APHODEX* tiennent compte, en particulier, des phonèmes candidats pour le segment précédent et le segment suivant (contexte gauche et contexte droit). Dans notre cas, le contexte étant similaire, cette information n'est pas nécessaire. La simplicité de nos règles découle de l'application considérée.

5. Combinaison d'un système à règles de production et d'une approche reconnaissance des formes

Les résultats des deux approches à base de règles de production et de comparaison de formes sont combinés à l'aide d'une *stratégie de décision.* La décision finale est prise en tenant compte des résultats fournis par les deux passes du système, chaque passe fournissant des candidats potentiels pour le mot reconnu. Aux hypothèses générées par l'algorithme de programmation dynamique est associé un facteur de confiance dépendant des mots qui ont été reconnus et d'une matrice de confusion obtenue par apprentissage. Quant à la deuxième passe (fondée sur l'extraction d'indices) elle effectue une discrimination entre trois classes : {B, D, E}, {G, T, P}, {V, Z, C, FEED}. Celà signifie que lorsque les candidats trouvés par cette deuxième passe sont fournis, il y a aussi une décision prise quant à l'appartenance du mot de test à une de ces trois classes. La décision finale considère les candidats du système à base de connaissances (qui appartiennent à la classe retenue) et les candidats, parmi les trois premiers, de l'algorithme de programmation dynamique qui appartiennent à la classe identifiée par le système à base de connaissances. Un facteur de confiance global

est calculé grâce à une fonction combinant les différents facteurs de confiance des hypothèses générées pendant les deux passes :

$$CF\ global = CF1 + CF2 - CF1 \times CF2\ .$$

où $1 = première\ passe$ et $2 = deuxième\ passe$.

Cette fonction, fondée sur l'hypothèse que les résultats générés par les deux passes sont indépendants, permet d'accumuler des évidences progressivement. Le mot reconnu est celui auquel est associé le facteur de confiance le plus élevé. Cette stratégie de décision suit le principe d'*information prépondérante*, utilisé dans le système de reconnaissance de parole continue *HEARSAY II* [14], qui donne plus d'importance à la source de connaissance la plus fiable. Dans le système proposé, pour les classes de mots difficiles les candidats générés par la deuxième passe sont plus fiables que ceux générés par l'algorithme de programmation dynamique.

Ce système hybride a été testé sur le vocabulaire *E-SET*, qui est un sous-ensemble de la base de données *D1*. Les scores de reconnaissance obtenus sont très proches de 90 %, ce qui correspond à une diminution de l'erreur de reconnaissance de plus de 60 % par rapport aux résultats obtenus lors de la première passe. Dans la version actuelle du système, les facteurs de confiance associées aux indices extraits sont calculés mais ne sont pas encore pris en compte par le moteur d'inférence. De plus, les facteurs de confiance qui ont été utilisés (ceux qui sont associés aux règles) ont besoin d'être ajustés grâce à des tests plus intensifs. Ces dernières remarques laissent entrevoir un facteur supplémentaire de progression qui peut aussi être obtenu par le raffinement des connaissances utilisées.

6. Résumé et conclusions

ORION est un système hybride à deux passes qui utilise plusieurs sources de connaissances : *psychoacoustiques* (dans l'analyse *PLP*), *physiologiques* (par l'introduction de caractéristiques spectrales dynamiques) et *phonétiques*. Un tel système tient compte de notre connaissance sur la parole sans pour autant négliger ses limitations. Pour un vocabulaire de mots qui ne sont pas acoustiquement similaires, un modèle qui simule des propriétés du système auditif humain et accentue les transitions spectrales a fourni les meilleurs résultats parmi les modèles étudiés. Une évaluation de ce système sur une base de données de chiffres (*D3*) a permis d'obtenir des scores de reconnaissance supérieurs à 98 % (moyenne pour tous les locuteurs de test). Pour des mots acoustiquement similaires, l'introduction de connaissances phonétiques, l'utilisation de modèles perceptuels (pour la segmentation automatique et l'extraction d'indices acoustiques) et le développement d'une stratégie de décision élaborée diminuent considérablement l'erreur de reconnaissance (environ 60 %) et permettent de s'affranchir des limitations associées à l'algorithme de programmation dynamique.

Manuscrit reçu le 19 septembre 1989.

BIBLIOGRAPHIE

[1] P. Alinat, « Reconnaissance des Phonèmes au Moyen d'une Cochlée Artificielle », 1973. Thèse de Docteur Ingénieur.

[2] B. S. Atal, « Effectiveness of Linear Prediction Characteristics of the Speech Wave for Automatic Speaker Identification and Verification », *J. Acoust. Soc. Amer.*, 55 : 1304-1312, June 1974.

[3] G. L. Bradshaw, R. Cole and Z. Li, « A Comparison of Learning Techniques in Speech Recognition », in *ICASSP-82*, pp. 554-557, 1982.

[4] B. G. Buchanan and E. H. Shortliffe, *Rule-Based Expert Systems : the MYCIN Experiments of the Stanford Heuristic Programming Project*. Addison-Wesley, 1985.

[5] J. Caelen, « Un modèle d'Oreille ; Analyse de la Parole Continue ; Reconnaissance Phonémique », Université Paul Sabatier de Toulouse, 1979, Thèse d'État.

[6] Carbonell *et al.*, « APHODEX, Design and Implementation of an Acoustic-Phonetic decoding Expert System », in *ICASSP-86*, pp. 1201-1204, 1986.

[7] F. Casacuberta and E. Vidal, « Speech Recognition with Difficult Vocabularies », in H. Niemann et al., editor, « *Recent Advances in Speech Understanding and Dialog Systems* », pp. 279-283. Springer-Verlag Berlin Heidelberg, 1988.

[8] R. A. Cole, R. M. Stern and M. J. Lasry, « Performing Fine Phonetic Distinctions », in J. S. Perkell and D. H. Klatt, editors, *Variability and Invariance in Speech Processes*, pp. 325-345. Hillsdale, NJ, Lawrence Erlbaum Assoc, 1985.

[9] Cole *et al.*, « Feature-Based Speaker-Independent Recognition of Isolated English Letters », in *ICASSP-84*, pp. 731-733, 1983.

[10] Dallos *et al.*, « Cochlear Inner and Outer Hair Cells : Functional Differences », *Science*, 177 : 356-358, 1972.

[11] B. Delgutte, « Some Correlates of Phonetic Distinctions at the Level of the Auditory Nerve », in R. Carlson and B. Granström, editors, *The representation of Speech in the Peripheral Auditory System*, pp. 131-149. Elsevier Biomedical Press, 1982.

[12] B. Delgutte, « Codage de la Parole dans le Nerf Auditif », 1984. Thèse de Doctorat d'État.

[13] M. Dolmazon, « Représentation of Speech-like Sounds in the Peripheral Auditory System in Light of a Model », in R. Carlson and B. Granström, editors, *The Representation of Speech in the Peripheral Auditory System*, pp. 151-164. Elsevier Biomedical Press, Amsterdam, 1982.

[14] L. D. Erman, F. Hayes-Roth, V. R. Lesser and D. R. Reddy, « The HEARSAY-II Speech-Understanding System : Integrating Knowledge to Resolve Uncertainty », *Computing Surveys*, 12(2) : 213-253, 1980.

[15] H. Fletcher, « Auditory Patterns », *Review of Modern Physics*, pp. 47-65, 1940.

[16] Fohr *et al.*, « Paramétrisation Acoustique et Décodage Phonétique Fondé sur des Connaissances, pour la Parole Continue Multilocuteur », in *Décodage Acoustico-Phonétique, éditeur GRECO Communication Parlée du C.N.R.S.* 1988.

[17] S. Furui, « On the Role of Spectral Transition for Speech Perception », *J. Acoust. Soc. Am.* (80) : 1016-1025, 1986.

[18] B. A. Hanson and H. Wakita, « Spectral Slope Based Distortion Measures for All-Pole Models of Speech », in *ICASSP-86*, pp. 757-780, 1986.

[19] H. Hermansky, « An Efficient Speaker-Independent Automatic Speech Recognition by Simulation of Some Properties of Human Auditory Perception », in *ICASSP-87*, pp. 1159-1162, 1987.

[20] H. Hermansky, B. A. Hanson and H. Wakita, « Low-Dimensional Representation of Vowels Based on All-Pole Modeling in the Psychophysical Domain », *Speech Communication* (4) : 181-187, 1985.

[21] H. Hermansky and J. C. Junqua, « Optimization of Perceptually-Based ASR Front-End », in *ICASSP-88*, pp. 219-222, 1988.

[22] J. C. Junqua, « Evaluation of ASR Front-Ends in Speaker-Dependent and Speaker-Independent Recognition », *J. Acoust. Soc. Am.* (81 S1) : S93, 1987.

[23] J. C. Junqua, « Contribution à l'Amélioration de la Robustesse des Systèmes de Reconnaissance Automatique de Mots Isolés », Université de Nancy I, May 1989. Thèse d'Université.

[24] D. H. Klatt, « Prediction of Perceived Phonetic from Critical-Band Spectra : a First Step », in *ICASSP-82*, pp. 1278-1281, 1982.

[25] L. F. Lamel, and V. W. Zue, « Performance Improvement in a Dynamic-Programming-Based Isolated Word Recognition System for The Alpha-Digit Task », in *ICASSP-82*, pp. 558-561, 1982.

[26] R. F. Lyon, « A Computational Model of Filtering, Detection, and Compression in the Cochlea », in *ICASSP-82*, pp. 1282-1285, 1982.

[27] J. Makhoul, « Spectral Analysis of Speech by Linear Prediction », *IEEE Trans. ASSP-21* (3) : 140-148, 1973.

[28] J. Makhoul and R. Schwartz, « Ignorance Modeling : comments from Performing Fine Phonetic Distinctions, R. Cole, R. M. Stern and M. J. Lasry », in J. S. Perkell and D. H. Klatt, editors, *Variability and Invariance in Speech Processes.* Hillsdale, NJ, Lawrence Erlbaum Assoc, 1985.

[29] E. A. Martin, R. P. Lippmann and D. Paul, « Two-Stage Discriminant Analysis for Improved Isolated-Word Recognition », in *ICASSP-87*, pp. 709-712, 1987.

[30] C. Myers, L. R. Rabiner and A. E. Rosenberg, « Performance Tradeoffs in Dynamic Time Warping Algorithms for Isolated Word Recognition », *IEEE Trans. ASSP*, 28(6) : 623-634, 1980.

[31] K. K. Paliwal, « On the Performance of the Quefrency-Weighted Cepstral Coefficients in Vowel Recognition », *Speech Communication* (1) : 151-154, 1982.

[32] G. Perennou and M. de Calmes, « Segmentation en Événements Phonétiques et en Unités Syllabiques », in *XIV JEP Paris*, pp. 142-146, 1985.

[33] L. R. Rabiner and J. G. Wilpon, « Isolated Word Recognition Using a Two-Pass Pattern Recognition Approach », in *ICASSP-81*, pp. 724-727, 1981.

[34] D. W. Robinson and R. S. Dadson, « A Redetermination of the Equal-Loudness relations for Pure Tones », *British Journal of Applied Physics*, 7 : 166-181, 1956.

[35] S. Seneff, « A Computational Model for the Peripheral Auditory System : Application to Speech Recognition Research », in *ICASSP-86*, pp. 1983-1986, 1986.

[36] S. S. Stevens, « On the Psychophysical Law », *Phychological Review*, 64 : 153-181, 1957.

[37] V. W. Zue and L. F. Lamel, « An Expert Spectrogram Reader : A Knowledge-Based Approach To Speech Recognition », in *ICASSP-86*, pp. 1197-1200, 1986.