Décodage acoustico-phonétique : apport de la décomposition temporelle généralisée et de transformations spectrales non-linéaires

Application à la reconnaissance des mots épelés en continu

Thèse présentée en vue de l'obtention du

DOCTORAT DE L'E.N.S.T.

par

Claude MONTACIE

Soutenue le **13 février 1991** devant le jury composé de :

Président :	J.-J. Mariani
Rapporteurs :	L. Miclet
	C.J. Wellekens
Directeur :	X. Rodet
Correspondant :	G. Chollet
Examinateur :	H. Méloni

RÉSUMÉ

Le travail présenté dans cette thèse se situe dans le cadre plus général de la Reconnaissance Automatique de la Parole Continue. Il présente un protocole complet de décodage acoustico-phonétique, composé des modules permettant le passage du signal vocal à un treillis phonétique. Le système proposé permet un décodage robuste. Il a été évalué sur un corpus de mots épelés continûment. L'adaptation à un locuteur de test différent de celui de l'apprentissage est également étudié.

Ce système présente plusieurs aspects originaux :

— L'utilisation d'un modèle original de l'évolution des paramètres spectraux : **la décomposition temporelle généralisée**, c'est-à-dire étendue à une mesure de dissimilarité spectrale quelconque. La décomposition temporelle rend compte des aspects dynamiques et redondants du signal de parole pour localiser et calculer des cibles spectrales moins dépendantes d'événements locaux non pertinents que les paramètres originaux.

— La paramétrisation du signal vocal avec le développement et l'implémentation d'un algorithme de **lissage cepstral itératif**. On montre que certains facteurs jugés parasites (e.g., la mélodie) voient ainsi leur influence diminuer.

Ce protocole de décodage acoustico-phonétique a été appliqué à un corpus de mots épelés en mode continu à vitesse d'élocution rapide. Il a permis d'obtenir un taux de décodage acoustico-phonétique de 70% avec un taux d'insertion de 20% (au premier choix). Sans juger de la complexité intrinsèque de la base de données d'application, la comparaison avec les techniques les plus performantes, telles que les **machines de Markov cachées**, n'a pas donné des performances équivalentes sur ce même corpus.

L'extension de ce système, dans un cadre pluri-locuteur, a entraîné l'emploi de **techniques d'adaptation au locuteur**. Celles-ci présentent les caractéristiques suivantes :

— Le calcul de fonctions de **transformations spectrales non-linéaires** utilisant les **perceptrons multi-couches**. Cette technique a été testée par rapport à la Régression Linéaire Multiple et au filtrage quadratique.

— Le **choix automatique du vocabulaire d'adaptation** spécifique de chaque locuteur.

Ces techniques ont permis une augmentation du taux de décodage acoustico-phonétique de plus de 10% comparé à ce que l'on peut obtenir en l'absence d'adaptation, pour un locuteur de test différent de celui de référence.

SOMMAIRE

Résumé ……I

Sommaire ……III

Notations utilisées……V

Introduction …… 1

Chapitre 1. Les modèles de l'évolution spectrale ……11

Chapitre 2. La décomposition temporelle généralisée ……23

Chapitre 3. Paramétrisation spectrale adaptée au décodage acoustico-phonétique et à la décomposition temporelle généralisée.....49

Chapitre 4. Le processus de décodage acoustico-phonétique.....63

Chapitre 5. Adaptation au locuteur par transformations non-linéaires des cibles spectrales……75

Chapitre 6. Evaluation du système de décodage acoustico-phonétique..... 103

Conclusions et Perspectives..... 111

Références Bibliographiques..... 113

Annexe I. Distance d'Itakura-Saito..... 129

Annexe II. Résultats comparatifs du système de référence…… 131

Annexe III. Syntaxe et modèles du système de référence …… 135

Annexe IV. Calcul des vecteurs et des valeurs propres d'une matrice …… 141

Annexe V. Vocabulaire d'apprentissage..... 145

Annexe VI. Vocabulaire de test..... 149

Annexe VII. Algorithme de Wagner et Fisher 151

Annexe VIII. Article ICASSP 153

Annexe IX. Exemples de décomposition temporelle 157

Annexe X. Alphabet phonétique..... 161

Table des Figures..... 163

Table des Tableaux..... 165

Table des Matières..... 167

NOTATIONS UTILISÉES

$\mathcal{D}$ est une distance ou une mesure de dissimilarité.

$\delta(\phi, n)$ est la fonction de compacité de la fonction d'interpolation au temps t_n.

d_n est le $n^{ième}$ vecteur désiré par le perceptron multi-couches.

$\mathcal{E}$ est un espace métrique quelconque.

ϕ est une fonction d'interpolation.

g représente les cibles spectrales.

ℓ_2 est la distance euclidienne.

μ est la valeur maximale de la fonction de compacité δ.

NV représente la dimension de l'espace acoustique (nombre de points de quantification).

N représente le nombre de vecteurs spectraux composant la trajectoire spectrale.

p est la taille des vecteurs spectraux.

q est l'ordre du modèle de l'évolution des paramètres spectraux.

T est la largeur d'une fenêtre d'analyse.

t_n est l'instant correspondant au $n^{ième}$ paramètre de la trajectoire spectrale.

w_{ij} est le poids du perceptron multi-couches entre le noeud i et le noeud j.

ω est une fonction caractéristique.

x_n est le $n^{ième}$ vecteur présenté au perceptron multi-couches.

y^n est la $n^{ième}$ composante du vecteur spectral y.

y_n est le $n^{ième}$ paramètre de la trajectoire spectrale.

$\mathbf{y}_n$ est le $n^{ième}$ vecteur spectral de la trajectoire spectrale.

INTRODUCTION

La Reconnaissance de la Parole Continue consiste à transformer le "flot" acoustique du signal de parole en une représentation symbolique. Cette représentation doit être caractéristique du contenu linguistique. Les recherches en Reconnaissance Automatique de la Parole, domaine issu du cadre plus général de la Reconnaissance des Formes, ont suivi tout d'abord deux axes distincts : l'un procédant suivant une approche globale, l'autre suivant une approche analytique. Actuellement la distinction analytique/global a perdu une partie de sa netteté depuis que les méthodes globales, limitées au début aux mots isolés ont été généralisées aux mots connectés.

L'approche globale consiste à identifier directement les mots ou groupes de mots en les comparant à des représentations acoustiques de référence. Actuellement les techniques les plus performantes sont fondées sur la comparaison dynamique et sa généralisation, la modélisation par les machines de Markov cachées. L'approche globale s'applique principalement à la reconnaissance de mots isolés ou connectés.

Les effets de la coarticulation, que de nombreuses études constatent et répertorient [Petitot-Cocorda 84], apparaissent lors de la réalisation des phonèmes, ils rendent la parole naturelle plus complexe qu'une simple suite de mots isolés. Une approche analytique doit alors être envisagée pour la reconnaissance de la parole que l'on convient d'appeler "continue". Il s'agit de localiser et d'identifier sur un signal vocal, en mode d'élocution continue (i.e., sans pauses particulières entre les mots), des entités a priori caractéristiques du contenu linguistique. La définition de l'ensemble de ces entités doit permettre une reconstitution de l'énoncé prononcé, d'un point de vue phonétique, il s'agit alors du Décodage Acoustico-Phonétique (DAP). Les unités généralement utilisées sont les phonèmes (au nombre d'une trentaine en français), les diphones (au nombre de 1200), les syllabes (au nombre de 5000), les triphones (au nombre de 7000). Cette reconnaissance d'entités discriminantes utilise de nombreuses techniques issues des domaines du Traitement du Signal, de la Phonétique, de l'Intelligence Artificielle ou des Probabilités. Parmi ces techniques d'ailleurs utilisées dans l'approche globale, on peut citer la recherche de traits acoustiques (e.g., voisé/non-voisé, ouvert/fermé ...), les systèmes à base de règles faisant coopérer des connaissances de nature et de niveau différents, les machines de Markov cachées, les réseaux connexionistes.

A. LA RECONNAISSANCE DE LA PAROLE CONTINUE

La reconnaissance de la parole continue peut être basée directement sur une comparaison de formes nouvelles avec des références des mots à reconnaître (e.g., description complète des mots en termes de modèles acoustiques), ou bien sur l'identification d'un ensemble d'unités élémentaires (e.g., phonèmes, diphones, syllabes). Dans le premier cas, il s'agit d'une reconnaissance globale, dans le second cas d'une reconnaissance analytique.

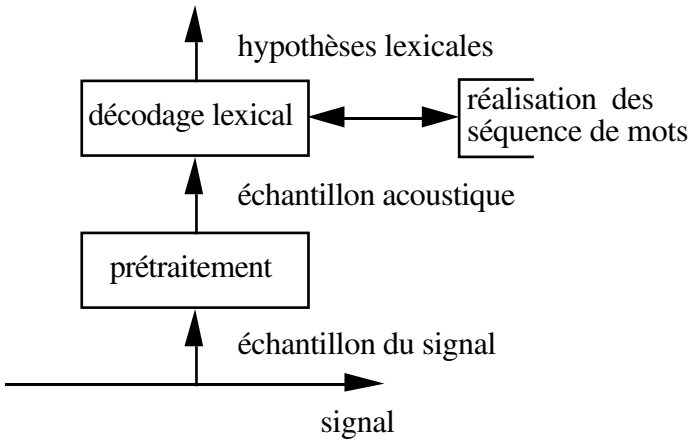


Figure A.1. Méthode de reconnaissance globale [Potage 86]

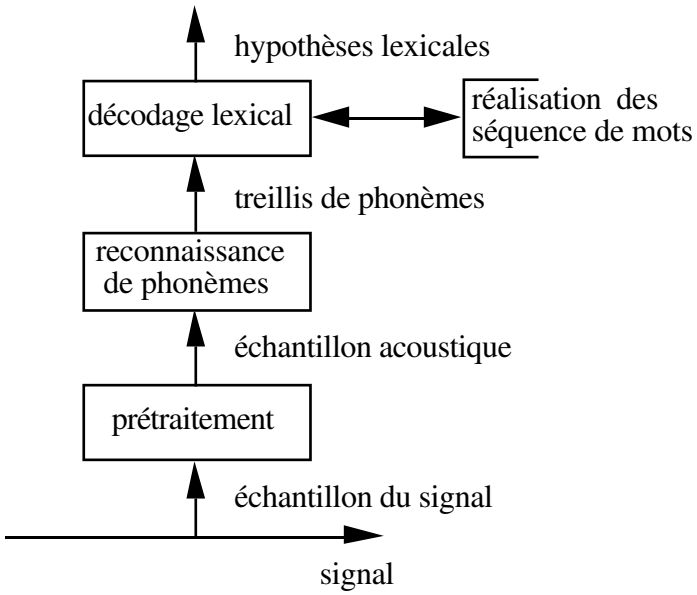


Figure A.2. Méthode avec reconnaissance de phonèmes [Potage 86]

L'approche globale a donné lieu ces trente dernières années à de nombreux systèmes opérationnels. Ils ont cependant quelques inconvénients majeurs :

— La mémoire et le temps de calcul demandés sont importants.

— La taille du vocabulaire reconnu est généralement faible. Elle dépasse rarement quelques centaines de mots. Ce vocabulaire est de plus difficilement modifiable.

— L'adaptation à des locuteurs est extrêmement lente et difficile.

Ces systèmes bien qu'imparfaits sont souvent cités dans la littérature car ce sont les premiers à avoir été opérationnels et pour certains commercialisés. Les systèmes de décodage acoustico-phonétique sont plus prometteurs, mais ils sont moins performants à l'heure actuelle.

Nous allons d'abord faire un bref état de l'art des Systèmes de Compréhension de la Parole Continue. Nous décrirons ensuite les bases de données utilisées pour l'évaluation de ces systèmes. Cette description de l'état de l'art a été faite à partir de [Mariani 82][Méloni 82][Potage 86][Mariani 89][Calliope 89][Waibel et Lee 89].

A.1. ETAT DE L'ART EN RECONNAISSANCE DE LA PAROLE CONTINUE

— Les premières réalisations

Si l'on excepte les systèmes des années cinquante à soixante, comme la "machine à écrire phonétique" d'Olson et Belar [Olson et Belar 56] ou la commande vocale d'un robot de Vicens [Vicens 69], qui avaient des performances très limitées (i.e., un vocabulaire de quelques mots), les premiers systèmes de reconnaissance de la parole continue ont été étudiés aux Etats-Unis dans le cadre du projet ARPA-SUR [Klatt 77]. Il s'agit des systèmes HARPY [Lowerre 76], BBN[Woods et al. 76], Hearsay II [Reddy 77], SDC. Le premier système peut être considéré comme global, les trois autres comme analytiques. Les performances du décodage acoustico-phonétique pour ces systèmes sont les suivantes:

— BBN pour 71 catégories phonétiques reconnait 52% des phonèmes

— Hearsay II pour 98 catégories phonétiques reconnait 42% des phonèmes

— SDC pour 40 catégories phonétiques reconnait 50% des phonèmes.

On peut aussi citer le système DRAGON [Baker 75], le premier à utiliser une modélisation à base de machines de Markov.

En Union Soviétique le système de l'Institut Cybernétique de KIEV annoncait de très bons résultats [Vintsjuk 76] au prix d'un long apprentissage. Il utilisait la technique de comparaison dynamique développée dans cet institut en 1968.

La recherche française était caractérisée par l'idée que les imperfections du décodage acoustico-phonétique n'empêchaient pas d'accéder au niveau supérieur de la compréhension. Au contraire les faiblesses des niveaux "inférieurs" pouvaient être épaulées par le recours à des niveaux plus "élevés" en utilisant les contraintes relatives au langage. Nous pouvons citer les systèmes suivants :

— Keal [Mercier 69][Mercier et al. 88]

La parole est d'abord segmentée en unités syllabiques en examinant l'énergie du signal. Une seconde segmentation localise les consonnes et les voyelles. Une fonction de discrimination linéaire est utilisée pour les voyelles. Un arbre d'identification est utilisé pour les consonnes et les semi-voyelles.

— Myrtille II [Haton et Pierrel 78]

L'analyse acoustique extrait différents paramètres comme les énergies absolues et quadratiques moyennes, le nombre de passages par zéro, le nombre moyen d'extrema par seconde ainsi que certains marqueurs prosodiques. Des procédures hiérarchisées extraient de ces différents indices un treillis de phonèmes.

— Esope [Mariani 82].

L'analyse acoustique extrait l'énergie dans 16 bandes de fréquence dont la largeur suit une échelle Mel [Zwicker 77]. Les instants de stabilité et de transition sont déterminés par la détection des minima et maxima de la courbe de variation de l'énergie dans chaque bande. L'étiquetage de chaque segment ainsi détecté est fait par comparaison avec des spectres de référence. On passe du treillis de phonèmes obtenu aux mots par une procédure de "vérification lexicale". La stratégie de cette procédure est descendante, gauche-droite, avec poursuite des meilleurs chemins et sans retour en arrière. Pour 34 catégories phonétiques, le pourcentage de phonèmes bien reconnus moins le pourcentage de phonèmes insérés est de 35% pour ESOPE.

— Les systèmes actuels

Ils sont pour la plupart à base de machines de Markov cachées. Ils sont capables de faire une reconnaissance suivant une approche analytique (une machine de Markov par

phonème) ou suivant une approche globale (une machine de Markov par mot ou groupe de mots). On peut citer les suivants :

BYBLOS [Kubala et al. 88] est un système de reconnaissance monolocuteur développé à BBN. Il est fondé sur les machines de Markov cachées. Sur un ensemble de phrases comprenant un vocabulaire de 997 mots, le taux de reconnaissance est de 67.6% de mots correctement reconnus pour une syntaxe non contrainte.

A Carnegie Mellon SPHINX [Lee 88] est un système de reconnaissance indépendant du locuteur à base de machines de Markov cachées. Il accepte un grand vocabulaire (i.e., mille mots). La particularité de ce système est l'utilisation de plusieurs types de paramètres (cepstraux, cepstraux différentiels, énergie, énergie différentielle) combinés dans plusieurs "codebooks" (i.e., ensemble de vecteurs représentatifs) au lieu d'un seul généralement. Les locuteurs sont eux aussi classés en plusieurs ensembles par une autre machine de Markov cachée. Sur le même corpus que le système précédent, mais en mode indépendant du locuteur, le taux de reconnaissance est de 70.6% de mots correctement reconnus pour une syntaxe non contrainte. Sur la base de donnée TIMIT, les résultats de ce système sont un taux de DAP, en mode indépendant du locuteur, de 65% pour 30 catégories phonétiques [Lee et Hon 89].

Le système TANGORA [Averbuch et al. 87] développé à IBM est conçu pour les vocabulaires de très grande taille (i.e., 20 000 mots). Il fonctionne en temps réel suivant une approche globale utilisant des machines de Markov cachées. Il annonce un taux de reconnaissance par mots de 93.8% en mode mono-locuteur.

En Europe, le système SPICOS [Ney et al. 88] utilise des machines de Markov continues. Il annonce un taux de reconnaissance monolocuteur de 92% sur un ensemble de phrases composées à partir de 917 mots en allemand. La syntaxe est cependant contrainte. L'originalité de ce système est l'utilisation de fonctions de probabilités de Laplace au lieu des fonctions de probabilités gaussiennes employées habituellement.

En France, le système Phil86 [Jouvet 88] développé au CNET est aussi à base de machines de Markov cachées continues. Il fonctionne en mode indépendant du locuteur sur un vocabulaire de 50 à 100 mots. Ce système a été implanté sur un processeur de signal [Monné 88]. Les performances annoncées sont de 95% de mots reconnus, mais pour une syntaxe contrainte. Nous avons utilisé l'implantation temps réel du système Phil86 pour évaluer notre système de décodage acoustico-phonétique (cf. Tableau 6.2). Les résultats sont un taux de DAP de 60% pour 22 catégories phonétiques, en mode monolocuteur.

Quelques systèmes utilisant des processeurs de comparaison dynamique spécialisés donnent cependant de très bonnes performances en temps réel. On peut citer en France le système MOZART et Amadeus développé au Limsi [Gauvain et Mariani 82].

D'autres équipes développent des systèmes de décodage acoustico-phonétique suivant des approches différentes. Certains utilisent les techniques d'Intelligence Artificielle comme les systèmes à base de règles :

— L'une des premières utilisations des systèmes experts pour le décodage acoustico-phonétique est due à R. De Mori [De Mori et al. 83]. Il s'agissait d'un système de règles pour l'extraction de traits acoustiques.

Une autre application des systèmes experts est la lecture automatique des spectrogrammes de parole : la connaissance d'un phonéticien est alors mises sous forme de règles. On peut citer le système du Limsi [Stern et al.86] (un taux de DAP de 63% pour 13 catégories phonétiques), celui du MIT [Zue et Lamel 86] (un taux de DAP de 84% pour 6 catégories phonétiques), ainsi que le système Aphodex [Carbonell et al. 84][François et Fohr 90].

Les systèmes experts permettent aussi de faire coopérer des stratégies différentes de DAP. L'approche complémentaire de plusieurs équipes devrait apporter un enrichissement des systèmes pris isolément. Des études préliminaire ont été décrites dans [Méloni et Bulot 87][Bulot et Méloni 88][Caelen et al. 88].

D'autres systèmes utilisent des techniques connexionnistes comme les TDNNs (Time Delay Neural Networks) [Waibel et al. 89a][Hataoka et Waibel 89]. Cette dernière approche prometteuse n'a pas encore donné de meilleurs résultats que les machines de Markov cachées. La tendance actuelle est d'associer les deux méthodes [Bourlard et Wellekens 88][Franzini et al. 90].

A.3. Les bases de données

L'évaluation des algorithmes de décodage acoustico-phonétique et plus généralement des systèmes de reconnaissance doit être faite sur des bases de données de test communes et facilement accessibles. Elle doivent être préalablement étiquetées et segmentées. De nombreuses bases de données de parole ont été enregistrées depuis plusieurs années : notamment pour la France BDSON et pour les Etats-Unis DARPA TIMIT. Nous avons aussi utilisé une base de mots isolés TEXAS que nous décrirons au chapitre 5.

— BDSON

Les corpus BDSON ont été définis par le GRECO "Communication Parlée" du CNRS [Cervantes et Sérignat 86]. Ils sont divisés en une partie Acoustique et une partie Evaluation. Quelques corpus peuvent être utilisés pour le décodage acoustico-phonétique. Ils s'agit des corpus PEQ (i.e., 50 phrases phonétiquement équilibrées), SYL (i.e., 192 phrases dédiées à l'étude des diphones [CV]), LABISA (texte d'environ 40 secondes) et REC (i.e., 50 noms épelés en continu) . C'est ce dernier corpus que nous avons utilisé pour évaluer notre système de décodage acoustico-phonétique. Cette base de données est en langue française et comporte 32 locuteurs. Elle est disponible depuis 1987 sur bandes vidéo BETAMAX et depuis 1991 sur 7 CD-ROM. A ce jour, elle n'est pas encore entièrement étiquetée.

— DARPA TIMIT

La base de données de parole DARPA TIMIT [Fisher et al. 86][Fisher et al. 87] a été spécifiquement développée pour l'évaluation du décodage acoustico-phonétique. Elle est en langue anglaise et comporte 630 locuteurs. Chacun des locuteurs a prononcé dix phrases. Deux phrases sont les mêmes pour tous les locuteurs. Les huit autres phrases sont tirées au hasard, cinq parmi un premier ensemble restreint de phrases, trois à partir d'un second ensemble plus grand. Cette base de données complètement étiquetée est disponible depuis 1989 sur CD-ROM pour les 420 premiers locuteurs.

B. CONCLUSIONS ET PLAN

De l'analyse des différents systèmes réalisés, nous retiendrons quelques idées générales qui semblent orienter la conception de futurs systèmes de reconnaissance et qui expliquent notre motivation de recherche.

— Quelle que soit la stratégie adoptée, et l'apport des divers modules, l'information de base provient du décodage acoustico-phonétique : sa qualité, conditionnant le décodage lexical, influe de façon très significative sur les performances du système. Seules des sévères contraintes lexicales et syntaxiques permettent d'obtenir actuellement des taux de reconnaissance acceptables. L'impression qui se dégage de l'ensemble des réalisations, au vu des résultats du décodage acoustico-phonétique, est qu'il semble indispensable d'obtenir à ce niveau de meilleurs scores (i.e., supérieur à 85%, comme le montrent les

expériences de J.-S. Liénard, J.-J. Mariani et G. Renard [Liénard et al. 77] sur l'intelligibilité des phrases synthétiques altérées).

— L'immense variété des situations, qu'il s'agisse du type de locuteur, des conditions d'élocution ou des conditions d'enregistrement, rend impossible la conception d'un système universel. Une procédure d'adaptation, performante et rapide, aux situations rencontrées nous semble indispensable.

Ainsi, la nécessité d'un décodage acoustico-phonétique de meilleure qualité et d'une procédure d'adaptation plus performante a motivé les travaux de recherche exposés dans cette thèse.

B.1. Plan

Le premier chapitre de cette thèse est consacré à un état de l'art des modélisations de l'évolution spectrale. L'intérêt de ces méthodes résident en ce qu'elles permettent d'avoir une description structurée des trajectoires des paramètres spectraux. La décomposition temporelle (classique) est une de ces méthodes.

Le deuxième chapitre décrit un modèle original de l'évolution spectrale, la décomposition temporelle généralisée. Il s'agit de l'extension à une mesure quelconque de dissimilarité spectrale de la technique classique de décomposition temporelle. En effet, certaines paramétrisations (formants, coefficients de prédiction) ne permettent pas une bonne discrimination lorsqu'elles sont associées à une distance euclidienne.

Dans le troisième chapitre, le problème classique du choix de la paramétrisation est abordé. Après un état de l'art des paramétrisations les plus performantes pour le décodage acoustico-phonétique, nous décrivons une nouvelle paramétrisation : les coefficients cepstraux avec lissage cepstral itératif. Cette paramétrisation est moins sensible aux variations de la mélodie que les coefficients cepstraux classiques.

Dans le quatrième chapitre, notre système complet de décodage acoustico-phonétique est décrit avec ses principaux modules. Celui-ci utilise les cibles calculées par l'algorithme de décomposition temporelle, ce qui en fait une de ses particularités. La constitution du dictionnaire des cibles de référence est détaillée. La sélection des cibles identifiées grâce à des paramètres complémentaires (e.g., indice d'énergie, d'instabilité spectrale, de durée) est expliquée.

Dans le cinquième chapitre, les problèmes d'adaptation au locuteur sont abordés, un protocole complet d'adaptation est proposé. Ce protocole doit permettre la

reconnaissance de locuteurs différents de celui de l'apprentissage. Après un état de l'art des fonctions de transformations spectrales linéaires généralement utilisées, nous introduisons le concept de fonctions de transformations spectrales non-linéaires avec l'utilisation du perceptron multi-couches et du filtrage quadratique. Nous définissons ensuite une procédure automatique de choix du vocabulaire d'adaptation (i.e., le choix du vocabulaire prononcé par le nouveau locuteur)

Enfin dans le sixième chapitre, deux modes d'évaluation de notre système de décodage acoustico-phonétique sont présentés. Le premier correspond aux performances du système sur un corpus de noms épelés en continu. Le second vise à la comparaison de nos résultats avec ceux d'un système de référence choisi, fondé sur des machines de Markov connectées.

CHAPITRE 1. LES MODÈLES DE L'ÉVOLUTION SPECTRALE

La plupart des méthodes de représentation du signal vocal reposent sur une propriété fondamentale de ce signal : la stationnarité. Le signal vocal est analysé sur des segments de courte durée à l'intérieur desquels il peut être considéré comme stationnaire. L'évolution du signal est alors décrite par une suite de vecteurs de paramètres d'analyse. Cependant cette représentation discrétisée du signal ne rend pas compte de l'évolution continue du spectre du signal de parole. Pour cette raison, cette représentation présente vraisemblablement des insuffisances pour effectuer un décodage acoustico-phonétique de qualité.

Une description plus structurée des trajectoires de ces vecteurs de paramètres doit apporter une amélioration des performances du décodage acoustico-phonétique. Elle doit aussi permettre une simplification de l'algorithme de décodage employé. Une étude des modèles de l'évolution spectrale va permettre de justifier le choix du modèle retenu pour notre système de décodage acoustico-phonétique.

1.1. LA PRÉDICTION LINÉAIRE VECTORIELLE

La prédiction linéaire (cf. chapitre 3) est une modélisation du signal couramment utilisée dans le domaine du traitement de la parole. Une extension de cette technique consiste à calculer un modèle autorégressif vectoriel pour décrire les trajectoires suivies par des vecteurs d'analyse de la parole.

Soit $\{y_n\}$, (n = 1 ,.., N) une suite de N vecteurs spectraux d'ordre p à moyenne nulle. Son évolution est représentée par un modèle autorégressif vectoriel d'ordre q :

$$y_n = \sum_{i=1}^{q} \Phi_i \, y_{n-i} + \varepsilon_n, \qquad (1.01)$$

où $\{\Phi_i\}$, (i = 1 ,.., q) sont des matrices p x p et $\{\varepsilon_n\}$, (n = 1 ,.., N) est une fonction d'excitation représentée par un bruit blanc vectoriel centré. L'estimation des matrices $\{\Phi_i\}$ est faite par l'algorithme de Levinson généralisé au cas vectoriel [Whittle 63].

Cette approche est aussi utilisée pour le problème de la reconnaissance d'un locuteur parmi plusieurs locuteurs [Grenier 80][1][Artieres et al. 91].

L'utilisation de ce modèle de l'évolution spectrale en reconnaissance de la parole pose cependant trois problèmes :

— L'ordre optimal q_{opt} du modèle de la prédiction linéaire vectorielle est difficile à estimer. Un choix possible est d'utiliser le critère d'Akaike [Akaike 71]. [2]

— Le choix de la durée de la trajectoire spectrale à modéliser est difficile. Une durée trop faible (i.e., quelques secondes) ne permet pas d'estimer précisément les matrices $\{\Phi_i\}$. Une durée trop longue (i.e., plusieurs phrases) lisserait trop la trajectoire des paramètres d'analyse.

---

[1] Les vecteurs $\{y_n\}$ sont des coefficients cepstraux calculés chaque centi-seconde. Un modèle $\{\Phi_i^j\}$, (i = 1 ,.., p) (j = 1 ,.., L) est calculé pour chacun des L locuteurs sur un ensemble de 14 phrases. La phase de reconnaissance du locuteur consiste en un filtrage inverse de la phrase test, représentée par la suite des vecteurs $\{Y_n^{test}\}$ (n = 1 ,.., N_{test}), par les modèles des différents locuteurs. On obtient alors les résidus suivants ε_n^j pour le modèle $\{\Phi_i^j\}$:

$$\varepsilon_n^j = Y_n^{test} - \sum_{i=1}^{p} \Phi_i^j \, Y_{n-i}^{test} \qquad (1.02)$$

Le locuteur reconnu est celui correspondant au modèle $\{\Phi_i^{jopt}\}$ tel que :

$$j_{opt} = \underset{j}{\text{Argmin}} \; \text{trace} \left(\sum_{n=1}^{N_{test}} \varepsilon_n^j \, (\varepsilon_n^j)^T \right) \qquad (1.03)$$

(i.e., celui dont le résiduel vectoriel est le plus faible)

Sur 70 phrases de test, pour 10 locuteurs, le taux de réussite a été de 100%. Y. Grenier considère que "le modèle calculé sur la voix d'un locuteur modélise les capacités articulatoires du locuteur, du moins en première approximation".

[2] La méthode consiste à choisir l'entier q_{opt} minimisant l'expression suivante, fonction de la vraisemblance $p(\{y_n\} / \{\Phi_i\})$ et d'une fonction Q(N) du nombre de vecteurs spectraux :

$$q_{opt} = \underset{p}{\text{Argmin}} \left(-2 \log \left(\max p(\{y_n\} / \{\Phi_i\}) \right) + q \; Q(N) \right) \qquad (1.04)$$

Dans le cas d'un modèle de prédiction linéaire des signaux gaussiens, en notant E_q l'énergie résiduelle du modèle, l'expression se simplifie en N log (E_q) + 2 q. Cette expression est cependant difficilement généralisable au cas vectoriel.

— Un opérateur de reconnaissance sur l'ensemble des matrices $\{\Phi_i\}$ et des vecteurs d'excitation $\{\varepsilon_n\}$ nous a semblé a priori difficile à définir.

1.2. LE MODÈLE A CIBLES SPECTRALES

Le modèle à cibles spectrales est une variante du modèle précédent (cf. formule 1.01). La fonction d'excitation $\{\varepsilon_n\}$ considérée n'est plus un bruit blanc vectoriel centré mais une fonction vectorielle constante par morceaux. Le terme de cible spectrale est employé pour décrire la réponse d'un tel modèle à chaque palier de la fonction d'excitation $\{\varepsilon_n\}$. La valeur de la cible spectrale g_χ correspond à la valeur de $y_{n->\infty}$ quand la fonction d'excitation est une constante χ.

$$g_\chi = (I_p - \sum_{i=1}^{q} \Phi_i)^{-1} \chi \qquad (1.05)$$

(avec I_p la matrice identité d'ordre p)

Une méthode itérative [De Lima 89], en deux étapes, a été proposée pour estimer le nombre Nc et l'emplacement des paliers de la fonction d'excitation, les matrices $\{\Phi_i\}$ et les cibles spectrales $\{g_j\}$ (j = 1,.., Nc). Chaque palier est localisé en examinant la trajectoire résiduelle des paramètres spectraux, obtenue par la différence entre la trajectoire originelle des paramètres spectraux et la sortie du modèle autorégressif vectoriel quand on l'excite avec la fonction formée par les paliers déjà localisés. Le critère d'arrêt de cette méthode itérative est basé sur une minimisation de la trajectoire résiduelle des paramètres spectraux.

Ce modèle serait plus adapté que le modèle précédent pour la reconnaissance de la parole. Un opérateur de reconnaissance possible est de définir une métrique simple (e.g., la distance euclidienne) entre les cibles spectrales et d'étiqueter une cible spectrale inconnue par l'étiquette de la cible de référence la plus proche. L'utilisation de ce modèle se heurte encore à deux des problèmes qui sont ceux du modèle précédent : le choix de la durée de la trajectoire à modéliser et l'ordre optimal du modèle.

1.3. LE MODÈLE ÉVOLUTIF

Pour ce modèle, la suite des vecteurs spectraux $\{y_n\}$, (n = 1 ,.., N), est décomposée sur une base de fonctions $\{\phi_i\}$, (i = 1 ,.., q) fixées a priori. En principe, aucune contrainte n'est imposée aux fonction $\{\phi_i\}$. Néanmoins une base orthogonale représente de manière plus efficace l'évolution temporelle des paramètres en minimisant le nombre de fonctions choisies. Les fonctions choisies sont généralement des polynômes de Legendre, des fonctions trigonométriques ou des fonctions sphéroïdales aplaties. Les vecteurs $\{g_i\}$, (i = 1 ,.., q) représentent la contribution de chaque fonction ϕ_i dans la trajectoire des vecteurs spectraux estimés $\{\hat{y}_n\}$.

$$\hat{y}_n = \sum_{i=1}^{q} g_i \, \phi_i(n) \qquad (1.06)$$

Introduite par L.A. Liporace [Liporace 75], développée par Y. Grenier [Grenier 84], cette approche a donné lieu à de nombreuses applications en traitement de la parole, que ce soit en codage [Chevalier et al. 85] ou en segmentation [André-Obrecht et Delyon 89]. Des expériences préliminaires ont été conduites sur la modélisation et la reconnaissance des transitions consonnes-voyelles [Bouliane et al. 86]: les vecteurs $\{y_n\}$ sont des coefficients cepstraux, ils sont décomposés sur une base formée par les quatre premiers polynômes de Legendre. La phase de reconnaissance consiste à chercher le segment de référence $\{y_n^{j_{opt}}\}$ le plus proche du segment de test $\{y_n^{test}\}$ suivant une métrique $\mathcal{D}$ sur les vecteurs $\{g_i\}$:

$$j_{opt} = \underset{j}{\text{Argmin}}\ \mathcal{D}(\{g_n^j\}\ ,\{g_n^{test}\}) \qquad (1.07)$$

Sur un corpus monolocuteur de 10x18 transitions consonnes-(e) ((le) / (Re) / (je) / (we) / (ye) / (ve) / (fe) / (se) / (xe) / (ze) / (me) / (ne) / (pe) / (te) / (ke) / (be) / (de) / (ge)), le taux de décodage acoustico-phonétique était de 58% pour le premier choix. La métrique $\mathcal{D}$ utilisée était une distance de Mahalanobis sur les vecteurs $\{g_i\}$ concaténés.

polynômes de Legendre	fonctions trigonométriques	fonctions sphéroïdales aplaties
$\phi_0(n) = 1$ $\phi_1(n) = n$ $\phi_i(n) = ((2\ i\ -1)\ n\ \phi_{i-1}(n) + (i-1)\ \phi_{i-2}(n)\ ) / i$	$\phi_0(n) = 1$ $\phi_{2i-1}(n) = \cos(i\ n),\ i > 0$ $\phi_{2i}(n) = \sin(i\ n),\ i > 0$	Largeur de bande $B \in [0, 1]$ Minimisation du nombre de fonctions en fonction de B

Tableau 1.1. Bases de fonctions orthogonales

Un tel modèle peut donner de bons résultats pour la reconnaissance de la parole comme le montrent les résultats précédents. Cependant à notre avis, il présente deux inconvénients :

— Le choix de la base de fonctions orthogonales n'est guidé que par le problème de minimisation du nombre de fonctions choisies. Le fait qu'il s'agisse d'un signal de parole avec des propriétés bien spécifiques par rapport aux autres signaux n'est pas pris en considération.

— Le choix d'opérateur de reconnaissance sur les vecteurs $\{g_i\}$ nous semble a priori difficile à définir.

1.4. LES MACHINES DE MARKOV CACHÉES

J.K. Baker [Baker 75] et F. Jelinek [Jelinek et al. 82] ont montré comment le décodage de la parole pouvait être considéré comme un problème plus général de décodage en théorie de l'information. D'après cette transposition du problème, les codes qui représentent le signal de parole sont émis par une source (e.g., une chaîne de Markov cachée). La source peut être "linguistique" (i.e., modèle de langage) ou "acoustique" (i.e., modèle de l'évolution spectrale). C'est actuellement la méthode reconnue comme la plus performante et dont l'utilisation s'est largement répandue en Reconnaissance Automatique de la Parole [Rabiner et Juang 86][Jouvet 88][Lee et Hon 89][Picone 90].

C'est le système de référence choisi pour évaluer les performances de notre système de décodage acoustico-phonétique. Les raisons de ce choix sont la facilité de l'apprentissage entièrement automatique et les bonnes performances de reconnaissance généralement obtenues.

1.4.1. Le modèle de Markov

Un modèle de Markov caché peut être représenté par un ensemble de NQ états discrets $\{q_i\}$, (i = 1 ,.., NQ) et de transitions $\{t_{ij}\}$, (i,j = 1 ,.., NQ) entre ces états. A chaque transition t_{ij} sont associées deux probabilités : la probabilité de transition a_{ij}, (i,j = 1 ,.., NQ) et la probabilité d'émission B_{ij}, (i,j = 1 ,.., NQ). La probabilité de transition a_{ij} représente la probabilité de passer dans l'état q_j alors que l'on est dans l'état q_i. La probabilité d'émission $B_{ij}(y_n)$ est une fonction de densité de probabilité correspondant à la probabilité d'émettre sur une transition, un symbole acoustique en fonction du vecteur spectral y_n présenté. Cette fonction de densité de probabilité peut être choisie discrète ou continue.

— Densité de probabilité discrète

L'espace acoustique est discret et fini. Un ensemble de vecteurs spectraux représentatif de l'espace acoustique a été quantifié en un ensemble de vecteurs-code $\{V_k\}$, (k = 1 ,.., NV). Chaque densité de probabilité B_{ij}, (i,j = 1 ,.., NQ) est entièrement définie par la connaissance de la probabilité d'émission de chacun des points V_k (ou centroïde).

$$\mathbf{Bij = \{b_{ijk}\} \text{ avec } b_{ijk} = P(V_k/t_{ij}),} \sum_{k=1}^{NV} \mathbf{b_{ijk} = 1} \qquad \mathbf{(1.08)}$$

— Densité de probabilité continue

Le symbole acoustique émis est le vecteur y_n lui-même. Plusieurs choix sont possibles pour les fonctions de densité de probabilité. Les fonctions choisies sont généralement des sommes pondérées de Ng gaussiennes $\{\mathcal{N}(\mu_{ijk};\Sigma_{ijk};y)\}$, (k = 1 ,.., Ng). Dans ce cas, la fonction de densité de probabilité s'écrit :

$$\mathbf{B_{ij}(y_n) = \sum_{k=1}^{Ng} c_k \, . \, \mathcal{N}(\mu_{ijk};\Sigma_{ijk};y_n), \text{ avec } \sum_{k=1}^{Ng} c_k = 1} \qquad \mathbf{(1.09)},$$

où $\mathcal{N}(\mu_{ijk};\Sigma_{ijk};y_n)$ représente la valeur pour le vecteur spectral y_n de la gaussienne $\mathcal{N}(\mu_{ijk};\Sigma_{ijk})$.

Chacune des gaussiennes $\mathcal{N}(\mu_{ijk};\Sigma_{ijk})$ est caractérisée par son vecteur moyen μ_{ijk} et sa matrice de covariance Σ_{ijk}. Une hypothèse réductrice, généralement admise, pour la

fonction $B_{ij}(y_n)$ est de considérer des fonctions de densité de probabilité gaussiennes vectorielles caractérisées par une matrice de covariance $\sum_{ijk}$ diagonale.

1.4.2. L'apprentissage d'un modèle

L'estimation du modèle revient à la recherche, pour chaque transition, des deux ensembles de probabilité $\{a_{ij}\}$ et $\{B_{ij}\}$. Les algorithmes de Baum-Welch [Baum 72] et de Viterbi [Viterbi 67] sont utilisés pour cette estimation. La structure de la machine de Markov (i.e., type de fonction de probabilités, nombre d'états et de transitions) est définie a priori. L'ensemble des travaux sur cette méthode a permis de définir les topologies les plus performantes pour la reconnaissance de la parole. Le modèle de Bakis, par exemple, permet des bouclages sur un même état et des saut d'états. Il permet ainsi de prendre en considération toutes les possibilités de répétition ou d'omission d'un vecteur spectral.

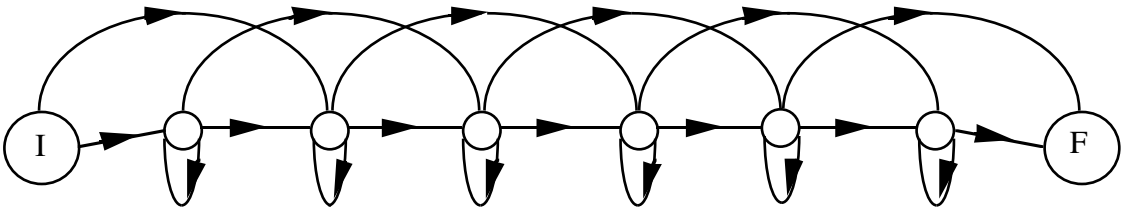


Figure 1.1. Modèle de Bakis

Une carte de reconnaissance de parole utilisant cette technique [Monné 88] est utilisée pour la comparaison expérimentale de notre système de décodage acoustico-phonétique. L'ensemble des protocoles utilisés (i.e., types de modèles, nombre d'états et de transitions) est décrit au chapitre 6.

1.5. LA DÉCOMPOSITION TEMPORELLE CLASSIQUE

La décomposition temporelle classique modélise l'évolution spectrale d'un segment de parole, décrit par les vecteurs spectraux $\{y_n\}$, (n = 1 ,.., N), par un modèle d'interpolation linéaire :

$$\hat{y}_n = \sum_{i=1}^{q} g_i \, \phi_i(n) \qquad (1.10)$$

Le $n^{ième}$ vecteur spectral y_n est estimé par $\hat{y}_n$, c'est-à-dire en tant que combinaison linéaire d'un nombre limité de vecteurs spectraux $\{g_i\}$, (i = 1,.., q), appelés "cibles spectrales". Une fonction d'interpolation à support compact ϕ_i est associée à chaque cible g_i. Les cibles $\{g_i\}$ expriment le contenu spectral du signal, la fonction d'interpolation $\phi_i(n)$ mesure l'influence de la $i^{ème}$ cible pour l'estimation du $n^{ième}$ vecteur spectral y_n. Les fonctions d'interpolation $\{\phi_i(n)\}$ ne sont pas fixées a priori, contrairement à celles utilisées pour le modèle évolutif (cf. paragraphe 1.3).

Le but de la décomposition temporelle est d'estimer à la fois les fonctions d'interpolation $\{\phi_i\}$ et les cibles spectrales $\{g_i\}$. La seule contrainte imposée a priori concerne la forme des fonctions ϕ_i: elles doivent être positives et compactes dans le temps, c'est-à-dire avoir des valeurs non-nulles sur un intervalle de temps fini.

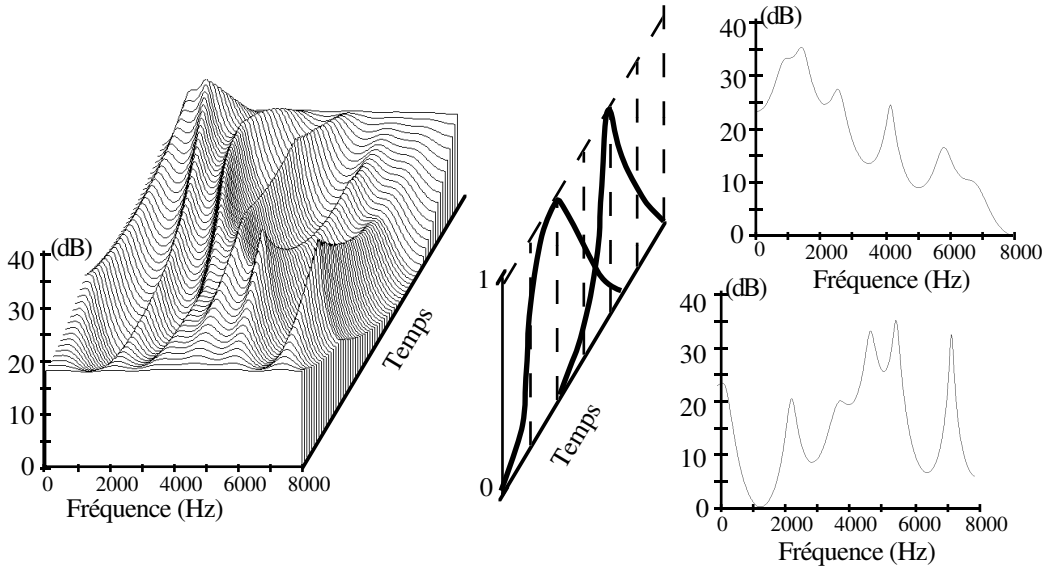


Figure 1.2. Exemple de décomposition temporelle sur le segment (ia)

Introduit comme technique de codage par B.-S. Atal [Atal 83], ce modèle de l'évolution spectrale a été amélioré par S.-M. Marcus [Marcus et Lieshout 84]. A l'origine sa formulation était motivée par le fait que les paramètres calculés sur des fenêtres successives sont fortement corrélés. Il était donc possible de représenter leur évolution par un nombre plus petit de paramètres.

On peut également considérer cette technique comme un moyen de détecter les courbures de la trajectoire des vecteurs spectraux $\{y_n\}$, les cibles $\{g_i\}$ représentant une succession

d'attracteurs locaux (ou points d'accumulations) par lesquels la trajectoire passe ou semble se diriger successivement.

La décomposition temporelle a été développée et utilisée dans différents laboratoires pour de nombreuses applications du traitement de la parole :

— Le codage [Atal 83][Cheng et O'Shaugnessy 90]
— La synthèse [Chollet et al. 86][Bimbot 88],
— Les liens avec les machines de Markov cachées [Niranjan et Fallside 87],
— La segmentation et l'étiquetage [Zuk 84][Marteau 88],
— Le décodage acoustico-phonétique [Marcus et Lieshout 84][Chollet et al. 87] [Bimbot et al. 88][Deléglise et al. 88][Montacié et Chollet 88],
— La recherche d'une représentation plus proche des phénomènes articulatoires [Djik et al. 89].

L'algorithme classique de décomposition temporelle s'effectue en trois étapes :

— Recherche des fonctions d'interpolation les plus compactes dans le temps représentant un événement acoustique "significatif" (cette notion est expliquée plus précisément dans le prochain chapitre).
— Constitution d'un ensemble cohérent et limité de fonctions d'interpolation par sélection et fusion des fonctions de l'ensemble constitué à la première étape.
— Calcul des cibles spectrales.

1.5.1. Principes de résolution [Atal 83]

Soit p l'ordre des vecteurs spectraux, q le nombre de fonctions d'interpolation et t_n l'instant correspondant au vecteur spectral y_n. L'équation (1.10) peut s'exprimer sous la forme matricielle suivante :

$$\mathbf{Y = G\ \Phi,} \qquad \mathbf{(1.11)}$$

où Y est une matrice pxN dont l'élément $Y_{j,n}$ est $\hat{y}_n^j$, G est matrice pxq dont l'élément $G_{j,i}$ est g_i^j, et Φ est une matrice qxN dont l'élément $\Phi_{j,n}$ est $\phi_i(n)$.

Si le rang de la matrice Y est supérieur à q, l'équation (1.11) peut être réécrite sous la forme suivante :

$$\mathbf{\Phi = (G^t\ G)^{-1}\ G^t\ Y}. \qquad \mathbf{(1.12)}$$

Si l'on note W la matrice (qxp) égal à $(G^t G)^{-1} G$ (pseudo-inverse de G) alors :

$$\hat{\phi}_i(n) = \sum_{j=1}^{p} W_{i,j}\ y_n^j \qquad (1.13)$$

Cela veut dire que les fonctions d'interpolation $\{\phi_i\}$ sont estimées par une combinaison linéaire des composantes des vecteurs $\{y_n\}$. La recherche des fonctions d'interpolation consiste à trouver les coefficients $W_{i,j}$ qui garantissent que les $\{\hat{\phi}_i(n)\}$ soient positives et compactes.

Nous allons maintenant exposer la méthode décrite par B.-S. Atal pour le calcul des ϕ_i. Comme nous le verrons dans le prochain chapitre, la méthode que nous employons diffère sur quelques points (e.g., la forme de la fonction caractéristique et le choix des fonctions d'interpolation retenues).

B.-S. Atal recherche d'abord N fonctions ϕ_n, chacune centrée sur l'instant t_n. La contrainte de compacité est ramenée à la minimisation de la fonction $\delta(\phi, n)$. Cette fonction mesure la dispersion de la fonction ϕ à l'instant t_n:

$$\delta(\phi, n) = \sqrt{\sum_{j=1}^{N} (\omega_n(j)\ \phi^2(j)) \ /\ \sum_{j=1}^{N} \phi^2(j)}, \qquad (1.14)$$

appelée fonction de compacité,

avec

$$\omega_n(j) = (j - n)^2,\ 1 \le j \le N, \text{ appelée fonction caractéristique.} \qquad (1.15)$$

$\delta(\phi, n)$ est en quelque sorte une fonction qui mesure le taux de compacité de la fonction ϕ par rapport à la fonction caractéristique $\omega_n(j)$. Cette fonction caractéristique est centrée sur le vecteur y_n: plus la fonction ϕ est compacte et moins ses valeurs, loin de l'instant t_n, ont une amplitude élevée.

Donc

$$\phi_n = \underset{\phi}{\text{Argmin}}\ \delta(\phi, n) \qquad (1.16)$$

Les fonctions d'interpolation $\{\phi_i\}$ (i = 1,.., q) sont une sélection des fonctions $\{\phi_n\}$ (n = 1,.., N). Posons :

$$v(\phi_n, k) = \left(\sum_{j=1}^{N} (k - n)\phi_n^2(j)\right) / \sum_{j=1}^{N} \phi_n^2(j) \qquad (1.17)$$

Les fonctions $\{\phi_n\}$ sélectionnées sont celles pour lesquelles la fonction v (ϕ_n, k) passe d'une valeur positive à une valeur négative au point n : v (ϕ_n, k) est une fonction barycentre qui indique si la fonction ϕ_n est située plutôt dans le passé, cas où v(ϕ_n, k) est positive, ou plutôt dans le futur, cas où v (ϕ_n, k) est négative. Lorsque ϕ_n passe par zéro en décroissant, cela indique que l'on a atteint le centre d'une fonction à sélectionner.

Une fois obtenues les fonctions d'interpolation $\{\phi_i\}$, les cibles spectrales $\{g_i\}$ sont calculées pour minimiser l'erreur de reconstruction E entre la trajectoire spectrale originale $\{y_n\}$ et la trajectoire spectrale originelle estimée $\{\hat{y}_n\}$:

$$E = \sum_{n=1}^{N} ||\hat{y}_n - y_n||^2 \qquad (1.18)$$

On trouve expérimentalement, sur un long signal de parole (i.e., qq minutes), qu'une bonne modélisation peut être obtenue par un nombre q de fonctions $\{\phi_i\}$ en moyenne proche de N/5. C'est-à-dire que la trajectoire spectrale est décrite par N/5 vecteurs spectraux $\{g_i\}$ et par N/5 fonctions d'interpolation $\{\phi_i\}$ au lieu des N vecteurs spectraux $\{y_n\}$ originaux. La décomposition temporelle permet donc une réduction du nombre de paramètres pour la description d'une trajectoire spectrale.

Mais cette compression n'est pas toujours le seul critère : les contraintes d'optimalité de la description d'une trajectoire spectrale doivent dépendre des buts recherchés : dans le cas du décodage acoustico-phonétique les cibles spectrales $\{g_i\}$ doivent être en nombre limité et représenter les "signatures spectrales", signatures que nous espérons plus invariantes que les paramètres spectraux originaux, des phonèmes.

CONCLUSIONS

De l'étude des différents modèles de l'évolution spectrale, nous avons dégagé quelques idées qui ont orienté notre démarche dans le choix du modèle employé pour notre système de décodage acoustico-phonétique :

— Les modèles les plus aptes, a priori, à permettre un décodage acoustico-phonétique performant sont la décomposition temporelle et les machines de Markov cachées. La décomposition temporelle sera utilisée comme prétraitement de notre système de décodage acoustico-phonétique. Les machines de Markov cachées nous serviront de système de référence pour évaluer notre protocole de décodage acoustico-phonétique.

— Tous les modèles de l'évolution spectrale (exceptées les machines de Markov cachées utilisant des densités de probabilités discrètes) font appel, pour leur estimation, à des techniques classiques d'analyse linéaire (régression linéaire multiple, décomposition en valeurs singulières). Cette estimation a généralement comme contrainte principale, la minimisation d'une erreur quadratique entre la trajectoire estimée des paramètres spectraux et la trajectoire originale. Pour permettre l'utilisation d'une mesure de dissimilarité spectrale différente de la distance euclidienne, nous avons développé une extension de la décomposition temporelle classique : la décomposition temporelle généralisée.

CHAPITRE 2. LA DÉCOMPOSITION TEMPORELLE GÉNÉRALISÉE

La décomposition temporelle classique utilise un modèle d'interpolation linéaire et des techniques d'analyse ayant comme critère la minimisation d'une erreur quadratique. L'opérateur de reconnaissance, si l'on veut rester rigoureux, devrait alors être fondé sur la distance euclidienne. Cette distance, si elle est simple à mettre en oeuvre (d'où son succès), n'est pas la plus performante des mesures de dissimilarité spectrale. Notre but n'est pas pour l'instant de rechercher la mesure de dissimilarité la plus efficace, mais d'apporter les outils nécessaires pour utiliser la décomposition temporelle avec n'importe quelle mesure de dissimilarité. Cela permettra peut-être de construire des opérateurs de reconnaissance plus performants. La recherche du couple optimal (représentation paramétrique, mesure de dissimilarité) est alors rendue plus aisée. Il suffit au préalable de définir une mesure de dissimilarité $\mathcal{D}$ adaptée à l'analyse particulière employée.

Nous présenterons d'abord quelques études montrant que la distance euclidienne entre les paramètres spectraux n'est pas la plus performante. Nous décrirons ensuite les bases mathématiques de l'interpolation linéaire généralisée à une mesure de dissimilarité quelconque, en utilisant les géodésiques. Nous verrons que la solution exacte est extrêmement difficile à utiliser. Deux solutions approchées de l'interpolation linéaire généralisée sont alors détaillées : l'une s'appuyant sur la notion de centroïde généralisé, l'autre sur l'analyse factorielle. L'algorithme complet de décomposition temporelle généralisée utilisant l'analyse factorielle est présenté ensuite.

2.1. LES OPÉRATEURS DE RECONNAISSANCE

Déterminer une mesure de dissimilarité entre deux spectres instantanés est un problème fondamental dans l'analyse de parole. C'est un opérateur privilégié pour la plupart des systèmes de Reconnaissance Automatique. Or, la distance euclidienne n'est pas toujours la meilleure mesure de dissimilarité, au sens des performances de reconnaissance, comme le montrent les études suivantes :

— Dans leur étude sur la reconnaissance multilocuteur (4 locuteurs : 2 hommes et 2 femmes) de mots isolés, utilisant la comparaison dynamique comme technique d'alignement temporel, N. Nocerino, F.-H. Soong, L.-R. Rabiner et D.-H. Klatt [Nocerino et al. 85] ont comparé :

— plusieurs mesures dérivées de la prédiction linéaire (d'ordre 8) :

- le logarithme du rapport de vraisemblance,
- le rapport de vraisemblance,
- la distance euclidienne sur les paramètres cepstraux,

— Deux mesures de dissimilarité fondées sur des critères perceptifs :
- le rapport de vraisemblance pondéré,
- la mesure de pente pondérée.

Ces mesures de dissimilarité se classent de la façon suivante dans l'ordre décroissant des performances :

1 - logarithme du rapport de vraisemblance et mesure de pente pondérée,
3 - rapport de vraisemblance,
4 - distance euclidienne sur les paramètres cepstraux,
5 - rapport de vraisemblance pondéré.

— Dans son étude sur la reconnaissance monolocuteur de noyaux vocaliques, M.-J. Caraty [Caraty 87] a comparé :
- le logarithme du rapport de vraisemblance,
- la distance euclidienne sur les coefficients cepstraux en échelle Mel,
- l'Ajustement de Pics Spectraux (APS), une mesure de dissimilarité sur les maxima spectraux, fondée sur des critères perceptifs.

Ces mesures de dissimilarité se classent de la façon suivante dans l'ordre décroissant des performances :

1 - Ajustement des Pics Spectraux,
2 - distance euclidienne sur les coefficients cepstraux en échelle Mel,
3 - logarithme du rapport de vraisemblance.

— Dans leur étude sur la reconnaissance indépendante du locuteur de mots isolés dans une ambiance bruitée, D. Mansour et B.-H. Juang [Mansour et Juang 89] ont comparé plusieurs mesures de dissimilarité sur les coefficients cepstraux dérivés de la prédiction linéaire (d'ordre 8):
- la distance euclidienne,
- la projection cepstrale.

Ces mesures de dissimilarité se classent de la façon suivante dans l'ordre décroissant des performances :

1 - projection cepstrale,
2 - distance euclidienne.

On peut remarquer dans chaque étude, que la distance euclidienne sur les paramètres cepstraux, pourtant l'une des plus utilisées, n'a jamais donné les meilleures performances de reconnaissance. A l'inverse, les mesures de dissimilarité qui sont les plus efficaces ne possèdent pas les propriétés d'une distance!

Nous allons voir maintenant, comment l'interpolation linéaire, définie à partir d'une distance euclidienne, peut être généralisée à une mesure de dissimilarité quelconque.

2.2. LA PROBLÈMATIQUE DE L'INTERPOLATION LINEAIRE GENERALISÉE

La décomposition temporelle classique modélise l'évolution spectrale d'un segment de parole, décrit par les vecteurs spectraux $\{y_n\}$, (n = 1 ,.., N), par un modèle d'interpolation linéaire (cf. équation 1.10). La décomposition temporelle généralisée, pour sa part, doit utiliser un modèle d'interpolation linéaire elle-même généralisée à une mesure de dissimilarité $\mathcal{D}$ quelconque. L'équation (1.10) doit alors être réécrite (cf. équation 2.04 ou équation 2.12). Dans le cas où la résolution de l'équation généralisée n'est pas possible (ou très difficile), nous proposons une approximation du modèle d'interpolation linéaire généralisée en utilisant l'analyse factorielle. Il s'agit de projeter une partie de la trajectoire spectrale dans un nouvel espace, muni de la distance euclidienne, dans lequel on utilisera l'équation (1.10).

2.2.1. Les géodésiques

Dans un espace muni de la distance euclidienne, l'interpolation linéaire entre deux vecteurs est le plus court chemin entre ces deux vecteurs . Ce n'est plus le cas en général quand l'espace est muni d'une autre distance. Cependant le plus court chemin entre deux vecteurs existe toujours : il porte le nom de géodésique.

Nous allons d'abord définir la notion de chemin dans un espace métrique quelconque. Nous allons montrer qu'il est possible de calculer la longueur de ce chemin. Nous allons ensuite exposer la méthode de résolution exacte quand l'espace est $\mathbb{R}^p$ et que la distance choisie est deux fois continûment dérivable. Pour plus de détails on pourra consulter [Lelong-Ferrand 63][Choquet 73].

— Rappels

On appelle ESPACE METRIQUE tout couple constitué par un ensemble $\mathcal{E}$ et une application $\mathcal{D} : (a, b) \rightarrow \mathcal{D}(a, b)$ de $\mathcal{E}$ x $\mathcal{E}$ dans $\mathbb{R}_+$, possédant les propriétés suivantes :

— $(a = b) \Leftrightarrow (\mathcal{D}(a, b) = 0)$.

— $\mathcal{D}(a, b) = \mathcal{D}(b, a)$.

— $\mathcal{D}(a, b) \leq \mathcal{D}(a, c) + \mathcal{D}(b, c)$.

Soit $\mathcal{T}$ un ensemble totalement ordonné (e.g., $\mathbb{R}^p$) et f une application de $\mathcal{T}$ dans un espace métrique. Pour toute partie finie $\sigma = \{t_i, i = 1,.., n)$ de $\mathcal{T}$, on appelle VARIATION de f sur σ le nombre $\mathcal{V}_{f,\sigma}$:

$$\mathcal{V}_{f,\sigma} = \sum_{i=1}^{n-1} \mathcal{D}\,(f(t_i), f(t_{i+1})), \qquad (2.01)$$

Dans le cas d'un espace euclidien, la variation de la fonction f sur l'ensemble des points $\{t_i\}$ correspond à la longueur de la ligne brisée reliant tous les points images par f, dans l'ordre de leur indice.

Pour tout $\mathcal{A}$ inclus dans $\mathcal{T}$, on appelle VARIATION TOTALE de f sur $\mathcal{A}$ le nombre positif, fini ou infini $\mathcal{V}_{f,\mathcal{A}}$, ainsi défini :

$$\mathcal{V}_{f,\mathcal{A}} = \textbf{sup } \mathcal{V}_{f,\sigma} \textbf{ (pour tout les } \sigma \textbf{ finis inclus dans } \mathcal{A}) \qquad (2.02)$$

Lorsque $\mathcal{V}_{f,\mathcal{A}} < \infty$ on dit que f est à VARIATION BORNEE sur $\mathcal{A}$.

Dans le cas où $\mathcal{T}$ est l'intervalle [0, 1] de $\mathbb{R}$ et f une fonction continue, les valeurs de f sur cet intervalle définissent une courbe paramétrée ou CHEMIN. Si $V_{f,\mathcal{T}}$ est à variation bornée, $V_{f,\mathcal{T}}$ est alors la LONGUEUR de ce chemin. Soient a et b, deux points de $\mathcal{E}$. On appelle GEODESIQUES sur $\mathcal{E}$ entre a et b, les chemins de longueur inférieure ou égale à celle de tout autre chemin ayant mêmes extrémités a et b. Une géodésique entre a et b a une longueur $\mathcal{LG}_{\mathcal{D}}(a, b)$ et est définie par les valeurs de la fonction $f_{\text{géodésique}}$ sur l'intervalle [0, 1].

$$f_{\textbf{géodésique}} = \underset{f}{\textbf{Argmin}}\ V_{f,\mathcal{T}}\textbf{, avec } f(0) = a \textbf{ et } f(1) = b \qquad (2.03)$$

Soit la trajectoire spectrale représentée par les $\{y_n\}$ (n = 1,.., N), éléments de l'espace métrique $\mathcal{E}$. La décomposition temporelle généralisée modélise cette trajectoire par un modèle d'interpolation linéaire généralisée à la distance $\mathcal{D}$:

$$\hat{y}_n = \underset{y}{\mathbf{Argmin}} \sum_{i=1}^{q} \phi_i(n) \; [\mathcal{L}_{\mathcal{G}_{\mathcal{D}}} (y \; , g_i \sum_{j=1}^{q} \phi_j(n))]^2 \qquad (2.04)$$

Le $n^{ième}$ élément de la trajectoire spectrale, y_n, est estimé par $\hat{y}_n$, c'est-à-dire comme combinaison linéaire généralisée à la distance $\mathcal{D}$ d'un nombre limité d'éléments de $\mathcal{E}$, les cibles spectrales g_i (i = 1,.., q).

— Nous allons montrer que dans le cas où $\mathcal{E} = \mathbb{R}^p$ et la distance $\mathcal{D}$ choisie est la distance euclidienne ℓ_2, nous retrouvons l'équation (1.10) :

Les éléments y_n, $\hat{y}_n$ et g_i deviennent respectivement les vecteurs spectraux y_n, $\hat{y}_n$ et g_i. Comme $\mathcal{L}_{\mathcal{G}_{\ell_2}}(y \, , g_i \sum_{j=1}^{q} \phi_j(n))$ est égal à $\ell_2 (y \, , g_i \sum_{j=1}^{q} \phi_j(n))$, l'équation (2.04) devient :

$$\hat{y}_n = \underset{y}{\mathbf{Argmin}} \sum_{i=1}^{q} \phi_i(n) \; \ell_2 \; (y \; , \; g_i \sum_{j=1}^{q} \phi_j(n))^2 \qquad (2.05)$$

Soient $\hat{y}_n^k$ et g_i^k, les $k^{ièmes}$ composantes respectives des vecteurs $\hat{y}_n$ et g_i. Comme l'expression $[\ell_2 (y, g_i . \sum_{j=1}^{q} \phi_j(n))]^2$ est différentiable par rapport à $\hat{y}_n^k$, nous pouvons réécrire l'équation (2.05) sous la forme suivante :

$$\partial \Big(\sum_{i=1}^{q} \phi_i(n) \; [\ell_2(\hat{y}_n \; , \; g_i \sum_{j=1}^{q} \phi_j(n))]^2 \Big) \; / \; \partial(\hat{y}_n^k) = 0, \text{ pour } 1 \le k \le p \qquad (2.06)$$

D'où :

$$\sum_{i=1}^{q} \phi_i(n)\ (\hat{y}_n^k - g_i^k \sum_{j=1}^{q} \phi_j(n)) = 0, \text{ pour } 1 \le k \le p. \qquad (2.07)$$

Ce qui donne :

$$\hat{y}_n^k = \sum_{i=1}^{q} \phi_i(n)\ g_i^k \text{ , pour } 1 \le k \le p \text{ et } n = 1 \text{ à } N. \qquad (2.08)$$

Ce qui correspond à l'équation (1.10).

— Dans le cas où $\mathcal{E}$ est égal à $\mathbb{R}^p$ et la distance $\mathcal{D}$ choisie est deux fois continûment dérivable (i.e., de classe C^2). Soient x et y, deux vecteurs de $\mathbb{R}^p$, on peut déterminer les valeurs de la fonction f sur l'intervalle [0, 1] en résolvant un système d'équations différentielles [Lelong-Ferrand 63].

Soit $z = f(t)$, avec $0 \le t \le 1$, un vecteur de $\mathbb{R}^p$ situé sur la géodésique entre les vecteurs spectraux x et y. On désigne par z^i sa $i^{ème}$ composante. Soient les $\mathcal{M}_{ij}(z)$ appelés tenseurs métriques[1]. L'expression de la longueur de la géodésique $\mathcal{LG}_{\mathcal{D}}(x, y)$ est alors la suivante :

$$\mathcal{LG}_{\mathcal{D}}(x, y) = \mathcal{LG}_{\mathcal{D}}(f(0), f(1)) = \int_{t=0}^{t=1} \sqrt{\sum_{i,j=1}^{p} \mathcal{M}_{ij}(z)\ \frac{dz^i}{dt}\ \frac{dz^j}{dt}}\ dt \qquad (2.09)$$

M. Sugiyama [Sugiyama 90] a calculé les matrices des tenseurs métriques $\mathcal{M}_{ij}(z)$ pour différentes mesures de distorsion. Il a cependant renoncé à calculer les géodésiques explicitement et a préféré une solution approchée en employant des techniques de quantification vectorielle.

---

[1] les $\mathcal{M}_{ij}(z)$ sont appelés tenseurs métriques, ils sont calculés en employant la formule suivante :

$$\mathcal{M}_{ij}(z) = \frac{\partial^2 \mathcal{D}(z,u)}{\partial u^i\ \partial u^j}_{u = z} \qquad (2.10)$$

Quand $\mathcal{D}$ est une distance euclidienne, les valeurs de $\mathcal{M}_{ij}(z)$ ne dépendent pas du point z : $\mathcal{M}_{ij}(z)=1$ si i=j sinon 0.

— Quand $\mathcal{E}$ est un espace métrique quelconque, il n'est en général pas possible de calculer $\mathcal{LG}_{\mathcal{D}}(x, y)$ pour x et y, éléments de $\mathcal{E}$: l'équation 2.04 ne peut alors être résolue.

Nous présentons maintenant une approximation de l'interpolation linéaire généralisée, plus simple à employer que l'expression (2.04) en utilisant la notion de centroïde généralisé.

2.2.2. Une approximation de l'interpolation linéaire généralisée

Le centroïde généralisé de a et b est le point c tel que :

$$c = \underset{x}{\mathbf{Argmin}}\ \mathcal{D}^2(x, a) + \mathcal{D}^2(x, b) \qquad (2.10)$$

La solution approchée que nous proposons pour calculer la géodésique entre a et b repose sur l'hypothèse approchée que le point c fait partie lui-même de cette géodésique. Les valeurs de la fonction f pour t variant entre 0 et 1, représentant la géodésique, peuvent être calculées par l'expression suivante :

$$f(\mathbf{t}) = \underset{x}{\mathbf{Argmin}}\ [\mathbf{t}\ \mathcal{D}(x, a)^2 + \mathbf{(1\text{-}t)}\ \mathcal{D}(x, b)^2], \qquad (2.11)$$

pour t variant entre 0 et 1.

— Quelques contre-exemples peuvent être trouvés dans le cas de topologies à trous [Lelong-Ferrand 63]. Mais on peut faire, en première approximation, l'hypothèse que de telles topologies ne se rencontrent pas en traitement de la parole : par exemple, l'ensemble des filtres stables est égal à $[-1\ +1]^p$ dans l'espace des coefficients de réflexions k^i (i = 1,..,p) : ce n'est pas une topologie à trous.

En appliquant cette approximation de l'interpolation linéaire généralisée. On peut réécrire l'équation de la décomposition temporelle, sous la forme suivante :

$$\hat{y}_{\mathbf{n}} = \underset{y}{\mathbf{Argmin}} \sum_{\mathbf{i=1}}^{\mathbf{q}} \phi_{\mathbf{i}}(\mathbf{n})\ [\mathcal{D}\ (y\ , g_{\mathbf{i}} \sum_{\mathbf{j=1}}^{\mathbf{q}} \phi_{\mathbf{j}}(\mathbf{n}))]^2, \qquad (2.12)$$

où l'on note que la distance $\mathcal{D}$ a remplaçé la longueur de la géodésique $\mathcal{LG}_{\mathcal{D}}$.

L'intérêt de cette approximation est qu'elle est plus simple que l'équation exacte (2.04) et qu'elle s'applique, contrairement à l'équation (2.09), à une mesure de dissimilarité quelconque . On peut remarquer que l'on retrouve toujours l'équation de la décomposition temporelle, dans le cas de $\mathbb{R}^p$ muni de la distance euclidienne.

Dans le cas où l'équation (2.12) reste trop difficile à utiliser, nous proposons une autre solution. Il s'agit de projeter une partie de la trajectoire spectrale dans un nouvel espace, muni de la distance euclidienne : l'espace factoriel. Dans cette espace intermédiaire l'équation (1.10) sera utilisée sans modifications pour modéliser la trajectoire des paramètres spectraux.

2.2.3. L'analyse factorielle

Soit la trajectoire spectrale représentée par les $\{y_n\}$ (n = 1,.., N), éléments de l'espace métrique $\mathcal{E}$ muni de la distance $\mathcal{D}$. Soit $I_{\mathcal{D}}(\{y_n\})$ l'inertie de cette trajectoire par rapport à son centroïde généralisé :

$$\mathbf{I_{\mathcal{D}}\ (\{y_n\}) = 1/2 \sum_{i,j=1}^{N} \mathcal{D}\ (y_i, y_j)^2} \qquad \mathbf{(2.13)}$$

L'analyse factorielle [Benzecri 70][Pau 71][Mailles 80][Mokeddem 85] consiste à "projeter" les paramètres spectraux $\{y_n\}$ dans $\mathbb{R}^N$, muni de la distance euclidienne. Le but de cette projection est de conserver au mieux les distances initiales entre les $\{y_n\}$ (i.e., $\mathcal{D}(y_i, y_j)$ i,j = 1 ,.., N) dans l'expression de la distance euclidienne entre les paramètres spectraux projetés $\{y_n\}$ (i.e., $\ell_2(y_i, y_j)$ i,j = 1 ,.., N). Les vecteurs $\{y_n\}$ sont ceux donnés par les expressions suivantes :

$$\mathbf{\{y_n\} = \underset{\{x_n\}}{Argmin} \sum_{i,j=1}^{N} (\mathcal{D}\ (y_i, y_j) - \ell_2\ (x_i, x_j))^2,} \qquad \mathbf{(2.14)}$$

et

$$\mathbf{I_{\mathcal{D}}\ (\{y_n\}) = I_{\ell 2}\ (\{y_n\}).} \qquad \mathbf{(2.15)}$$

Si les distances restent inchangées, les vecteurs $\{y_n\}$ représentent l'image euclidienne de $(\{y_n\}, \mathcal{D})$.

La résolution de ce problème passe par la diagonalisation de la matrice S de taille N x N dont les éléments $\{S_{i,j}\}$ (i,j = 1 ,.., N), sont calculés par la formule suivante :

$$S_{i,j} = -\frac{\mathcal{D}(y_i, y_j)^2}{2} + \frac{1}{2N}\left(\sum_{k=1}^{N}\mathcal{D}(y_i, y_k)^2 + \sum_{k=1}^{N}\mathcal{D}(y_k, y_j)^2\right) - \frac{1}{2N^2}\sum_{k,l=1}^{N}\mathcal{D}(y_k, y_l)^2 \qquad (2.16)$$

Dans le cas où $\mathcal{D}$ est la distance euclidienne, on peut remarquer que la matrice S est la matrice des produits scalaires entre les vecteurs spectraux $\{y_n\}$ centrés.

— Calcul de la projection

La matrice S est diagonalisée sous la forme suivante :

$$S = U^t A^2 U , \qquad (2.17)$$

où A est une matrice NxN diagonale, dont les valeurs propres $\{\lambda_n\}$, (n = 1,.., N) sont généralement positives ou nulles, et U une matrice NxN orthogonale représentant les vecteurs propres de S , $\{u_n\}$, (n = 1 ,.., N).

Les $\{y_n\}$ sont alors calculés par la formule suivante [Benzécri 70] :

$$y_n^j = \sqrt{\lambda_j}\, u_n^j \qquad (2.18)$$

La qualité de la projection factorielle dépend de la matrice S. Elle est parfaite dans le cas où S est semi-définie positive (i.e., toutes les valeurs propres sont positives ou nulles). Deux cas doivent être considérés : la matrice S est semi-définie positive ou non.

— Projection parfaite : S est semi-définie positive

La reconstruction de la matrice des distances est complète. La part d'inertie de la trajectoire spectrale, représentée par les k premiers axes principaux d'inertie, est égale à :

$$\left(\sum_{j=1}^{k}\lambda_j\right) / \text{trace (S), où trace (S)} = \sum_{j=1}^{N}\lambda_j . \qquad (2.19)$$

Cette valeur est comprise entre 0 et 1, et n'est égale à 1 que si les distances sont rigoureusement reconstruites dans l'espace de projection. En général, la plus grande partie de l'inertie de la trajectoire spectrale (à peu près 99%) est contenue dans les dix premiers axes principaux d'inertie.

— Cas général : S n'est pas semi-définie positive

S admet alors des valeurs propres négatives. La projection est imparfaite, la mesure de dissimilarité choisie n'est pas définie à partir d'un produit scalaire, ou même ne vérifie pas les propriétés d'une distance. Si les valeurs propres négatives sont petites en valeur absolue, le critère de la qualité de la projection choisie est :

$$\left(\sum_{j=1}^{k} \lambda_j\right) / \left(\sum_{j=1}^{n} |\lambda_j|\right) \qquad (2.20)$$

Il donne une mesure de la qualité de la projection factorielle sur les k premiers axes d'inertie. Seuls les axes correspondant à des valeurs propres positives sont gardés. Ceux-ci représentent la matrice semi-définie positive T_S la plus proche de S au sens d'une distance euclidienne [Eckart et Young 36]. C'est-à-dire une matrice T_S telle que :

$$T_S = \underset{T}{\text{Argmin}} \|S - T\| = \sum_{i,j}^{N} (S_{i,j} - T_{i,j})^2 \qquad (2.21)$$

La trajectoire spectrale représentée par les $\{y_n\}$ peut être projetée dans $\mathbb{R}^N$.

Nous allons voir au paragraphe 2.2.4 que l'analyse factorielle peut servir pour modéliser cette trajectoire spectrale en utilisant une variante de l'algorithme de décomposition temporelle définie au paragraphe 1.5.

2.2.4. Conclusions et adaptation de la décomposition temporelle à une mesure de dissimilarité quelconque

Nous allons d'abord décrire les modifications apportées à l'algorithme d'Atal pour le calcul d'une fonction d'interpolation ϕ : elles concernent la fonction caractéristique employée $\omega_n(j)$ $1 \le j \le N$, et la fonction de compacité utilisée $\delta(\phi, n)$, $1 \le n \le N$. Soit t_n l'instant correspondant à l'élément y_n de la trajectoire spectrale :

— Nous n'employons pas la fonction caractéristique d'Atal (cf. formule 1.15). Les principales différences sont :

- Notre fonction caractéristique n'est pas définie pour toute la durée N de la trajectoire, mais seulement sur un intervalle $I_1 = [T_1, T_2]$ de taille M. De plus cette fonction est partout nulle sauf sur un sous-intervalle $I_2 = [t_1, t_2]$. Ce sous-intervalle est centré sur l'instant t_n. I_2 est inclus dans I_1.

- La forme de la fonction n'est plus concave, mais convexe.

La fonction caractéristique sera maintenant appelée $\omega_{I1,I2}(j)$, $T_1 \leq j \leq T_2$.

La forme des fonctions caractéristiques $\omega_{I1,I2}(j)$ est variable (cf. Figure 2.1). La plus simple est la fenêtre rectangulaire (i.e., $\omega_{I1,I2}(j) = 1$ sur I_2). Une fonction caractéristique trapézoidale a aussi été utilisée [Ahlbom et al. 87]. Elle est obtenue par squeletisation de la fonction d'interpolation calculée à l'itération précédente. Le coût de calcul de cette méthode est toutefois élevé. D'après la littérature aucune forme ne semble être plus performante qu'une autre. Dans nos expériences, nous avons choisi une fonction caractéristique rectangulaire.

Figure 2.1. différentes fonctions caractéristiques

— De la même façon nous utilisons la fonction de compacité $\delta(\phi, I_1, I_2)$ au lieu de celle définie par Atal (cf. formule 1.14) :

$$\delta(\phi, I_1, I_2) = \frac{\sum_{j=1}^{M} \omega_{I1,I2}^2(j)\, \phi^2(j)}{\sum_{j=1}^{M} \omega_{I1,I2}^2(j) \sum_{j=1}^{M} \phi^2(j)} \qquad (2.22)$$

Puisque la fonction caractéristique est convexe, plus $\delta(\phi, I_1, I_2)$ sera élevé, plus la fonction sera compacte et centrée sur l'instant t_n (rappelons l'intervalle I_2 est centré sur l'instant t_n). Donc $\phi_{I1,I2}$ la fonction choisie (cf. figure 2.2) sera :

$$\phi_{I1,I2} = \underset{\phi}{\mathbf{Argmax}}\ \delta(\phi, I_1, I_2) \qquad (2.23)$$

On peut considérer que $\phi_{I1,I2}$ est la fonction d'interpolation la plus corrélée avec la fonction caractéristique compacte $\omega_{I1,I2}(j)$.

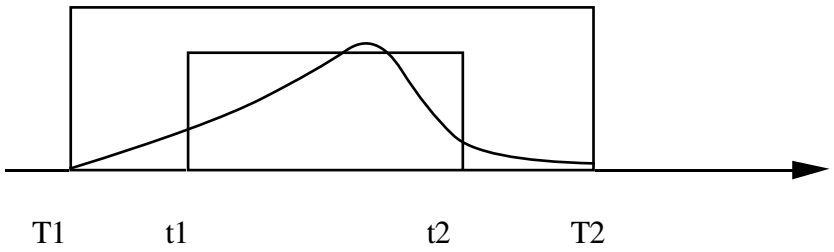


Figure 2.2. Fonction d'interpolation pour deux intervalles $I_1 = [T_1, T_2]$ et $I_2 = [t_1, t_2]$

— Calcul de la fonction d'interpolation $\phi_{I1,I2}$

Nous allons d'abord chercher l'espace le plus adapté à la résolution de l'équation (2.23): il s'agit de l'espace formé par les axes factoriels de la trajectoire spectrale sur l'intervalle I_1 (ou espace factoriel).

Soit la trajectoire spectrale sur l'intervalle I_1 représentée par les éléments de $\mathcal{E}$ $\{y_n\}$, $T_1 \leq n \leq T_2$. On peut projeter cette trajectoire dans $\mathbb{R}^M$, sur les vecteurs $\{y_n\}$, $1 \leq n \leq M$, en diagonalisant la matrice S (cf. équation 2.16) de taille MxM :

$$\mathbf{S = U^t\ A^2\ U,} \qquad \mathbf{(2.24)}$$

où A est une matrice MxM diagonale, et U une matrice MxM orthogonale qui forme la base dans laquelle la trajectoire spectrale sur l'intervalle I_1, est projetée (cf. équation 2.18). Dans cette nouvelle base on peut reprendre l'équation 1.13 et l'écrire sous la forme suivante :

$$\hat{\phi}(n) = \sum_{j=1}^{M} b_j\ U_{n,j}\ , \qquad \mathbf{(2.25)}$$

avec b un vecteur d'ordre M.

Cela veut dire que ϕ est une approximation d'une combinaison linéaire des composantes des axes factoriels $\{U_n\}$. La recherche des fonctions d'interpolation consiste à trouver les coefficients $B_{n,j}$ qui garantissent que $\delta(\phi, I_1, I_2)$ soit maximal.

Soit $\mu = \delta(\phi_{I1,I2}, I_1, I_2)$. En dérivant $\delta(\phi_{I1,I2}, I_1, I_2)$ par rapport à bj, on peut réécrire l'équation (2.23) sous la forme :

$$\sum_{j=1}^{M} \omega_{I1,I2}{}^2(j) \frac{\partial \phi^2(j)}{\partial b^r} = \mu \sum_{j=1}^{M} \frac{\partial \phi^2(j)}{\partial b^r}, \text{ pour } 1 \le r \le M \qquad (2.26)$$

En utilisant l'équation 2.25, nous pouvons écrire :

$$\frac{\partial \phi^2(j)}{\partial b^r} = 2 \sum_{j=1}^{M} b^j \, U_{n,j} \, U_{n,r} \text{ , pour } 1 \le r \le M \qquad (2.27)$$

Ce qui donne :

$$\sum_{j=1}^{M} b^j \sum_{n=1}^{M} \omega_{I1,I2}{}^2(j) \, U_{n,j} \, U_{n,r} = \mu \, b^r \qquad (2.28)$$

Ou sous forme matricielle

$$R \, b = \mu \, b, \qquad (2.29)$$

avec R une matrice MxM :

$$R_{ij} = \sum_{i,j=1}^{M} \omega_{I1,I2}(j) \, U_{i,j} \, U_{i,j} \qquad (2.30)$$

La résolution de l'équation 2.29 revient à chercher le vecteur propre b associé à λ la plus grande valeur propre de la matrice R.

L'algorithme calculant la fonction d'interpolation $\phi_{I1,I2}$ est résumé ci-dessous. Comme nous le verrons au paragraphe 2.3.1, seuls les vecteurs U_j correspondant aux plus grandes valeurs propres λ_j sont utilisés :

Etape 0: — La matrice S est calculée soit directement pour une distance euclidienne (calcul de la matrice des produits scalaires), soit en utilisant la matrice des distances entre les vecteurs de paramètres pour une mesure de dissimilarité.

La trace Tr de la matrice (i.e., la somme des valeurs propres) est calculée.

— j est initialisé à 1.

Etape j: La valeur propre λ_j de module maximal et le vecteur propre correspondant U_j sont calculés par la méthode de la puissance itérée [cf. Annexe IV].
Deux cas sont à considérer :

— Si λ_j est suffisamment faible par rapport à Tr, la matrice R est calculée avec l'ensemble des valeurs propres déjà trouvées $\{U_n\}$ (n = 1 ,.., j-1). L'équation (2.29) est résolue pour trouver l'indice de compacité μ maximum. La fonction d'interpolation correspondante $\phi_{I1,I2}$ est alors calculée et l'algorithme s'arrête.

— Dans l'autre cas, on effectue une déflation de la matrice S (i.e., création d'une matrice comprenant les valeurs propres $\{U_n\}$ (n = j+1 ,.., m)) par la méthode de Hotelling [cf. Annexe IV].
j prend la valeur j+1 et on continue au début de l'étape j.

2.3. ESTIMATION D'UNE FONCTION D'INTERPOLATION SIGNIFICATIVE SUR UN INTERVALLE

Une fonction d'interpolation est significative si elle représente fidèlement les événements acoustiques se produisant dans l'intervalle I_2. La fonction $\phi_{I1,I2}$ est celle qui est la plus corrélée avec la fonction caractéristique $w_{I1,I2}$. C'est celle dont la fonction de compacité est maximale.

Le critère de la fonction de compacité maximale n'est cependant pas suffisant car le nombre de fonctions $\phi_{I1,I2}$ conservées serait trop important. La solution choisie est de réduire l'espace spectral et de choisir un nombre déterminé d'intervalles de calcul (i.e., I_1 et I_2).

2.3.1. Réduction de l'espace spectral

La recherche des "meilleures" fonctions ϕ sur un intervalle de temps [1, N] ne doit pas être seulement guidée par la notion d'indice de compacité maximum. Elle doit faire intervenir des paramètres supplémentaires car dans le cas contraire la solution triviale présentée ci-dessous serait possible :

$$\Phi \ / \ \phi_i \quad (0 \leq i \leq N) \text{ avec } \phi_i(n) = 0 \text{ si } i \neq n, \ 1 \text{ si } i = n. \qquad (2.31)$$

(i.e., un ensemble de fonctions d'interpolation formant un peigne de Dirac)

On peut remarquer que chacune des fonctions d'interpolation ϕ_i prise séparément a un indice de compacité maximum (i.e., égal à 1). Mais cette solution ne réduit nullement le nombre de paramètres décrivant l'évolution spectrale. Une solution simple serait de se fixer une taille minimum pour les fonctions d'interpolation. Elle n'est cependant pas satisfaisante car la durée des événements dans le signal de parole est extrêmement variable.

Le choix d'Atal est de ne pas prendre en compte les axes d'inertie de faible contribution (i.e., associés aux valeurs propres les plus faibles). Les courbures de la trajectoire proches du bruit de la mesure, peu significatives, sont ainsi éliminées. Le choix du nombre d'axes d'inertie gardés peut être fait en prenant un seuil absolu d'inertie [Deléglise 91] ou relatif (i.e., pourcentage donné de l'inertie totale). Le premier critère est stable pour un type d'enregistrement et de paramétrisation donné. Les fonctions d'interpolation obtenues représentent toujours des courbures significatives de la trajectoire spectrale. Un nouveau seuil est cependant nécessaire (en pratique) pour tout changement de prise de son et d'analyse du signal. Le seuil relatif a été préféré pour ne privilégier aucune paramétrisation; il est plus imprécis mais répond mieux à notre approche indépendante du type d'analyse et de la mesure de dissimilarité choisie.

Le critère de maximisation de la valeur μ sur un espace spectral réduit ne suffit pas à assurer la compacité des fonctions d'interpolation obtenues. En pratique, celles-ci sont rendues compactes par l'élimination pure et simple des valeurs faibles ou négatives.

2.3.2. Choix des intervalles de calcul

A chaque couple d'intervalles (I_1, I_2) correspond une fonction d'interpolation ϕ_{I_1,I_2} significative et compacte. Pour éviter de tester tous les couples d'intervalles, un sous-ensemble des couples doit être choisi de manière à détecter tous les événements acoustiques significatifs.

On peut distinguer deux familles d'intervalles: celle à fenêtrage fixe et celle à fenêtrage adaptatif.

2.3.2.1. Méthodes classiques

Deux méthodes ont été principalement étudiées. La première introduite par B.-S. Atal, consiste, pour un intervalle I_1 constant et fixe, à faire glisser les sous-intervalles I_2 du début à la fin de I_1 et à conserver les fonctions d'interpolation centrées sur le milieu de

I_1. Cette méthode a ensuite été améliorée [Marcus et Lieshout 84] [Zuk 84] en prenant un sous-intervalle de I_2 centré sur I_1 glissant. Le choix des fonctions se fait a posteriori, par une sélection hiérarchique. Dans les deux méthodes la taille de I_1 (250 ms) et de I_2 (80 ms) est fixe.

L'inconvénient du fenêtrage fixe est de ne pas détecter tous les événements acoustiques significatifs :

— Les événements de très courte durée (e.g., les plosives) sont masqués par le comportement global des paramètres .

— Deux phonèmes voisins temporellement et de même nature ne sont pas systématiquement décorrélés.

Pour pallier cet inconvénient, le fenêtrage adaptatif [Ahlbom et al. 87][Bimbot et al. 88] [Dijk et Marcus 89] a été introduit.

2.3.3.2. Fenêtrage adaptatif

Bien qu'un peu plus coûteux en temps de calcul, le fenêtrage adaptatif présente l'avantage de centrer automatiquement I_1 sur les principaux événements acoustiques . A partir d'un intervalle de départ $I_2^0 = [t_1^0, t_2^0]$ de taille δ^0, on procède de la façon suivante:

Etape 0 : On initialise $I_1^0 = I_2^0$, et $k = 0$.

Etape k : A l'étape k, on choisit l'intervalle I_1^k de même centre que l'intervalle I_2^k mais de taille $\Delta T^k = a\delta^k + b$ ($1 \leq a \leq 2$, $1 \leq b \leq 5$), avec δ^k la taille de I_1^k.
On recherche la fonction d'interpolation $\phi_{I1^k,I2^k}$ la plus compacte et l'on détermine son support temporel I_ϕ^k après troncature.
Plusieurs cas se présentent alors :

— Si $I_2^k = I_\phi^k$, on conserve la fonction ϕ trouvée et l'algorithme s'arrête.

— Si des périodicités sont trouvées dans la suite des (I_2^k), on conserve la fonction ϕ de plus grande compacité et l'algorithme s'arrête.

— Dans tous les autres cas, k est incrémenté.

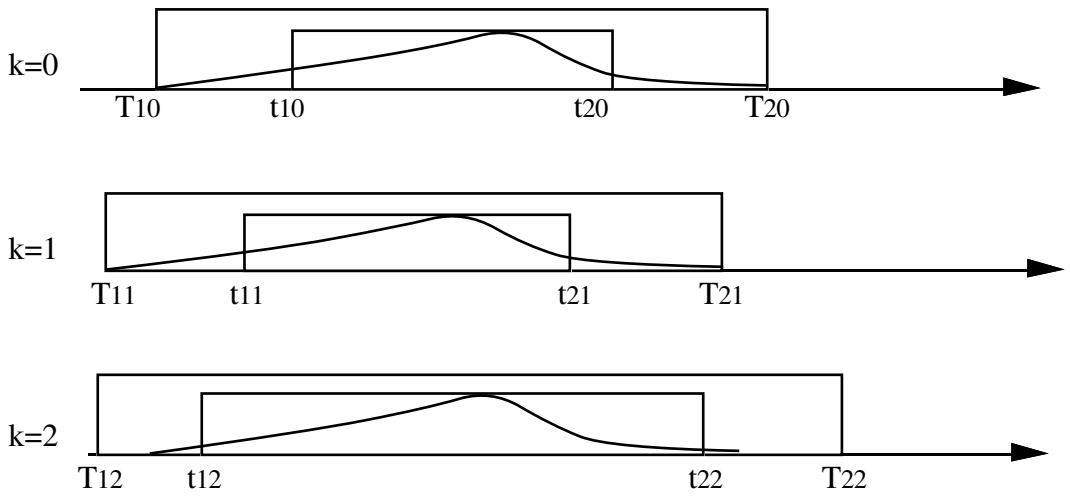


Figure 2.3. Itérations du fenêtrage adaptatif

Le système n'est pas sensible au choix de la taille de l'intervalle de départ δ^0 ($3 \leq \delta^0 \leq 15$). La convergence de l'algorithme n'a pas été démontrée, mais aucune divergence n'a été observée à ce jour. Le nombre moyen d'itérations est cinq. Un maximum d'itérations (i.e., 20) est prévu. Ce chiffre n'a été atteint dans aucune des expériences (i.e., une analyse de quelques dizaines de minutes de signal de la parole).

2.4. CONSTITUTION D'UN ENSEMBLE COHÉRENT DE FONCTIONS D'INTERPOLATION

L'algorithme à fenêtrage adaptatif n'omet pratiquement aucun événement acoustique, il a l'inconvénient en revanche d'en détecter trop. Pour cette raison nous avons dû développer des procédures pour éliminer, fusionner ou diviser les fonctions d'interpolation excédentaires.

Une fois obtenu l'ensemble des fonctions $\{\phi_i\}$ sur un segment de parole [1, N], les fonctions dont le support temporel est trop petit (i.e., quelques centisecondes) ou correspondant à une zone de silence sont éliminées. Une liste L_ϕ est obtenue par tri des fonctions d'interpolation suivant leur centre de gravité[1]. La corrélation $Crl(\phi_i, \phi_j)$ est calculée entre chaque couple de fonctions d'interpolation adjacentes ϕ_i et ϕ_{i+1}:

---

[1]Le centre de gravité Grav (ϕ) d'une fonction d'interpolation $\phi(n)$, définie entre 1 et N, est :

$$\mathbf{Grav}\ (\phi) = \sum_{n=1}^{N} n\ \phi(n) \ /\ \sum_{n=1}^{N} \phi(n) \qquad (2.31)$$

$$\mathbf{Crl}(\phi_i, \phi_j) = \frac{\sum_{n=1}^{N} \phi_i(n)\ \phi_j(n)\ \sum_{t=1}^{N} \phi_i(n)\ \phi_j(n)}{\sum_{n=1}^{N} \phi_i^2(n)\ \sum_{n=1}^{N} \phi_j^2(n)} \qquad (2.32)$$

Dans la procédure d'obtention de l'ensemble des fonctions d'interpolation, trois choix fondés sur les corrélations sont retenus:

Choix 1 : Si les fonctions adjacentes ϕ_i et ϕ_{i+1} sont décorrélées (i.e., $Crl(\phi_i, \phi_{i+1}) < 0.25$), l'une et l'autre sont conservées. C'est le cas le plus courant.

Choix 2 : Si les fonctions adjacentes ϕ_i et ϕ_{i+1} sont extrêmement corrélées (i.e., $Crl(\phi_i, \phi_{i+1}) > 0.75$), la moins compacte est éliminée.

Choix 3 : Si les fonctions adjacentes ϕ_i et ϕ_{i+1} sont proches sans être extrêmement corrélées (i.e., $0.25 < Crl(\phi_i, \phi_{i+1}) < 0.75$). Notre choix est de soustraire la contribution de la fonction la plus compacte à l'autre. Ce cas peut être interprété comme le regroupement de plusieurs événements acoustiques successifs par une fonction de faible compacité. Les règles de troncature sont appliquées : il peut y avoir division d'une onction d'interpolation en deux. Cette solution a été préférée à l'interpolation en termes de cibles non atteintes [Taillanter 88] résolue par l'interpolation géométrique de la trajectoire spectrale.

Après un parcours de la liste L_ϕ, la liste est retriée. On réitère la procédure tant que les choix 2 et 3 se présentent. En général deux ou trois itérations sont nécessaires.

L'ensemble de ces procédures donne un ensemble cohérent de fonctions d'interpolation qui vont permettre le calcul des cibles spectrales.

2.5. LE CALCUL DES CIBLES SPECTRALES

Nous utilisons deux méthodes pour le calcul des cibles spectrales : l'une globale et l'autre locale. La première est initialement appliquée de façon itérative, la seconde intervient pour affiner les résultats. Les cibles obtenues sont des vecteurs de l'espace factoriel local. Il est nécessaire, pour obtenir des cibles dans l'espace initial, de faire une projection factorielle inverse.

— La méthode globale

Il s'agit de minimiser l'erreur de reconstruction des paramètres spectraux sur l'ensemble du signal de parole analysé (cf. équation 1.18).

D'après l'équation (1.10) dans l'espace factoriel local, on a :

$$\mathbf{G = Y\ \Phi^{\#}} \qquad \mathbf{(2.33)}$$

où $\Phi^{\#}$ est la pseudo-inverse de la matrice Φ des fonctions d'interpolation.

Il est alors possible de recalculer l'ensemble des fonctions d'interpolation par la formule :

$$\mathbf{\Phi = G^{\#}\ Y} \qquad \mathbf{(2.34)}$$

Les deux procédures sont répétées alternativement jusqu'à ce que l'erreur de reconstruction se stabilise. La méthode locale est alors utilisée pour affiner les résultats et obtenir les cibles spectrales définitives.

— La méthode locale

La procédure de minimisation de l'erreur de reconstruction est toujours employée, sur chaque fonction séparément, les autres fonctions restant constantes. L'influence des fonctions d'interpolation voisines est ainsi éliminée. Il n'y a pas de recalcul des fonctions d'interpolation.

2.5.1. Projection Factorielle inverse

Le dernier problème à résoudre est l'obtention des cibles $\{g_i\}$ dans l'espace initial $\mathcal{E}$. Il s'agit de retrouver les composantes d'un vecteur dont on connait les coordonnées dans le sous-espace factoriel $\mathbb{R}^N$. Si le problème de la projection est résolu d'une manière simple par l'analyse factorielle, le problème inverse est difficile.

Il peut être formulé de la façon suivante : soit la trajectoire spectrale représentée par les $\{y_n\}$, (n= 1 ,.., N). Sa projection dans l'espace factoriel est représentée par les

vecteurs $\{y_n\}$, (n = 1 ,.., N). Le problème est de trouver la cible g, élément de $\mathcal{E}$, dont la projection suivant les axes factoriels serait g. La réduction de dimension apportée par la projection factorielle ne nous permet pas d'avoir une solution unique : la projection inverse de la cible g est un sous-espace de dimension égale à la dimension initiale moins la dimension de l'espace projeté. Nous avons dû introduire une contrainte supplémentaire : la cible g choisie sera l'élément de la projection inverse de la cible g, le plus proche de la trajectoire spectrale $\{y_n\}$. La contrainte choisie est d'obtenir les mêmes distances dans l'espace projeté (i.e., $\ell_2(g_i, y_n)$) que dans l'espace initial (i.e., $\mathcal{D}(g, y_n)$). C'est-à-dire de calculer g telle que :

$$g = \underset{\{x\}}{\text{Argmin}}\ E(g, \{y_n\}), \qquad (2.34)$$

avec

$$E(g, \{y_n\}) = \sum_{i=1}^{N} (\mathcal{D}(g, y_i) - \ell_2(g, x_i))^2, \qquad (2.35)$$

appelée erreur de reconstruction.

Ce problème de minimisation ne peut être résolu, dans le cas général, que par des techniques itératives d'optimisation numérique. Nos essais montrent qu'une méthode d'optimisation à base de gradient ne donne pas de résultats acceptables. La surface formée par l'erreur $E(g, \{y_n\})$ présente un grand nombre d'accidents (e.g., "ravins") et l'algorithme diverge rapidement. Une méthode d'optimisation basée sur une recherche dans un maillage autour des $\{y_n\}$ a été utilisée. Cette méthode permet une bonne minimisation du critère $E(g, \{y_n\})$, le coût de calcul est toutefois important. Une projection factorielle inverse demande généralement le calcul de quelques dizaines à quelques centaines de distances $\mathcal{D}$.

L'algorithme employé est résumé ci-dessous :

Etape 0: Les $\{y_n\}$ sont triés dans l'ordre décroissant de l'erreur de reconstruction $E(y_n, \{y_n\})$.
La liste $\{L_j\}$ (j = 1 ,.., N) est obtenue avec $L_j = \{j^{ème}\ y_n\}$
i est initialisé à 1.

Etape 1: La valeur de la cible g est initialisé à L_i.
j est initialisé à 1.

Etape 2: g^* est calculé à partir de g par recherche sur un maillage (cf. paragraphe 2.5.1.1).
Si $E(g, \{y_n\})$ est supérieur à $E(g^*, \{y_n\})$ g prend la valeur g^*.
Si $E(g, \{y_n\}) < \varepsilon$ donné on conserve g, fin de l'algorithme.
Si j > Maxj donné, passage à l'étape suivante sinon j est incrémenté.

Etape 3: Si i > N, la recherche est arrêtée. Dans le cas contraire i est incrémenté, retour à l'étape 1.

Le bon fonctionnement de l'algorithme dépend d'un choix judicieux de Maxj et de ε. Leurs valeurs au cours de nos expériences ont été de 50 pour Maxj et de 10^{-3} pour ε.

2.5.1.1. Méthode d'optimisation

La cible g cherchée doit minimiser l'erreur de reconstruction $E(g, \{y_n\})$. Nous admettrons que g est un vecteur à p composantes. Notre algorithme de recherche explore un maillage dans le voisinage de la cible g [Gottsried et Weisman 73]. La taille de la maille est proportionnelle à la norme du gradient de $E(g, \{y_n\})$. Les essais utilisant directement la direction du gradient n'ont pas donné de résultats satisfaisants. La surface de l'erreur de reconstruction $E(g, \{y_n\})$, trop chaotique, entraîne de nombreux phénomènes de divergence. L'algorithme est le suivant :

Etape 0: La taille de la maille est initialisée à $T_m = 10^{-3}$

Etape 1: Les listes $\{\Delta_i^+\}$ $\{\Delta_i^-\}$ (i = 1 ,.., p) sont construites.
$\Delta_i^+ = E(g + \delta_i, \{y_n\})$.
$\Delta_i^- = E(g - \delta_i, \{y_n\})$.
δ_i est un vecteur dont toutes les composantes sont nulles sauf la $i^{ème}$ qui est égale à T_m.
La valeur Δ correspondant à la plus faible valeur des $\{\Delta_i^+\}$ et$\{\Delta_i^-\}$ est calculée.

Etape 2: La norme du gradient de $E(g, \{y_n\})$ est calculée :

$$\| \text{grad } E(g, \{y_n\}) \| = \frac{1}{2 \text{ pas}} \sum_{i=1}^{p} | \Delta_i^+ - \Delta_i^- |.$$

Suivant la valeur de Δ deux cas se présentent alors :

— Δ est plus grand que $E(g, \{y_n\})$. L'exploration du maillage proche n'a pas donné de meilleure solution pour g, T_m est considéré comme trop grand. Il est divisé par deux.

— Δ est plus petit que $E(g, \{y_n\})$. L'exploration du maillage proche a donné une meilleure solution pour g. Si $\| \text{grad } E(g, \{y_n\}) \|$ est suffisamment grand :
t_m est initialisée à $E(g, \{y_n\}) / \| \text{grad } E(g, \{y_n\}) \|$,
sinon t_m reste inchangé.
g prend la valeur $g \pm \delta_i$ correspondant à Δ.

Cet algorithme ne diverge jamais car la valeur de g ne change que dans le cas d'une diminution de $(g, \{y_n\})$. La technique de pas variable permet une convergence rapide. Le nombre de distances à calculer est toutefois élevé (i.e., quelques centaines).

2.6. EXPÉRIENCES PRÉLIMINAIRES SUR LA TECHNIQUE DE DÉCOMPOSITION TEMPORELLE GÉNÉRALISÉE

Les expériences préliminaires ont consisté à comparer les différentes décompositions temporelles du mot "LUROT", épelé en continu (i.e., (èlyèroté)), pour quatre couples (paramétrisation, mesure de dissimilarité) :

— Les Coefficients Ceptraux lissés itérativement en échelle Mel (MFCC) associés à une distance euclidienne. La taille de la fenêtre est de 30 ms.

— Les Logarithmes des Rapports d'Aires (LAR) associés à une distance euclidienne. La taille de la fenêtre est de 20 ms.

— Les coefficients de corrélation associés à une distance euclidienne. La taille de la fenêtre est de 20 ms.

— Les coefficients de corrélation associés à une distance d'Itakura et Saito symétrique. La taille de la fenêtre est de 20 ms.

Pour plus de précisions sur ces paramétrisations et les mesures de dissimilarité associées, on se réfèrera au chapitre suivant.

— La comparaison des différents couples a d'abord porté sur l'adéquation des différentes décompositions temporelles au décodage acoustico-phonétique. Les critères retenus ont été le nombre de fonctions d'interpolation trouvées et leur durée moyenne. Les principaux défauts de la décomposition temporelle sont un trop grand nombre de fonctions d'interpolation (sur-segmentation) et une trop grande durée moyenne (regroupement de plusieurs événements par une seule fonction). Les résultats sont assez semblables que ce soit pour le nombre de fonctions d'interpolation ou leur durée moyenne.

Paramétrisation	MFCC	LAR	Coefficients de Corrélation	
Mesure de dissimilarité	ℓ_2	ℓ_2	ℓ_2	Itakura et Saito symétrique
Nombre de fonctions	13	13	14	15
Durée Moyenne (cs)	15.5	12.1	16.4	14.5

Tableau 2.1. Nombre de fonctions d'interpolation et moyenne de leur durée en fonction de la paramétrisation et la mesure de dissimilarité

— Nous avons testé ensuite les différents couples sur leur capacité à détecter le son (r) dans le triphone (èro) tiré de l'épellation du mot "LUROT" (cf. figure 2.4), ce qui est un problème assez difficile.

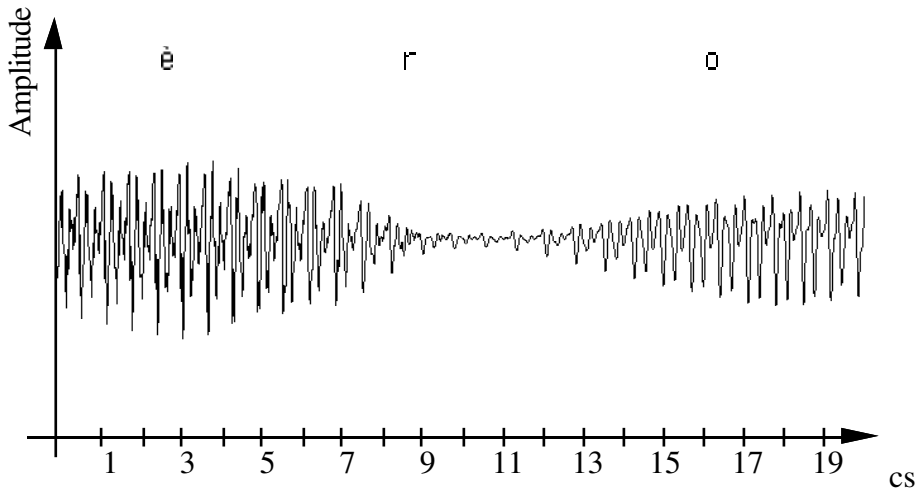


Figure 2.4. Chronogramme du triphone (èro) tiré de l'épellation du mot "LUROT"

Les figures 2.5. à 2.8. représentent la trajectoire des paramètres provenant de l'analyse du triphone (ère) dans le plan factoriel principal. La flèche donne le sens du temps. Le support temporel de la trajectoire représente I_1, c'est-à-dire l'étendue de la grande fenêtre de la décomposition temporelle. Le trait épais représente I_2 la petite fenêtre de la décomposition temporelle (i.e., le support compact non nul de la fonction d'interpolation associé), la petite fenêtre est au centre de la grande fenêtre.

Le critère retenu, pour juger de la qualité de ces décompositions temporelles, est leur aptitude à discriminer la cible spectrale (si elle existe) du son (r) par rapport au reste de la trajectoire spectrale.

On voit sur les figures suivantes que la meilleure décomposition du son (r) a été obtenue par les coefficients de corrélation associés à la distance d'Itakura-Saito. On pouvait cependant le prévoir car cette mesure de dissimilarité est spécifiquement adaptée à la comparaison de masses spectrales (cf. Annexe I). On peut aussi remarquer que le plan principal d'inertie de la trajectoire des coefficients de corrélation associés à la distance d'Itakura et Saito ne contient que 50% de l'inertie totale de la trajectoire.

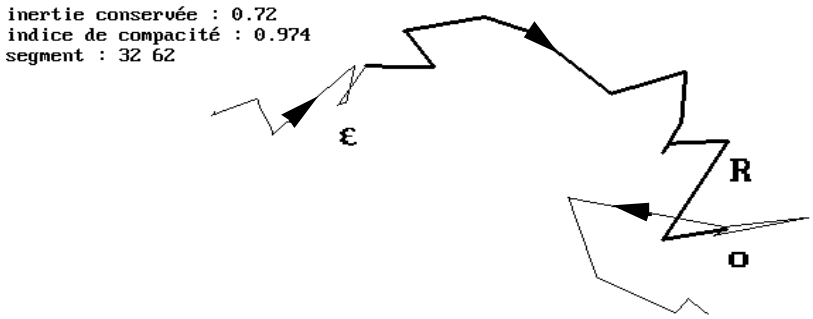


Figure 2.5. Trajectoire dans le plan principal d'inertie pour une paramétrisation MFCC, associée à une distance euclidienne

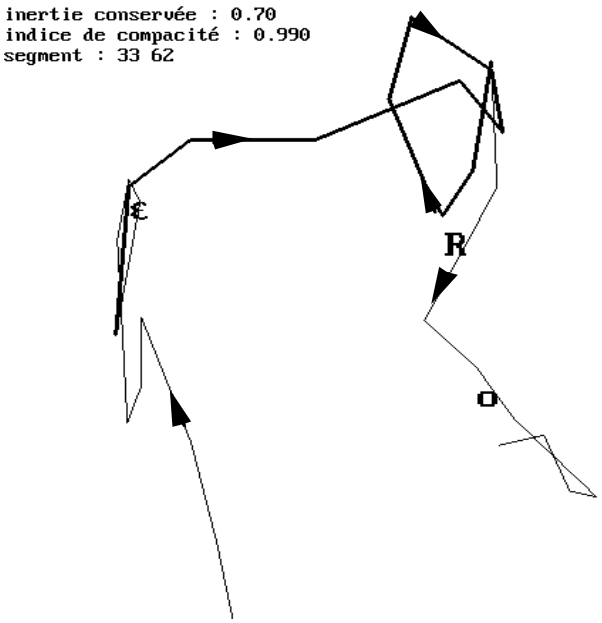


Figure 2.6. Trajectoire dans le plan principal d'inertie pour une paramétrisation LAR, associée à une distance euclidienne

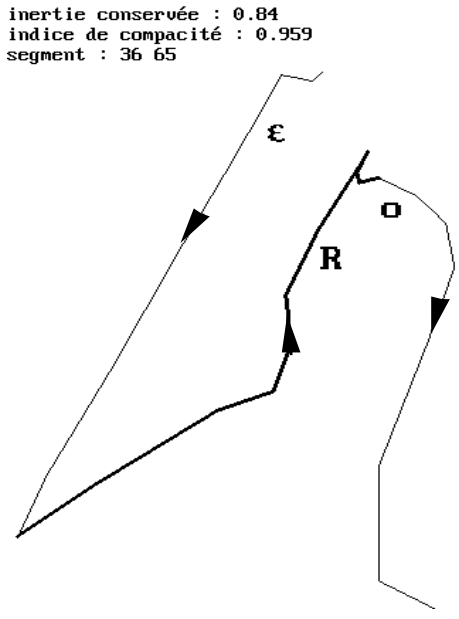


Figure 2.7. Trajectoire dans le plan principal d'inertie pour une paramétrisation avec des coefficients de corrélation, associée à une distance euclidienne

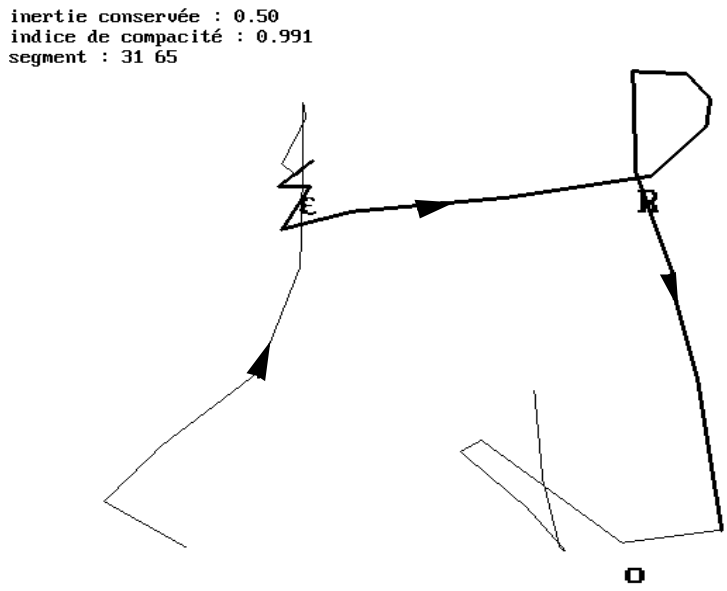


Figure 2.8. Trajectoire dans le plan principal d'inertie pour une paramétrisation avec des coefficients de corrélation, associée à une mesure de dissimilarité d'Itakura et Saito

CONCLUSIONS

L'objectif qui était de parvenir à construire un opérateur de reconnaissance utilisant une mesure de dissimilarité quelconque nous a amenés à définir la décomposition temporelle généralisée.

Pour mettre en oeuvre cette technique, nous avons notamment utilisé l'analyse factorielle et introduit le concept de projection factorielle inverse. Quelques expériences avec plusieurs paramétrisations et différentes mesures de dissimilarité montrent la faisabilité de cette technique. Nous allons maintenant détailler les paramétrisations et les mesures de dissimilarité employées. L'une d'elles retiendra plus particulièrement notre attention : les coefficients cepstraux en échelle Mel avec un lissage cepstral itératif.

CHAPITRE 3. PARAMÉTRISATION SPECTRALE ADAPTÉE AU DÉCODAGE ACOUSTICO-PHONÉTIQUE ET À LA DÉCOMPOSITION TEMPORELLE GÉNÉRALISÉE : IMPLANTATION D'UN LISSAGE CEPSTRAL ITÉRATIF

Le choix de la représentation spectrale pour la décomposition temporelle dans le but du décodage acoustico-phonétique est un problème important. C'est un compromis entre les critères suivants :

— L'interpolation linéaire généralisée des paramètres spectraux doit avoir a priori un sens physique. Il est notamment souhaitable que la transition entre deux phonèmes puisse être représentée comme interpolation linéaire (généralisée) entre les cibles spectrales de deux phonèmes.

— La distance utilisée sur les paramètres spectraux doit être une mesure correcte, au sens de la perception, de la différence entre deux sons.

— Les paramètres doivent être significatifs au sens de l'analyse phonétique et peu sensibles à un changement de locuteur de manière à faciliter les problèmes d'adaptation à de nouveaux locuteurs.

Pour chaque critère, nous avons retenu un jeu de paramètres acceptables : Log Area Ratio (LAR), coefficients de prédiction, Weighted Mel Frequency Cepstral Coefficients (WMFCC), WMFCC avec lissage cepstral itératif. Ces représentations sont décrites avec leurs avantages et désavantages respectifs. Les WMFCC et les LAR sont associés à une distance euclidienne.

3.1. LA PRÉDICTION LINÉAIRE

La prédiction linéaire ou modélisation autorégressive consiste à modéliser le signal de parole par un filtre autorégressif (AR) d'ordre p, H(Z) excité par un signal d'excitation blanc $\{\varepsilon_t\}$. Le signal de parole $\{s_t\}$, (t = 1 ,.., T), est alors représenté par l'équation récurrente suivante :

$$s_t = \sum_{i=1}^{p} a_i \, s_{t-i} + \Gamma \, \varepsilon_t \text{ , avec } \Gamma \text{ le pseudo-gain du filtre} \qquad (3.01)$$

$$H(Z) = \Gamma \, / \, (1 - \sum_{i=1}^{p} a_i \, Z^{-i}) \qquad (3.02)$$

Rappellons que cela consiste à dire que le signal à un moment donné peut être prédit comme combinaison linéaire des p derniers échantillons passés.

La recherche du modèle AR revient à chercher les coefficients de prédiction $\{a_i\}$, (i = 1 ,.., p). Deux algorithmes permettent de calculer les coefficients de prédiction qui minimisent l'erreur moyenne quadratique (i.e., $(\sum_{t=1}^{T} \varepsilon_t^2)$ / T). Il s'agit de la méthode d'autocorrélation et de la méthode de covariance. La première donne un filtre toujours stable, mais présente l'inconvénient d'une moins grande précision temporelle. Elle a été cependant choisie pour la simplicité de sa mise en oeuvre.

La taille des fenêtres d'analyse est de 20 ms (i.e., au moins deux périodes de la période fondamentale). Le prétraitement du signal [Markel et Gray 75] consiste en une préemphase de 90% et un fenêtrage de Hamming.

3.1.1. La paramétrisation LAR (Log Area Ratio)

Les LAR ou Logarithme du Rapport d'Aires sont alors obtenus par la formule suivante :

$$\mathbf{LAR_i = \frac{\log(1+k_i)}{\log(1-k_i)},} \qquad \mathbf{(3.03)}$$

avec $\{k_i\}$ $1 \leq i \leq p$, les coefficients de réflexion

Les paramètres LAR présentent les propriétés suivantes :

— La distance euclidienne sur les paramètres LAR correspond à la moyenne des valeurs absolues des différences entre deux spectres [Viswanathan et al. 75].

— L'interpolation linéaire entre deux sons proches donne des schémas de transitions formantiques réalistes, du moins pour les voyelles [Bimbot et al. 87].

Cependant nos expériences sur la reconnaissance de mots connectés montrent que cette représentation n'a pas de bonnes propriétés discriminantes. Elle est de plus relativement caractéristique du locuteur.

3.1.2. La distance d'Itakura et Saito

Elle compare l'erreur de prédiction d'un signal test prédit par les coefficients de prédiction $\{a_i\}$ au résultat de l'analyse LPC d'un signal de référence [Itakura 75].
L'erreur résiduelle de prédiction α pour le signal de référence $\{sr_t\}$ est la suivante :

$$\alpha = \sum_{t=1}^{T} \left(sr_t - \sum_{i=1}^{p} a_k \, sr_{t-i} \right)^2 \qquad (3.04)$$

Si l'on essaye de prédire le signal test $\{st_t\}$ avec les mêmes coefficients $\{a_i\}$, on obtient comme erreur de prédiction β :

$$\beta = \sum_{t=1}^{T} \left(st_n - \sum_{i=1}^{p} a_k \, st_{t-i} \right)^2 \qquad (3.05)$$

D'après la définition des coefficients de prédiction $\{a_i\}$ (i.e., minimisation de α par rapport aux $\{a_i\}$), α est toujours inférieur ou égal à β.
La distance d'Itakura et Saito D_{IS} consiste à prendre le logarithme du rapport β/α comme mesure de dissimilarité entre le signal test et le signal référence :

$$D_{IS}(\{st_t\},\{sr_t\}) = \log(\beta/\alpha) \qquad (3.06)$$

Cette mesure de dissimilarité est dissymétrique (i.e., $D_{IS}(\{st_t\},\{sr_t\}) \neq D_{IS}(sr_t\},\{st_t\})$)
Une interprétation de cette mesure en termes de rapport de vraisemblance se trouve dans l'Annexe I. Elle montre que cette mesure de dissimilarité est spécifiquement adaptée à la comparaison de masses spectrales.
Nous avons choisi d'utiliser une forme symétrique de la distance d'Itakura et Saito (appelée distance COSH [Miclet et Nehme 79]) dont la formulation est la suivante :

$$D_{COSH}(\{st_t\}, \{sr_t\}) = \sqrt{D_{IS}(\{st_t\},\{sr_t\}) \; D_{IS}(\{sr_t\},\{st_t\})} \qquad (3.07)$$

3.2. LA PARAMÉTRISATION WMFCC (WEIGHTED MEL FREQUENCY CEPSTRAL COEFFICIENTS)

Il n'y a plus modélisation du signal. Après une analyse spectrale de type Transformée de Fourier à court terme (typiquement 32 ms), l'énergie est calculée dans différentes bandes de fréquence au moyen de filtres à enveloppes triangulaires [Davis et al. 80]. La largeur des bandes passantes des filtres suit une courbe isotonique [Zwicker 77] pour tenir compte des phénomènes de perception. Le cepstre est ensuite calculé par la formule 3.08. La distance euclidienne sur les coefficients WMFCC présente de bonnes propriétés discriminantes.

Après une description de l'algorithme de calcul des WMFCC, nous montrerons que cette paramétrisation est sensible aux changements de la fréquence fondamentale du locuteur. Pour pallier ces inconvénients, un lissage cepstral itératif [Imai et Abe 81] du spectre initial est employé.

3.2.1. Calcul des WMFCC

Le spectre de Fourier à la base de la représentation, est calculé à partir d'une fenêtre d'analyse et traité par un banc de 24 filtres répartis sur une échelle de fréquence Mel. Les coefficients sont calculés sur la séquence $\{E_k\}$ (k = 1,.., 24), où E_k est l'énergie correspondant au $k^{ième}$ filtre.

$$cc_j = \sum_{k=1}^{24} \log (E_k) \cdot \cos [j (k - 1/2) \pi/24], \quad (j = 1 ,.., m) \quad (3.08)$$

L'échelle Mel est une échelle faite à partir de mesures perceptives [Zwicker 77]. Elle est linéaire jusqu'à 400 Hz (i.e., 400 Mels = 400 Hz). Au delà les mesures sont données dans le tableau ci-dessous :

Mel	500	600	700	800	900	1000	1100	1200	1300	1400	1500	1600	1700	1800	1900	2000	2100
Hz	510	630	770	920	1080	1270	1480	1720	2000	2320	2700	3150	3700	4400	5300	7700	9500

Tableau 3.1. Echelle de fréquence Mel

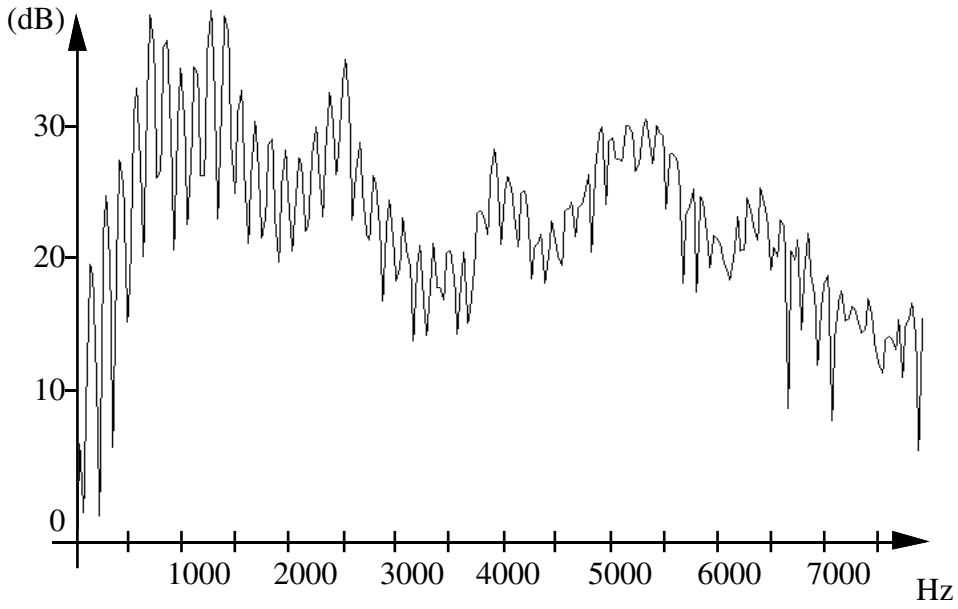


Figure 3.1. Spectre de la voyelle (a) avec une Echelle fréquentielle linéaire

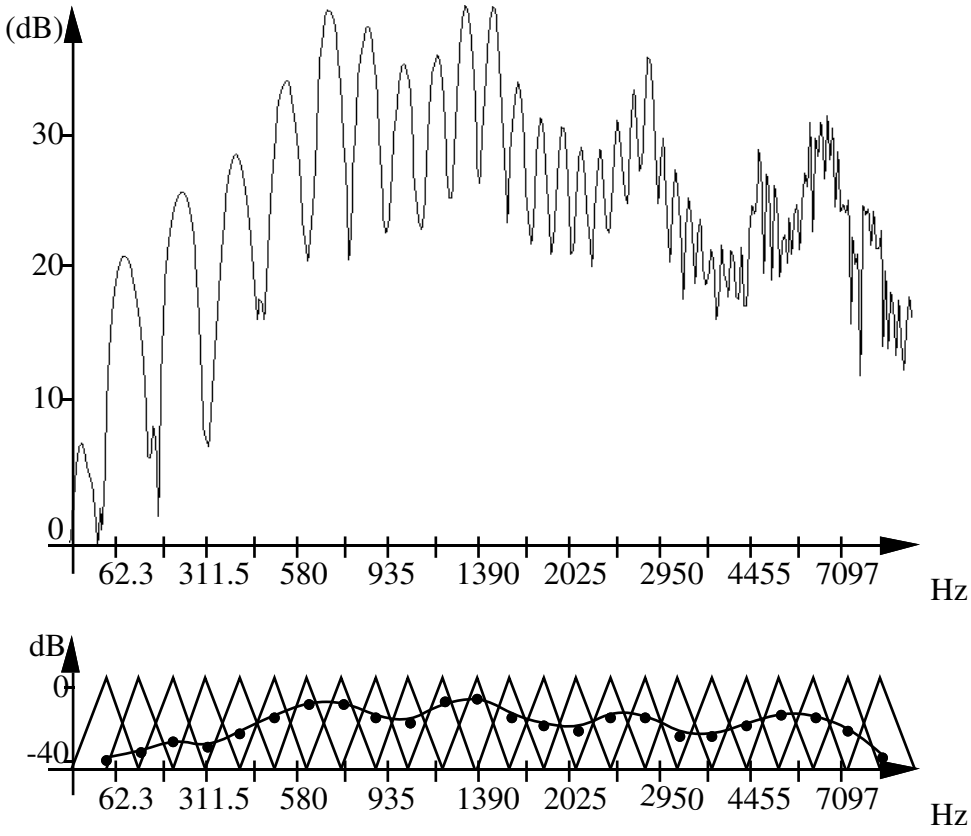


Figure 3.2. Spectre de la voyelle (a) et Banc de filtres avec une Echelle fréquentielle Mel

On peut remarquer sur la figure 3.2. que dans les basses fréquences (i.e., fréquences inférieures à 1000 Hz), la largeur des bancs de filtres est du même ordre de grandeur que les fluctuations du spectre dues à la contribution de la source. L'énergie calculée dans les premiers bancs de filtres est donc au moins autant influencée par la forme du conduit vocal que par la fréquence fondamentale de la source. Comme il est préférable d'obtenir une paramétrisation la plus indépendante possible des variations de la source, nous avons utilisé et développé le lissage cepstral itératif.

3.2.2 Lissage cepstral itératif

Le lissage cepstral classique sépare la contribution de la source de celle du conduit vocal. Après une détection de la fréquence fondamentale, on sépare sur le cepstre les deux contributions par une simple fenêtre rectangulaire.

Cette méthode n'enlève pas les fluctuations de la structure fine du spectre, particulièrement sensible dans le bas de l'échelle logarithmique, ces fluctuations sont dues, pour les voyelles, à la non-synchronisation de la période de la fréquence fondamentale avec celle de la fenêtre d'analyse.

La technique de lissage cepstral itératif [Imai et Abe 81][Halle 83] permet de s'affranchir de ces problèmes. Elle construit une "courbe idéale" qui passe par les maxima spectraux et minimise la contribution des valeurs spectrales fluctuant trop dans le bas du spectre. Cette méthode est extrêmement lourde en temps de calcul (temps de calcul dix fois supérieur au temps d'une simple transformée de Fourier). Mais la trajectoire spectrale obtenue est beaucoup moins erratique comparée aux WMFCC classiques. L'adaptation aux locuteurs entre homme et femme est rendue plus aisée (cf. chapitre 5).

Un algorithme de calcul direct [Galas et Rodet 90] vient d'être proposé. Cette méthode appelée "cepstre discret" a un coût de calcul inférieur à celles des WMFCC classiques. Elle n'a pas été utilisée, n'étant pas disponible lors de notre étude.

3.2.2.1. Algorithme de détection et calcul de la fréquence fondamentale

Les sons de la parole dits voisés sont produits par la vibration des cordes vocales. La mesure de la fréquence fondamentale (F_0) est une estimation à court terme de la pseudo-fréquence de cette vibration. Cette mesure peut se faire dans le domaine fréquentiel ou temporel [Hess 83]. Les principales difficultés résultent principalement de

la grande plage possible de variations (e.g., 60 - 450 Hz) et des interactions entre le conduit vocal et la source glottique. Les erreurs typiques dans la recherche de la fréquence fondamentale sont les erreurs d'harmonique (i.e., choix d'une fréquence n fois supérieure ($n \geq 2$) à celle recherchée). Nous avons conçu notre algorithme de calcul de la fréquence fondamentale pour éviter au maximum ce type d'erreur. La méthode employée est une amélioration de l'algorithme de Noll [Noll 67] fondée sur le cepstre. Elle est résumée ci-dessous pour un son échantilloné à 16 kHz et une largeur de la fenêtre d'analyse de 504 points :

Etape 0: Les coefficients cepstraux $\{c_i\}$ correspondant à la plage 70 -250 Hz sont calculés. Cela correspond à $\{c_i\}$ (i = 267 ,.., 64) pour un son échantillonné à 16 kHz.

Etape 1: La liste $\{c'_j\}$ (j = 1 ,.., 204) est obtenue par tri des $\{c_i\}$ par valeurs décroissantes. Pour chaque c'_j, on définit son amplitude $amp(c'_j)$ et son abcisse temporelle $temps(c'_j)$ (i.e., 64 ,.., 207).

Etape 2: Le premier k, tel que $amp(c'_k) < amp(c'_1/2)$, est recherché. Plusieurs cas se présentent alors :

— k est grand (i.e., $k \geq 25$) la fenêtre est déclarée non-voisée, la valeur de la fréquence fondamentale est fixée à 70 Hz, fin de l'algorithme.

— k est petit (i.e., $k < 25$) la fenêtre est déclarée voisée. Passage à l'étape suivante.

étape 3: C'est la recherche des harmoniques possibles de la fréquence fondamentale.

Calcul de la liste des$\{c''_l\}$ (j = 1 ,.., k) à partir des $\{c'_j\}$.

avec $c''_l = \sum_{m=1}^{k} amp(c'_m) \cdot modf(temps(c'_m), temps(c'_l))$.

où $modf(a,b) =$ (a modulo b) / b si b < a.
(b modulo a) / a si a > b.

La fréquence fondamentale choisie est celle correspondant au plus grand des $\{c''_l\}$.

La fréquence fondamentale maximale a été choisie à 250 Hz pour éviter qu'une sur-estimation de la fréquence fondamentale provoque un filtrage des premiers formants. Une éventuelle sous-estimation de la fréquence fondamentale nous semble moins préjudiciable pour le décodage acoustico-phonétique.

3.2.2.2. Lissage cepstral classique

Soit $\{x_k\}$ (k = 1,..T) une suite de T échantillons de signal après fenêtrage.

$\{X_k\}$ (k = 1,..T/2) = DFT($\{x_i\}$) sa transformée de Fourier discrète,

$\{S_k\}$ (k = 1,..T/2) où S_k = logarithme (S_k).

$\{c_k\}$ (k = 1,..T/2) = DFT^{-1}($\{S_k\}$) est la suite de coefficients cepstraux.

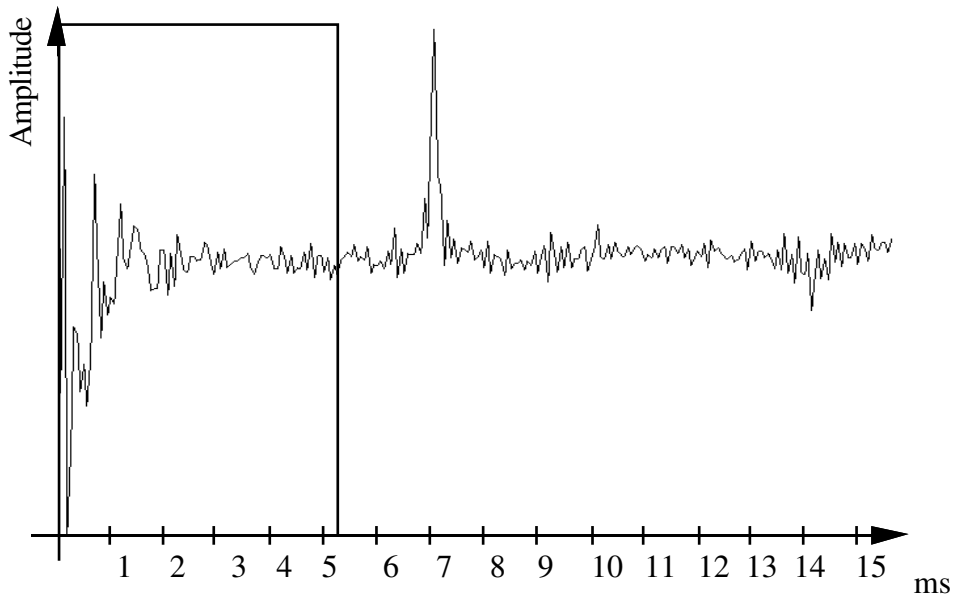


Figure 3.3. Cepstre de la voyelle (a) et gabarit du lissage cepstral

Soit h_k une fenêtre cepstrale symétrique par rapport à T/2 (i.e., $h_k = h_{M-k}$) avec $h_k = 0$ pour $M \leq m \leq T/2$.

$\{C_k\}$ (k = 0,..T-1) = h($\{c_k\}$) où $C_k = c_k\, h_k$

$\{\tilde{S}_k\}$ (k = 1,..T/2) = DFT($\{C_k\}$) est l'enveloppe obtenue par lissage cepstral.

En prenant M suffisamment petit on élimine les composantes dont la pulsation correspond à la structure harmonique du spectre (cas des sons voisés). Cette technique est satisfaisante dans le cas d'une voix masculine (e.g., fréquence fondamentale faible). Son utilisation sur des voix féminines (e.g., fréquence fondamentale élevée) est beaucoup plus délicate, car on constate expérimentalement que l'amplitude des formants est beaucoup

plus sous-estimée (cf. figure 3.6). Pour contourner ces difficultés un lissage cepstral itératif est proposé.

3.2.2.3. Lissage cepstral itératif

Il s'agit d'une extension de la méthode précédente qui vise à construire une enveloppe spectrale autour des pics stables. Le processus est itératif : il cherche à minimiser les écarts positifs entre la (DFT (x)) et l'enveloppe obtenue par lissage cepstral.

Etape 0: $V^{(0)} = \tilde{S}$ (lissage cepstral classique)

$E^{(0)} = g(S - \tilde{S})$

où $g(Y) = Y$ si $Y > 0$

$g(Y) = 0$ si $Y \leq 0$

$E^{(0)}$ représente les écarts positifs de S sur $\tilde{S}$.

$\tilde{E}^{(0)}$ est le lissage cepstral des écarts $E^{(0)}$.

$V^{(1)} = V^{(0)} + \tilde{E}^{(0)}$

Etape k: Soit $V^{(k)}$ l'enveloppe obtenue à l'itération précédente, $E^{(k)}$ et $\tilde{E}^{(k)}$ les écarts et lissage des écarts correspondants.

On calcule :

$$\mathbf{V^{(k+1)} = V^{(k)} + (1+\alpha)\, E^{(k)}} \qquad \mathbf{(3.09)}$$

$$\mathbf{E^{(k+1)} = g(E^{(k)} - (1+\alpha)\, \tilde{E}^{(k)})} \qquad \mathbf{(3.10)}$$

où α est un coefficient appelé "accélérateur de convergence".

Plusieurs cas se présentent alors :

— si $E^{(k+1)} < \varepsilon$ fixé, on garde le spectre lissé $V^{(k+1)}$ et l'algorithme s'arrête.

— Si k > MaxIter fixé, on ne conserve pas le spectre lissé, on reprend le spectre original, l'algorithme de lissage cepstral irératif n'a pas convergé; on peut remarquer qu'aux cours de nos expériences, ce cas ne s'est jamais produit.

— Dans tous les autres cas, k est incrémenté.

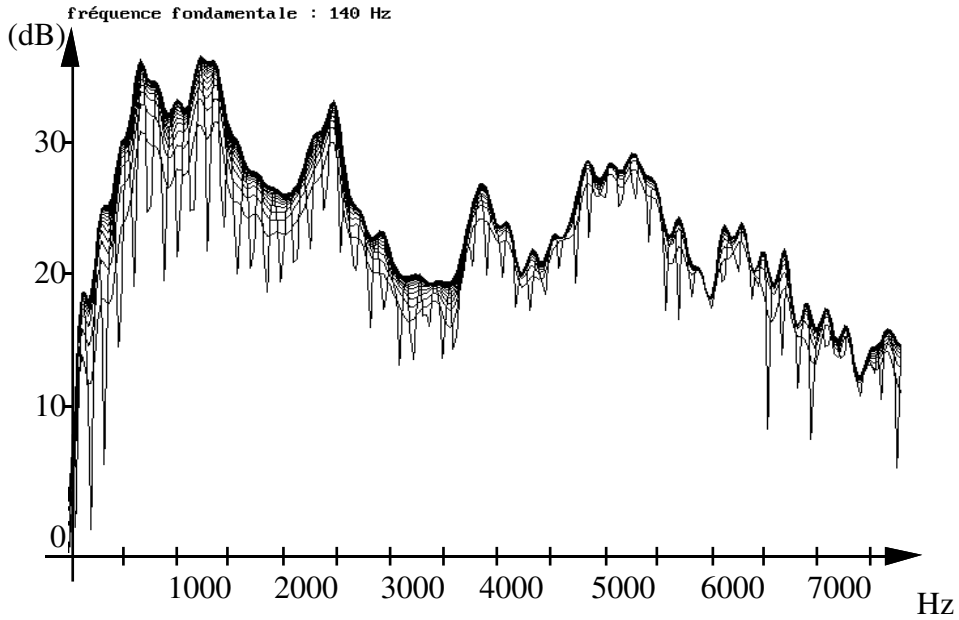


Figure 3.4. Lissage cepstral itératif pour un spectre de la voyelle (a) pour une voix masculine

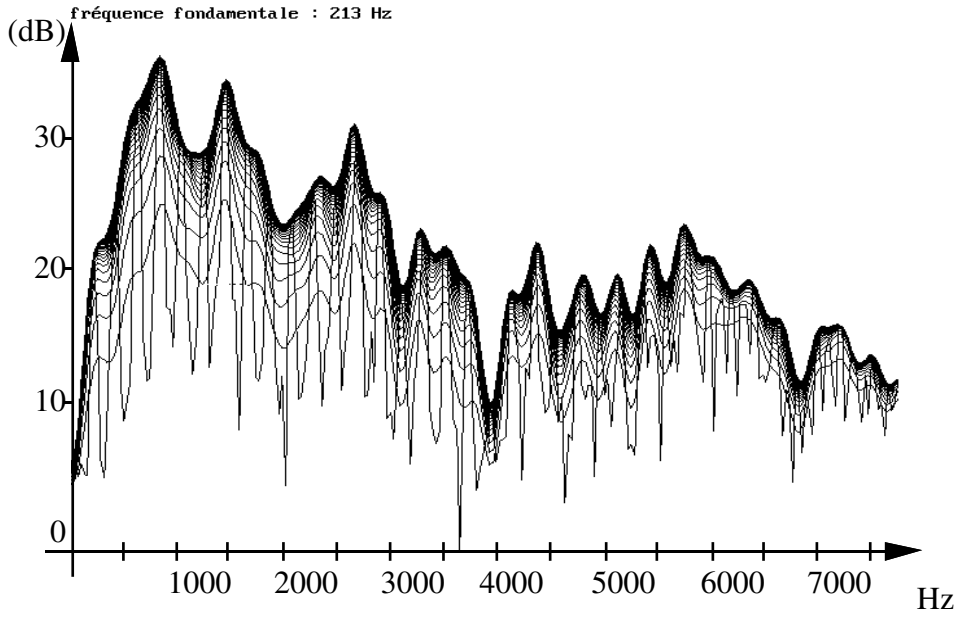


Figure 3.5. Lissage cepstral itératif pour un spectre de la voyelle (a) pour une voix féminine

Chaque spectre lissé correspond à une itération. Le spectre de plus faible énergie correspond au lissage cepstral classique. Le lissage cepstral classique élimine bien les fluctuations dans le bas de l'échelle logarithmique. L'amélioration apportée par cette

technique est plus évidente dans le cas des voix féminines que dans le cas des voix masculines.

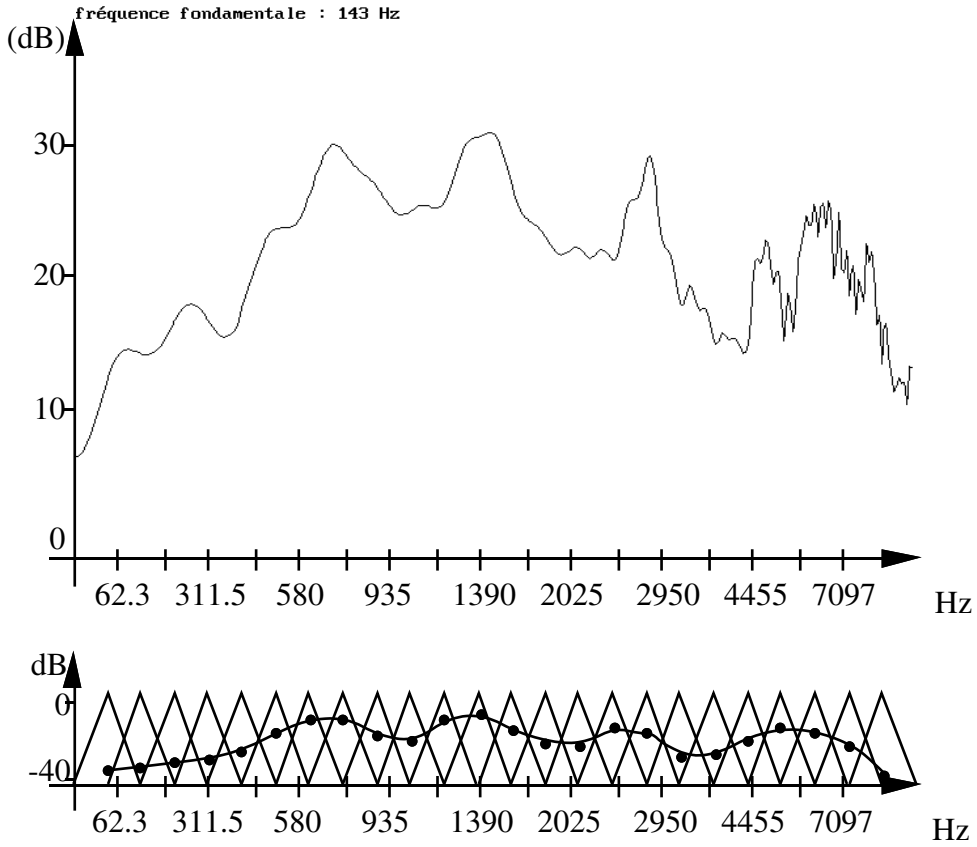


Figure 3.6. Lissage cepstral itératif pour un spectre de la voyelle (a) pour une voix masculine avec une Echelle fréquentielle Mel et un Banc de filtres triangulaires

Les fluctuations du spectre ont été sensiblement atténuées dans les basses fréquences. L'utilisation du banc de filtres apparait alors plus appropriée.

3.3. CALCUL D'UN INDICE D'INSTABILITÉ DE LA PARAMÉTRISATION

Un indice d'instabilité IndStab(t) est obtenu à partir d'une analyse glissante du signal de parole. Un vecteur de paramètres x(t) est calculé à chaque déplacement de la fenêtre d'analyse d'un échantillon. La comparaison de la variance locale et de la variance globale de chaque coefficient permet d'obtenir un indice d'instabilité.
Cet indice d'instabilité peut aussi nous indiquer si la paramétrisation employée représente bien le signal vocal qu'elle est censée modéliser : ce qui n'est plus le cas quand l'indice d'instabilité est élevé :

$$\mathbf{IndStab(t)} = \sum_{i=1}^{p} \frac{\mathbf{Var_L\ (x_i(t))}}{\mathbf{Var\ (x_i)}}, \qquad \mathbf{(3.11)}$$

où p représente la taille du vecteur de paramètre x. $Var_L\ (x_i(t))$ représente la variance locale du $i^{ème}$ coefficient du vecteur de paramètre x à l'instant t. Elle est obtenue de la manière suivante, en prenant T_v la taille de la fenêtre sur laquelle est calculée la variance :

$$\mathbf{Var_L\ (x_i(t))} = \Big(\sum_{i=-T_v/2}^{T_v/2} \mathbf{x_i(t)} \Big) \mathbf{\ /\ T_v\ -\ } \Big(\sum_{i=-T_v/2}^{T_v/2} \mathbf{x_i(t)^2} \Big) \mathbf{\ /\ T_v^2} \qquad \mathbf{(3.12)}$$

Var (x_i) représente la variance globale du $i^{ème}$ coefficient du vecteur de paramètres x. Elle est obtenue de la même façon que la variance locale en prenant pour la taille T_v, une durée beaucoup plus longue (e.g., sur le signal tout entier). L'estimation de la variance locale $Var_L\ (x_i(t))$ est faite sur trois fenêtres d'analyse (30 ms).

La mise en correspondance des courbes de l'indice d'instabilité obtenues pour chaque méthode de paramétrisation permet de déterminer leur variation à un changement de la position de la fenêtre d'analyse. Il semble que sur la figure ci-dessous que l'indice d'instabilité est assez élevé pour la paramétrisation WMFCC (particulièrement dans les zones fricatives). Le lissage cepstral itératif n'apporte visiblement pas de grandes modifications sur ce point.

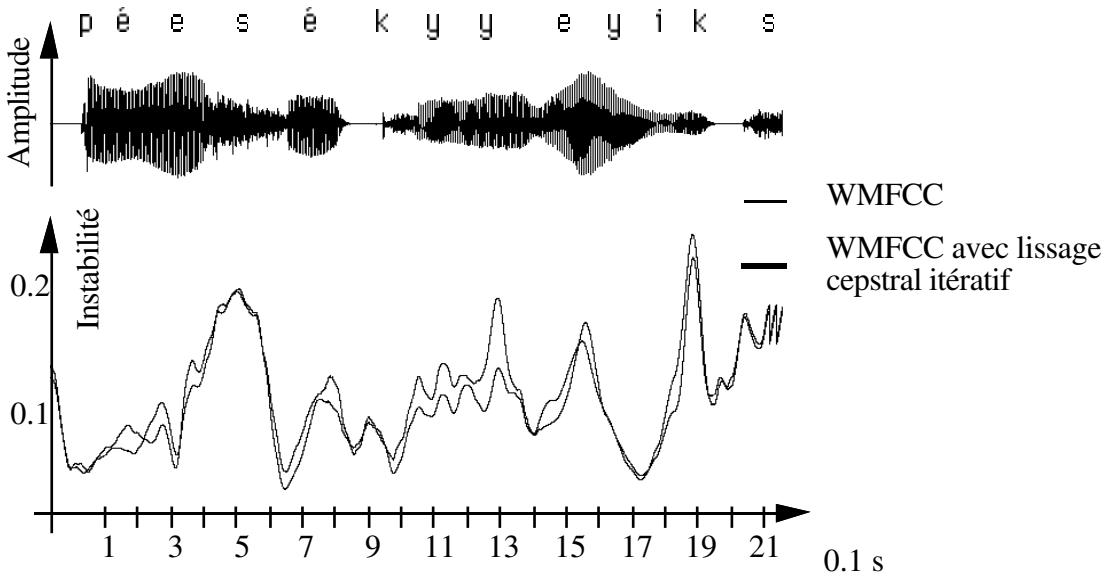


Figure 3.7. Chronogrammes de l'indice d'instabilité du mot épelé en continu "PECQUEUX" pour deux paramétrisations, WMFCC et WMFCC avec lissage cepstral itératif

CONCLUSIONS

L'étude des imperfections des paramétrisations généralement employées dans les systèmes de reconnaissance nous a amené à développer et à implanter un algorithme de calcul de coefficients cepstraux en échelle Mel avec un lissage cepstral itératif. Pour vérifier l'hypothèse que cette paramétrisation était plus stable, relativement au changement du point d'analyse, que la paramétrisation par les coefficients cepstraux classiques, nous avons introduit la notion d'indice d'instabilité de la paramétrisation. Mais cet indice ne s'est pas révélé significatif des différences entre les deux paramétrisations.

En revanche la nouvelle paramétrisation est moins sensible au changement de la mélodie que les paramètres cepstraux classiques. Elle va d'ailleurs nous permettre d'obtenir de meilleures cibles spectrales pour notre processus de décodage acoustico-phonétique.

CHAPITRE 4. LE PROCESSUS DE DÉCODAGE ACOUSTICO-PHONÉTIQUE

La décomposition temporelle fournit un ensemble de signatures spectrales (ou cibles) que l'on espère moins dépendantes d'événements locaux non pertinents que les paramètres originaux et qui doivent se prêter facilement au processus de décodage acoustico-phonétique. Les cibles du corpus de test sont identifiées par comparaison à un dictionnaire de cibles de référence obtenues à partir d'un corpus d'apprentissage représentatif de l'application.

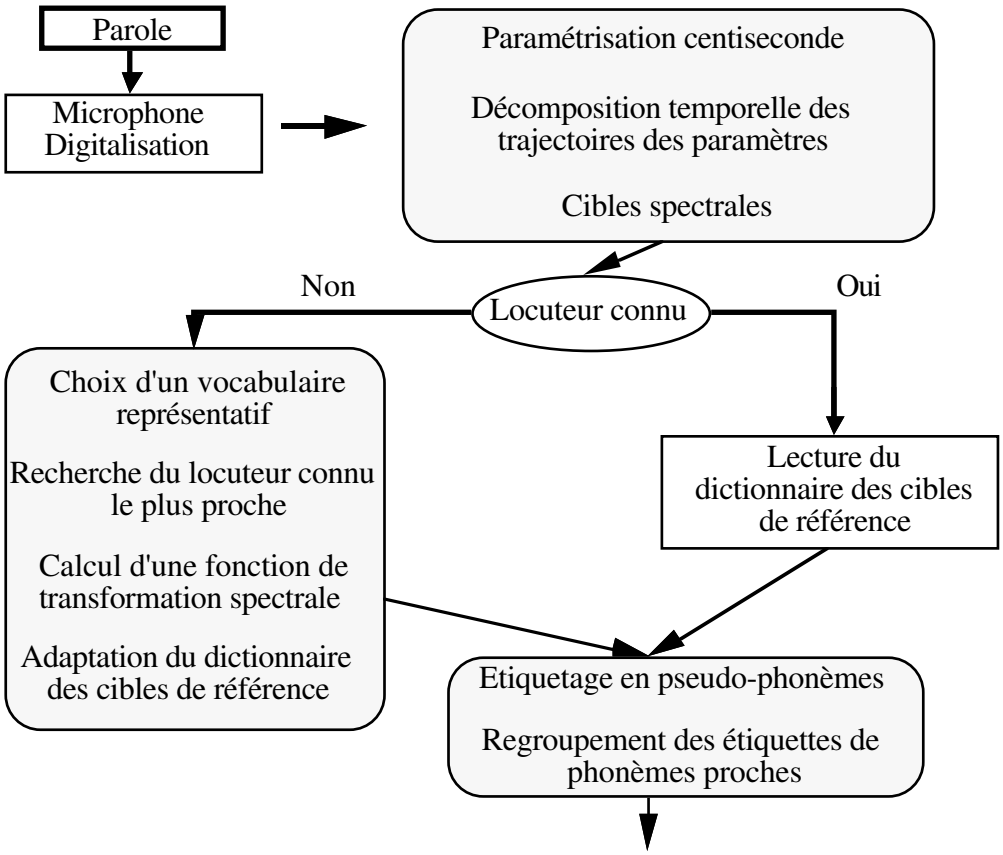


Figure 4.1. Synoptique du processus de décodage acoustico-phonétique

Nous exposerons d'abord les raisons qui ont guidé le choix du classificateur utilisé dans notre processus de décodage acoustico-phonétique. La constitution du dictionnaire des cibles de références est ensuite décrite. L'utilisation de paramètres complémentaires, pour sélectionner les cibles dues à des événements contextuels (transitions), est expliqué.

4.1. CHOIX DU CLASSIFICATEUR.

Les cibles spectrale du corpus de test sont identifiées par comparaison à un dictionnaire de cibles de référence. De nombreuses techniques permettent, à partir d'un ensemble d'apprentissage, la classification statistique de formes inconnues [Dubuisson 90]. L'approche statistique se sert de l'étude statistique des mesures que l'on a effectuées sur les objets à reconnaître. Les techniques statistiques se réfèrent à la règle de décision optimale : la règle de Bayes.

4.1.1. La règle de décision bayésienne

Soit $\{F_i\}$ (i = 1,..,M) l'ensemble des formes de références. La règle de décision bayésienne minimise la probabilité de mauvaise classification, à l'aide des probabilités a priori $p(F_i)$ et des probabilités conditionnelles $p(x/F_i)$ des formes x à classer. A chaque point x est associée la forme la plus probable F_i. Ceci peut s'exprimer par la règle suivante :

$$\mathbf{classe(x) = classe(F_{i_{opt}}) \text{ avec } i_{opt} = \underset{i}{Argmax}\ p(F_i/x).} \qquad \mathbf{(4.01)}$$

Le théorème de Bayes permet d'écrire :

$$\mathbf{p(F_i/x) = \frac{p(F_i)\ p(x/F_i)}{\sum_{j=1}^{M} p(F_j)\ p(x/F_j)},} \qquad \mathbf{(4.02)}$$

avec M le nombre de classes à reconnaître.

Si le calcul des $p(F_i)$ est simple à mettre en oeuvre (e.g., par simple numérotation). l'estimation des $p(x/F_i)$ est souvent impossible à faire directement .
Pour nos expériences deux classificateurs ont été utilisés, le perceptron multi-couches et la technique des k plus proches voisins (Kppv).

4.1.2. La règle du perceptron multi-couches

Le perceptron est un outil de classification et d'approximation de fonctions non-linéaires datant des années cinquante. Une version multi-couches a été développé récemment [Le Cun 87]. Une description complète en sera faite dans le chapitre 4. Le perceptron multi-couches est une extension de la séparation par hyper-plans. Les résultats

sont encore préliminaires [Bourlard et Wellekens 89][Baum 90] : ils ne permettent pas toujours une augmentation sensible des taux de reconnaissance par rapport aux techniques classiques. Certains obstacles théoriques sur le mode d'apprentissage, un algorithme itératif, semblent en limiter la portée : cet algorithme ne garantit pas un apprentissage optimal. Une expérience sur la classification des cibles spectrales [Bennani 88] a été faite dans notre laboratoire. Elle n'a pas donné de résultats sensiblement différents de la technique des K plus proches voisins que nous allons exposer maintenant. Cette méthode est très efficace et permet une bonne interprétation des résultats.

4.1.3. La règle des K plus proches voisins

La liste des étiquettes candidates est celle dont le représentant spectral est parmi les k plus proches, au sens de la distance spectrale utilisée, de l'événement à identifier. La classification finale consiste à effectuer un vote majoritaire sur l'ensemble de ces étiquettes. Quand l'ensemble d'apprentissage est suffisamment grand, cette technique tend vers l'estimation Bayésienne [Cover et Hart 67][Devijver et Kittler 82]. Son inconvénient est l'important volume de calcul qu'elle entraîne, par rapport aux estimations paramétriques (e.g., l'estimateur gaussien) ou la séparation par hyper-plans [Gaillat 81]. Quelques techniques permettent de réduire le nombre de calculs de distances. Il s'agit des méthodes de "condensations" [Hart 68], de "tesselations" [Delannoy 80] ou d'ordonnancement [Frieman et al. 75]. Récemment Kohonen [Kohonen 88] a introduit la technique de "LVQ" (Learning Vector Quantization) qui est une méthode de "condensation améliorée". Les résultats préliminaires en reconnaissance de la parole semblent très prometteurs [Mc Dermott et Katagiri 89][Poirier 90].

Dans [Niles et al. 89][Dubuisson 90], il est conseillé de prendre k égal à $\sqrt{N/2}$, où N est le nombre moyen d'éléments par classe d'apprentissage.

C'est cette méthode simple que nous utiliserons par la suite dans notre processus de décodage acoustico-phonétique. Nous avons choisi k=3 pour N = 15 à 30 représentants par classe.

4.2. CONSTITUTION DU DICTIONNAIRE DE RÉFÉRENCES

La première étape est de constituer un dictionnaire de cibles de référence en effectuant une décomposition temporelle d'un corpus d'apprentissage. Les cibles obtenues sont étiquetées par un code phonétique associé à la réalisation d'un pseudo-

phonème. La procédure d'étiquetage est semi-automatique, elle comprend une première passe automatique et une seconde passe manuelle pour les corrections éventuelles. Une analyse de l'autocohérence[1] du dictionnaire ainsi constitué permet de repérer les confusions les plus fréquentes.

4.2.1. Corpus d'apprentissage

L'apprentissage est conduit sur un ensemble de 269 logatomes [cf. Annexe V] porteurs de tous les diphones possibles dans une suite de lettres prononcées continûment. Outre les contextes (non utilisés pour l'apprentissage), ce corpus contient 513 sons et a été prononcé une fois par chacun des deux locuteurs (l'un féminin MJC, l'autre masculin FB).

Chaque son est représenté 15 ou 30 fois (environ), selon qu'il apparaît seulement en début ou en fin de certaines lettres, ou bien dans les deux positions. Les cibles de transitions sont conservées et sont au nombre d'une centaine pour ce corpus.

4.2.2. L'étiquetage du dictionnaire de références

L'étiquetage automatique consiste en l'utilisation de notre système de décodage acoustico-phonétique sur le corpus de référence. C'est un processus itératif, le dictionnaire de références est recréé à chaque itération avec l'ensemble des cibles bien décodées. Le corpus initial est celui d'un autre locuteur déjà étiqueté. Cet étiquetage n'est cependant pas satisfaisant, une correction manuelle est nécessaire :

Un système d'étiquetage manuel a été développé sur un ordinateur PC. L'interface écran graphique haute résolution et la souris permettent une vérification de l'étiquetage automatique. Les informations affichées sur le moniteur (chronogramme du signal de parole, fonctions d'interpolation, K plus proches voisins (Kppv) des cibles (cf. Figure 4.2)) donnent toutes les informations nécessaires à une correction des erreurs d'étiquetage.

---

[1] L'autocohérence est définie comme le taux de reconnaissance sur l'ensemble d'apprentissage en enlevant à chaque reconnaissance la forme à reconnaître de l'ensemble des formes de référence.

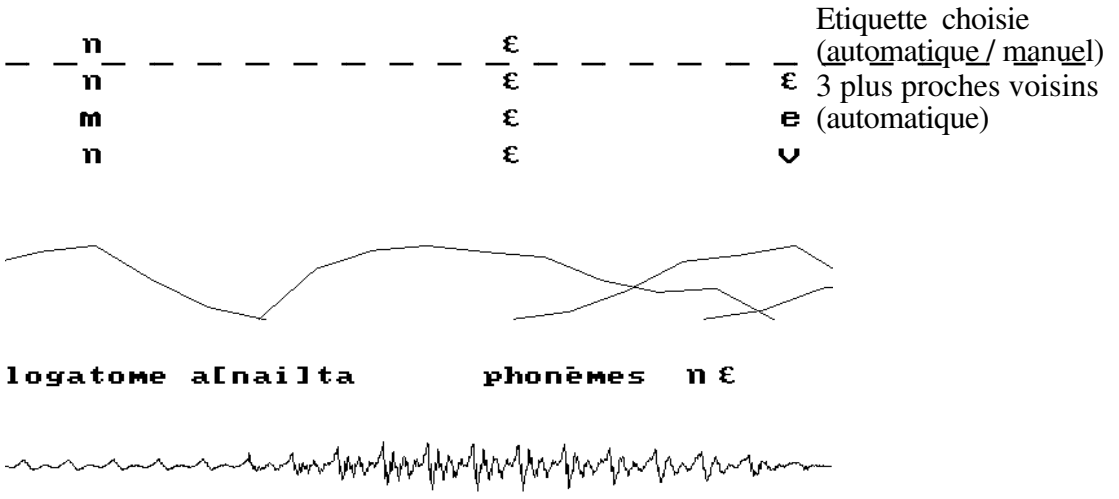


Figure 4.2. Illustration de l'étiquetage semi-automatique

Sur la figure 4.2 la première étiquette, si elle existe, correspond à celle de la cible. Les trois étiquettes suivantes sont celles des cibles du corpus d'apprentissage les plus proches, au sens de la distance spectrale utilisée, de cette cible.

Les régles de l'étiquetage manuel ont été les suivantes :

— Les occlusives voisées (i.e., (b), (d)) ont été étiquetées au centre de la tenue.

— Les occlusives sourdes (i.e., (p), (t), (k)) ont été étiquetées au début de l'explosion

— Tous les autres sons ont été étiquetés au centre de la fonction d'interpolation.

— Analyse du dictionnaire

Le dictionnaire de cibles a été analysé par la méthode des Kppv pour vérifier l'autocohérence de l'étiquetage. Les cibles dont les Kppv appartiennent à une classe trop différente de leurs représentants sont éliminées du dictionnaire. Les confusions inter macro-classes sont comptabilisées. La table des confusions les plus fréquentes initialise l'algorithme de Wagner et Fisher [Wagner et Fisher 74][Miclet 84][Annexe VII]. Cela permet un alignement lexical plus robuste, entre la succession des événements décodés et celle des sons attendus. Un exemple de table des confusions est donné ci-dessous :

Type de paramètres	Cibles spectrales		Vecteur spectral [1]	
Autocohérence[2]	84.4%		79.4%	
	Confusions	Pourcentage	Confusions	Pourcentage
	(m) <-> (n)	27%	(m) <-> (n)	35%
	(p) <-> (t)	27%	(l)<-> (e)	24%
	(x) <-> (z)	15%	(x) <-> (z)	17%
	(l)<-> (e)	13%	(p) <-> (t)	13%
	(b) <-> (d)	11%	(é) <-> (è)	10%

Tableau 4.1. Comparaison de l'autocohérence et des confusions entre les cibles spectrales et les vecteurs spectraux (WMFCC locuteur FB)

La plupart des confusions apparues entre les sons étiquetés de notre dictionnaire de cibles spectrales s'expliquent assez facilement phonétiquement. Les sons confondus ne diffèrent que d'un seul trait phonétique ou appartiennent à une même macro-classe.

Sur ce même tableau on peut remarquer que l'autocohérence du dictionnaire de cibles spectrales est plus élevée que celle du dictionnaire des vecteurs spectraux originaux. Cela confirme l'hypothèse selon laquelle les cibles spectrales devraient se prêter plus facilement au processus de décodage acoustico-phonétique que les vecteurs spectraux originaux.

---

[1] Le vecteur spectral choisi est celui correspondant au centre de gravité de la fonction d'interpolation associée à la cible spectrale.

[2] L'autocohérence est définie comme le taux de reconnaissance sur l'ensemble d'apprentissage en enlevant à chaque reconnaissance la forme à reconnaître de l'ensemble des formes de référence.

Lors de la décomposition temporelle du corpus d'apprentissage, toutes les cibles spectrales ne peuvent être étiquetées. Ces cibles sont dues généralement à des événements contextuels (e.g., transitions). Nous les avons appelées les "cibles de transition".

4.2.3. Cibles de transition

Il est difficile de trouver une interprétation phonétique aux cibles de transitions [Bimbot 88]. Elles apparaissent au niveau des transitions entre certains sons. Les cibles de transition ont un contenu spectral beaucoup moins prédictible que les cibles associées à des sons. Leur présence n'est pas systématique et leurs caractéristiques spectrales apparaissent très variables. Ces phénomènes ont pour conséquence l'existence d'un nombre important d'insertions lors du décodage direct des événements acoustiques détectés par la décomposition temporelle. Il est donc nécessaire de traiter les cibles de transition, c'est-à-dire de savoir les identifier en tant que telles dans la succession des événements. Deux modes de sélection sont utilisés pour cela :

— Pour le premier mode de sélection, l'hypothèse est posée qu'il existe une certaine invariance des cibles de transitions. Les cibles de transition du corpus d'apprentissage sont étiquetées par un symbole particulier et permettent de reconnaître une partie des cibles de transition du corpus test. La conservation des cibles de transition dans le dictionnaire permet environ 2/3 des identifications des cibles de transition dans le corpus test. Cependant le score de reconnaissance diminue sensiblement (environ 5%), certains pseudo-phonèmes étant identifiés comme des cibles de transitions. Ces résultats tendent à prouver qu'une majorité des cibles de transition a des caractéristiques propres (bien que très variables). Pour identifier le 1/3 restant des cibles de transition, un second mode de sélection est utilisé.

— Le second mode de sélection se sert des informations complémentaires définies dans le prochain paragraphe.

4.3. INFORMATIONS COMPLÉMENTAIRES SUR LES CIBLES.

L'adjonction d'informations complémentaires sur les cibles de référence et les cibles de test permet de mieux discriminer les événements phonétiques des événements contextuels (transitions) lors de la reconnaissance. Les informations complémentaires sont représentées par :

— Un paramètre d'énergie,

— Le nombre de passages par zéro (ppz),

— Un paramètre d'instabilité spectrale,

— Un paramètre de durée.

Chaque cible spectrale à reconnaître est donc associée à ces quatre paramètres. De même, chaque symbole phonétique (et non chaque cible) du dictionnaire est muni de caractéristiques complémentaires telles que la valeur moyenne et l'écart-type des paramètres complémentaires.

Au moment de la classification, on s'assure que les étiquettes phonétiques produites ne présentent pas d'incohérence avec les paramètres complémentaires. Dans le cas contraire, on substitue une étiquette de transition à l'étiquette d'origine.

L'intégration directe des informations complémentaires dans l'algorithme de décomposition temporelle ne donne pas de résultats satisfaisants avec les métriques choisies. Un phénomène de sur-segmentation apparaît car l'évolution des paramètres complémentaires n'est pas synchrone avec celle des paramètres spectraux.

4.3.1. L'énergie, le nombre de ppz, l'instabilité spectrale, la durée

— L'énergie du signal est un élément important pour la discrimination silence-parole. Il permet aussi la discrimination de nombreux sons. Deux types de paramétrisations de l'énergie, par une fenêtre de longueur Tfen, ont été utilisés :

$$E_1 = \sum_{t=0}^{Tfen} x(t)^2 \qquad (4.03)$$

$$E_2 = 10 \log_{10} \left(\sum_{t=0}^{Tfen} x(t)^2 \right) \qquad (4.04)$$

Une normalisation intervient ensuite pour fixer l'énergie moyenne sur le corpus d'apprentissage à 1 (ou 0 db). Cette normalisation intervient aussi sur l'ensemble de test pour obtenir une énergie moyenne constante. Cette procédure est indispensable car la diction des logatomes du corpus d'apprentissage n'est pas naturelle et a nécessité un effort plus important de nos locuteurs : ceux-ci ont tendance à parler moins fort. La paramétrisation en décibels (E_2) permet une meilleure discrimination entre sons : sa répartition par classe suit mieux la loi de probabilité employée (loi gaussienne).

— Le calcul des passages par zéro (ppz) est une méthode d'analyse opérant dans le domaine temporel. Elle est fondée sur la mesure de la densité de passages par zéro du signal vocal à moyenne nulle. Il suffit de compter le nombre de changements de signe du signal pendant un intervalle de temps.

Cette méthode très grossière permet par exemple de décider du caractère voisé ou non-voisé d'un segment de parole [Meloni 82]. Au cours de nos expériences, cette technique a surtout permis de distinguer les fricatives du reste des sons. Son utilisation est cependant restée marginale car les erreurs de substitutions sont rares pour les fricatives.

— L'instabilité spectrale est calculée à partir d'une distance interspectrale entre deux fenêtres consécutives. Le vecteur O représente le centroïde généralisé des cinq fenêtres précédentes. L'instabilité spectrale est donnée par la formule suivante :

$$Dn = \frac{D(y_n, y_{n+1})}{\sqrt{D(y_n, O)\ D(O, y_{n+1})}}, \qquad (4.05)$$

avec y_n vecteur spectral centré de la fenêtre n.

Les minima relatifs de cette fonction correspondent aux parties stables des sons, tandis que les maxima relatifs correspondent aux instants des transitions.

— La durée ou support temporel de la fonction d'interpolation ϕ_i discrimine surtout les sons stables comme les voyelles, les occlusives voisées, les consonnes nasales. Elle est moins utile sur les plosives où la fonction d'interpolation obtenue est très instable.

4.3.2. La pondération des paramètres

Les paramètres complémentaires p(t) sont calculés pour chaque trame de signal (typiquement 10ms). La pondération des paramètres par la fonction d'interpolation $\phi_i(n)$, ($n = T_1, .., T_2$), avec $[T_1, T_2]$ support de la fonction d'interpolation ϕ_i, permet de garder un seul représentant $P(\phi_i)$ par cible g_i étiquetée. Le calcul se fait par la méthode suivante :

$$P(\phi_i) = \frac{\sum_{n=T_1}^{T_2} \phi_i(n)\ p(n)}{\sum_{n=T_1}^{T_2} \phi_i(n)} \qquad (4.06)$$

Ces nouveaux paramètres sont plus discriminants pour le décodage acoustico-phonétique. L'importance des zones de coarticulation (frontière des fonctions

d'interpolation) est minimisée. Un autre choix consiste à prendre les paramètres du centre de gravité de la fonction d'interpolation (cette solution ne s'applique que pour des sons très stables et peut donner de plus mauvais résultats pour les autres sons).

4.3.3. La discrimination des caractéristiques complémentaires

Une statistique, par son, des paramètres compémentaires pondérés par les fonctions d'interpolation, est calculée. Un vecteur moyenne $Moy_{Par}(Ph)$ et un vecteur écart-type $Ect_{par}(Ph)$ du paramètre complémentaire considéré sont calculés pour chaque son Ph. La vraisemblance de l'appartenance du son x à la classe y est calculée par la formule suivante :

$$\mathbf{Vrais(y/x)} = \frac{\mathbf{1}}{\sum_{\mathbf{Par}} \mathbf{G_{Par}(y/x)}} \qquad \mathbf{(4.07)}$$

avec

$$\mathbf{G_{Par}(y/x)} = \frac{\mathbf{| Par(x) - Moy_{Par}(y) |}}{\mathbf{Ect_{Par}(y)}} \qquad \mathbf{(4.08)}$$

La loi de probabilité choisie est une gaussienne simplifiée par son, la matrice de covariance est considérée comme diagonale. On considère que le son x n'appartient pas à la classe y si Vrais(y/x) est inférieur à 1/2.

Dans le tableau 4.2 se trouvent les informations supplémentaires pour un locuteur masculin (FB). On retrouve les conclusions suivantes, souvent utilisées pour classifier les sons :

— La voyelle (a) possède la plus forte énergie moyenne, la tenue de la consonne (d) la plus faible énergie moyenne. L'énergie moyenne de l'explosion de la consonne (p) est extrèmement fluctuante. L'énergie discrimine parfaitement (z) de (c).

— La consonne (s) possède le plus grand nombre de ppz, la tenue de la consonne (b) le plus petit nombre de ppz. Le nombre de ppz discrimine parfaitement (x) de (v).

— La voyelle (é) est la plus stable, l'explosion de la consonne (p) la moins stable. On peut cependant remarquer que l'instabilité est le paramètre complémentaire le moins discriminant.

Sons	Nombre	Durée (cs)		Energie (db)		ppz		Instabilité	
		Moy	Ect	Moy	Ect	Moy	Ect	Moy	Ect
(a)	29	6.5	1.6	6.0	3.3	13.8	4.3	10.0%	2.6%
(é)	18	6.9	2.0	1.1	2.8	14.9	3.8	9.5%	2.0%
(è)	19	5.7	1.2	2.3	2.2	14.5	4.7	10.2%	1.8%
(i)	28	6.1	2.0	-9.0	4.3	31.4	19.2	10.7%	2.1%
(o)	29	6.0	2.3	1.7	3.0	8.2	2.8	11.7%	2.8%
(e)	25	7.0	2.7	1.6	3.1	11.0	5.7	9.7%	2.8%
(y)	27	6.6	2.3	-6.0	3.4	13.4	6.3	10.8%	2.4%
(p)[1]	11	4.0	1.4	-13.9	13.1	5.7	3.4	13.2%	3.3%
(t)[1]	10	4.6	2.7	-14.3	8.2	19.5	16.1	12.7%	2.9%
(k)[1]	24	3.5	1.0	-9.4	5.3	42.1	18.2	18.4%	4.6%
(b)[2]	14	5.6	1.1	-10.0	5.3	5.0	0.6	10.8%	1.7%
(d)[2]	29	4.5	1.3	-15.4	4.0	5.6	2.2	11.8%	2.3%
(f)	10	5.3	1.0	-12.5	2.9	50.6	9.8	9.7%	2.4%
(s)	27	6.6	2.0	-5.1	2.6	108.6	15.7	11.1%	1.5%
(v)	13	4.6	1.1	-4.1	4.4	16.1	3.4	9.6%	1.9%
(x)	16	6.6	2.5	0.3	2.7	68.8	11.4	10.6%	1.5%
(z)	16	5.9	1.3	-7.8	1.7	74.8	13.0	10.0%	2.1%
(c)	11	6.0	1.9	-0.3	1.9	49.4	13.7	10.9%	1.7%
(m)	15	6.2	1.0	-8.9	2.5	7.5	2.7	10.6%	1.5%
(n)	12	5.0	1.7	-8.0	3.2	7.5	2.3	11.7%	1.4%
(r)	10	4.3	0.6	-7.5	5.1	33.7	16.0	12.2%	2.4%
(l)	15	5.6	2.1	1.1	2.7	10.6	3.7	12.2%	2.1%

Tableau 4.2. Informations complémentaires (WMFCC locuteur FB)

On peut remarquer sur le tableau 4.3. que si chacun des paramètres complémentaires ne suffit pas individuellement pour lever toutes les ambiguités comme (b) / (d), (m) / (n), (t) / (k), la fonction de vraisemblance construite grâce à l'ensemble des paramètres complémentaires le permet.

---

[1] (x)1, il s'agit de l'explosion du son (x).

[2] (x)2, il s'agit dela tenue du son (x).

Vrais((b) / ($\bar{d}$))[1] = 0.27
Vrais((d) / ($\bar{b}$))[1] = 0.34
Vrais((t) / ($\bar{k}$))[1] = 0.23
Vrais((k) / ($\bar{t}$))[1] = 0.22
Vrais((m) / ($\bar{n}$))[1] = 0.49
Vrais((n) / ($\bar{m}$))[1] = 0.50

Tableau 4.3. Exemples de la discrimination obtenue par l'ensemble des paramètres complémentaires

CONCLUSIONS

Le processus de décodage acoustico-phonétique comprend deux étapes :

— L'étiquetage en pseudo-phonèmes des cibles spectrales de la décomposition temporelle par la technique des K plus proches voisins.

— La sélection des cibles spectrales dites de "transition", grâce à l'utilisation de paramètres complémentaires.

Le besoin d'un processus de décodage acoustico-phonétique performant nous a amenés à constituer un dictionnaire de cibles de référence pour chaque locuteur, ce qui est une lourde contrainte car le corpus d'apprentissage est important (269 logatomes). Pour permettre une reconnaissance rapide de locuteurs différents de ceux de l'apprentissage, une adaptation des cibles spectrales du dictionnaire est une alternative nécessaire.

---

[1] ($\bar{x}$) représente la valeur moyenne des paramètres complémentaires du son (x).

CHAPITRE 5. ADAPTATION AU LOCUTEUR PAR TRANSFORMATIONS NON-LINÉAIRES DES CIBLES SPECTRALES

La variabilité inter-locuteur est une source majeure d'erreurs en reconnaissance de la parole pluri-locuteurs. Les paramètres spectraux correspondant à un même son produit par deux locuteurs diffèrent le plus souvent. On observe d'ailleurs, pour un changement de locuteur, une détérioration notable des performances de notre système de décodage acoustico-phonétique.

Une approche pour contrôler cette variabilité consiste à apprendre une transformation spectrale sur un vocabulaire restreint prononcé par les deux locuteurs, puis à utiliser cette transformation pour adapter les vecteurs du dictionnaire de références. Une hypothèse est formulée qu'un tel "espace acoustique" existe pour chaque locuteur. Pour passer de l'espace d'un locuteur à l'espace d'un autre locuteur plusieurs choix sont possibles : si l'espace est discret et fini [Gray 84], la transformation est discréte sinon une fonction continue est généralement employée, le but est d'approximer au mieux cette fonction avec le nombre minimum de paramètres. La première technique a été étudiée dans [Goatcher et Mason 1986][Shikano et al. 86][Bonneau et Gauvain 87], elle nécessite un vocabulaire d'adaptation important. Nous avons préféré la seconde méthode introduite par Y. GRENIER [Grenier 77] qui demande un vocabulaire d'adaptation plus restreint.

Après un bref état de l'art de l'estimation d'une fonction à partir d'un ensemble de données d'entrée et de données de sortie, nous proposons la technique d'estimation d'une fonction non-linéaire, basée sur les perceptrons multi-couches. Nous exposons ensuite les problèmes rencontrés lors de la constitution de l'ensemble d'apprentissage (ensemble de données d'entrée et de données de sortie) : les limitations des techniques d'alignement de signaux de parole (la comparaison dynamique) et le choix du vocabulaire d'adaptation. Pour chacun de ces problèmes nous décrivons les réponses que nous y avons apportées : l'alignement dynamique itératif et une procédure automatique pour le choix du vocabulaire d'adaptation.

5.1. ESTIMATION D'UNE FONCTION DE TRANSFORMATION NON-LINÉAIRE

La formulation générale de la transformation qui fait correspondre un même son Ω prononcé par deux locuteurs différents A et B est la suivante :

$$\mathcal{F}(\Omega_A) \approx \mathcal{G}(\Omega_B) \text{ avec } \mathcal{F} \text{ et } \mathcal{G} \text{ deux fonctions quelconques.} \qquad (5.01)$$

Les régressions multiples et les corrélations canoniques permettent le calcul de $\mathcal{F}$ et $\mathcal{G}$ dans le cas où elles sont linéaires. Elles ont été étudiées dans [Grenier 77][Choukri 87] et ont permis une bonne adaptation dans le cas de la reconnaissance de mots isolés. Néanmoins leurs performances limitées nous incitent à nous orienter vers des techniques non-linéaires [Duvaut 89][Class et al. 90].

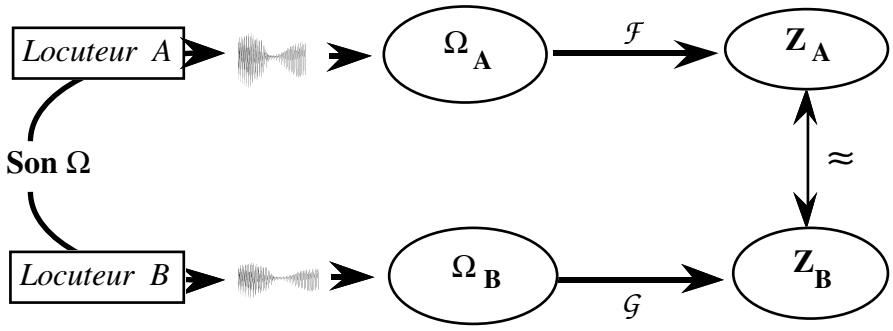


Figure 5.1. Synoptique de la transformation

Plusieurs techniques, désormais classiques, permettent l'estimation de fonctions non-linéaires. Généralement il s'agit d'une décomposition sur une base de "bonnes fonctions" (i.e., continûment dérivables), souvent des polynômes. Nous avons choisi une technique basée sur les perceptrons multi-couches dont la fonction de base est une sigmoïde (e.g., $f(x) = \frac{1-e^{-x}}{1+e^{-x}}$), fonction bornée et continûment dérivable. Ce choix semble bien s'adapter à la transformation d'espaces acoustiques eux-mêmes bornés. Dans un but d'évaluation nous avons développé et testé une technique de "filtrage" polynomial restreint à des polynômes de degré faible (≤ 3). A l'ordre 1 nous retrouvons les régressions linéaires multiples, les non-linéarités apparaissent à l'ordre 2 avec le filtrage quadratique. Nous n'avons étudié que les cas où la fonction $\mathcal{F}$ de la formule 5.01 est la fonction identité.

5.1.1. La régression linéaire multiple

C'est l'une des premières techniques d'adaptation utilisée [Furui 80]. Soient les vecteurs x et y d'ordre p. Ils représentent deux sons appariés pour deux locuteurs. On considère que le $i^{ème}$ coefficient de y du locuteur de test dépend linéairement des p coefficients de x :

$$y^i = \sum_{j=1}^{p} RM_{i,j}\, x^j, \qquad (5.02)$$

avec la matrice pxp RM, appelée matrice de régression multiple.

Les techniques de régression multiple permettent d'estimer la matrice de transformation RM si l'on dispose de suffisamment de couples de vecteurs appariés $\{(x_n, y_n)\}$ (n = 1 ,.., N) :

$$RM = Y\, X^{\#} \qquad (5.03)$$

avec $X^{\#}$ pseudo-inverse de X,

où Y est une matrice pxN dont l'élément $Y_{j,n}$ est égal à y_n^j, X est une matrice pxN dont l'élément $X_{j,n}$ est égal à x_n^j.

Les extensions de cette méthode sont :

— Les corrélations canoniques [Grenier 77][Choukri 87].

— La prise en compte des moments d'ordre supérieur à 2 par des techniques de filtrage polynomial ou de perceptrons multi-couches.

5.1.2. Le filtrage polynomial

Le filtrage polynomial consiste en une transformation linéaire d'une forme polynomiale des données d'entrée. Son calcul se fait par des techniques classiques d'algèbre linéaire. Soient N couples de vecteurs appariés $\{(x_n, y_n)\}$ (n = 1 ,.., N) et X et Y les matrices associées.

La forme polynomiale la plus simple est l'utilisation d'un forme polynomiale d'ordre 1. C'est la résolution du système linéaire Y=RM X. Nous retrouvons la technique des régressions multiples du paragraphe précédent. Le nombre de paramètres à estimer est p^2.

L'utilisation des formes polynomiales d'ordre 2 (filtrage quadratique) permet de modéliser des phénomènes non-linéaires. On considère que le $i^{ème}$ coefficient de y_n du locuteur de test ne dépend pas seulement des p coefficients de x_n, mais aussi des p^2 produits $x_n^j\, x_n^i$. Cela donne :

$$y^i = \sum_{j=1}^{p} FQ_{i,j}\, x^j + \sum_{j,k=1}^{p} FQ_{i,(i+1)p+k}\, x^j\, x^k), \qquad (5.04)$$

avec la matrice $(p+p^2)$xp FQ, appelée matrice de filtrage quadratique.

Les techniques de régression multiple permettent d'estimer la matrice de transformation FQ si l'on dispose de suffisamment de couples de vecteurs appariés $\{(x_n, y_n)\}$ (n = 1,..,N):

$$\mathbf{FQ = Y\ X'^{\#}} \qquad \mathbf{(5.05)}$$

avec $X'^{\#}$ pseudo-inverse de X',

où Y est une matrice pxN dont l'élément $Y_{j,n}$ est égal à y_n^j, X' est une matrice $(p+p^2)$xN dont l'élément $X_{j,n}$ est égal à x_n^j pour $1 \leq j \leq p$ et à $x_n^k\ x_n^l$, avec j = p+kp+l, pour j > p.
En utilisant la symétrie du produit, le nombre de paramètres à estimer est $p^2(p+3)/2$.

En augmentant le degré des polynômes employés, on peut définir le filtrage cubique, le filtrage d'ordre 4, etc... Le nombre de paramètres à estimer croît exponentiellement : nous nous sommes limités au filtrage quadratique. Nous avons essayé pour diminuer ce nombre de paramètres de réduire le modèle.

— Réduction de modèle

Dans le cas où le nombre de paramètres à estimer reste trop élevé, on peut effectuer une réduction de modèle en utilisant une décomposition en valeurs singulières des matrices d'entrée et de sortie. Seules les composantes principales correspondant à une inertie significative sont considérées.

N'ayant à notre disposition que quelques centaines de couples, nous avons utilisé cette réduction de modèle pour le filtrage quadratique.

5.1.3. Le perceptron multi-couches

Le perceptron multi-couches est un outil de traitement de l'information issu des recherches en connexionisme [Lippman 87]. Cet axe de recherche a produit depuis de nombreuses années des méthodes puissantes de Reconnaissance des Formes. Elles sont fondées sur l'utilisation de cellules (ou noeuds) dont le contenu (e.g., une valeur réelle) évolue dans le temps par interactions réciproques entre les cellules. Ces techniques font massivement appel à une architecture parallèle.

Le perceptron multi-couches est un réseau connexionniste directionnel où les unités d'entrée et les unités de sortie sont différenciées (figure 5.1). Il comporte des unités cachées (i.e., non reliées directement à une sortie ou à une entrée) qui permettent la modélisation de transformations complexes [Lapedes et Farber 87] :

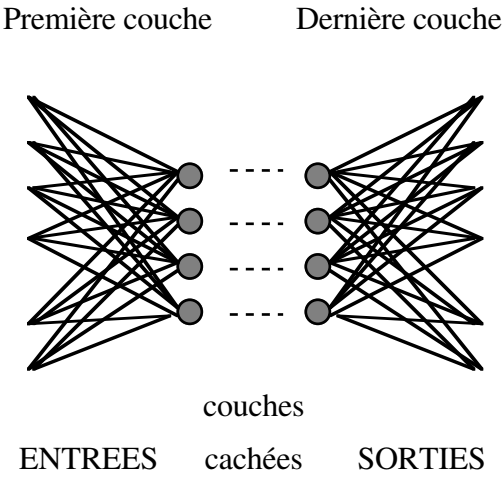


Figure 5.2. Perceptron multi-couches

La sortie de chaque noeud est fonction des sorties des noeuds de la couche précédente :

$$\mathbf{Sortie(noeud_j) = } f \Big(\sum_{\mathbf{i=0}}^{\mathbf{i=N}} \mathbf{w_{ij} * Entrée_i\ (noeud_j)} \Big), \qquad \mathbf{(5.06)}$$

où f est une fonction continûment dérivable (généralement une sigmoïde).

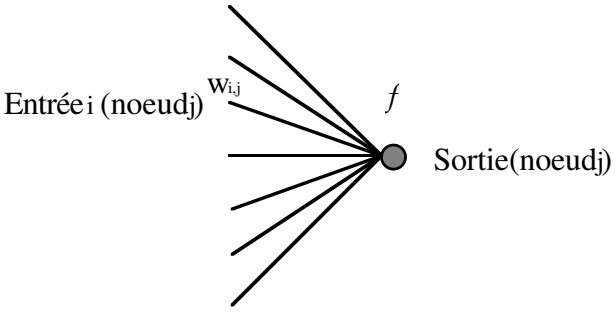


Figure 5.3. Noeud du perceptron multi-couches

La structure du perceptron étant fixée, un algorithme d'apprentissage itératif, la rétropropagation du gradient, estime les poids w_{ij}. L'algorithme minimise un critère d'erreur entre la sortie désirée et la sortie calculée. Quelques principes nous ont guidés dans la définition de la structure d'un perceptron d'adaptation :

— L'ajout d'une couche cachée ne peut qu'améliorer les performances. En effet, dans le pire des cas les noeuds supplémentaires fonctionnent en mode linéaire et ne modifient pas la fonction de transformation.

— Des théorèmes récents [Funahashi 89][Hornik 89] ont montré que l'ensemble des fonctions modélisées par un perceptron multi-couches avec une seule couche cachée était dense dans l'espace des fonctions. Malheureusement, ce théorème ne permet pas de spécifier le nombre de noeuds de la couche cachée nécessaires pour atteindre une précision donnée.

Considérant le faible nombre d'exemples (i.e., quelques centaines) de couples de vecteurs appariés $\{(x_n, y_n)\}$ à notre disposition, nous avons choisi, pour l'architecture du réseau, un nombre restreint de noeuds et une seule couche cachée. Dans le cas contraire nous risquions de faire un sur-apprentissage au lieu de calculer une fonction d'adaptation généralisable à d'autres sons que ceux de référence.

L'inconvénient de cette méthode est son important coût de calcul et la présence de nombreux minima locaux. En dépit d'une convergence incertaine vers le minimum global, cette technique a donné de meilleurs scores [Montacié et al. 89] que la technique des corrélations canoniques présentée comme la plus performante des techniques linéaires jusqu'à présent.

L'apprentissage d'un perceptron multi-couches demande un temps de calcul important. L'initialisation linéaire des poids permet de réduire considérablement le temps d'apprentissage avec des performances équivalentes.

5.1.3.1. La rétropropagation du gradient

Y. Le Cun [Le Cun 85] a développé cette méthode pour l'apprentissage des perceptrons multi-couches. C'est une extension de la règle de Widrow et Hopf [Widrow et Hopf 60] à une fonction de coût non-linéaire. Cette régle dite du "gradient stochastique" affirme, sous certaines hypothèses, que les deux procédures suivantes sont équivalentes:

— Une seule itération d'apprentissage pour l'ensemble des formes à apprendre.

— Une itération d'apprentissage pour chacune des formes à apprendre.

L'algorithme de rétroprogation du gradient permet de calculer les dérivées partielles de l'erreur de sortie par rapport à tous les poids w_{ij} du réseau. Une procédure de gradient est appliquée pour minimiser en fonction de tous les poids un critère d'erreur. Ce critère d'erreur est généralement la somme des écarts quadratiques entre les sorties désirées et les sorties calculées. Le calcul des dérivées partielles est effectué en sens inverse de la propagation, le gradient est donc rétropropagé. L'algorithme complet est le suivant :

Etape 0 : Initialisation des poids w_{ij} à une valeur non-nulle.

Etape 1: Un couple de vecteurs (i.e., un vecteur d'entrée x_n et un vecteur de sortie désiré d_n) est présenté au réseau. L'ordre des couples présentés est généralement aléatoire, cela évite des phénomènes d'oscillations.

Etape 2: Les $\{y^i\}$ ($i = 1 ,.., N_{noeud}$) et les $\{a^i\}$ ($n = 1 ,.., N_{noeud}$) correspondant à la réponse du réseau au vecteur d'entrée sont calculés.

$$\mathbf{a^i = \sum_{\{j\}} w_{ij}\ x_n^i,\ y^i = } f(\mathbf{a^i}) \qquad \mathbf{(5.07)}$$

où {j} représente l'ensemble des noeuds de la couche suivante qui sont reliés au noeud i.

Etape 3: L'erreur quadratique de reconstruction E, entre le vecteur y et le vecteur d_n correspondant à la réponse du réseau au vecteur d'entrée x_n, est calculée. Les gradients $\delta_i = \partial E/\partial a^i$ sont estimés par rétropropagation à l'aide des formules suivantes:

Pour les cellules de sortie

$$\delta_i = 2\, f'(a^i)(d_n^i - y^i) \qquad (5.08)$$

Pour les autres cellules

$$\delta_i = f'(a^i) \sum_{\{k\}} \delta_k\ w_{ik}, \qquad (5.09)$$

où {k} représente l'ensemble des noeuds des couches précédentes qui sont reliés au noeud i.

Les poids sont réajustés par la formule suivante :

$$w_{ij}\ (n+1) = w_{ij}\ (n) + \eta\ \delta_i\, y^j \qquad (5.10)$$

w_{ij} (n) représente le poids de la connection du noeud i au noeud j à la $n^{ième}$ itération.

η est une valeur de gain comprise entre 0 et 1.

y^j est la valeur de sortie du noeud j ou une valeur d'entrée.

$$\text{Avec}\ \ y^i = f(a^i) = \frac{1-e^{-a^i}}{1+e^{-a^i}} \qquad (5.11)$$

$$\text{On a } f'(a^i) = \frac{2\, e^{-a_i}}{(1+e^{-a_i})^2} \qquad (5.12)$$

$$\text{et } f'(a^i) = (1 + f(a^i))\,(1 - f(a^i))\,/\,2 \qquad (5.13)$$

On peut remarquer que le calcul de $f'(a^i)$ ne demande pas de calculs supplémentaires d'exponentielles et ne met en jeu que des multiplications et des additions.

Etape 4: Tant que les conditions d'arrêt ne sont pas vérifiées, aller à l'étape 1. Les conditions d'arrêt sont multiples (e.g., nombre maximum d'itérations, erreur de sortie minimale, etc.).

5.1.3.2. Préparation des données d'apprentissage

La sortie d'un perceptron multi-couches est bornée (e.g., entre -1 et +1). Les données d'apprentissage de la sortie n'ont pas de dynamique déterminée. Une préparation des données d'apprentissage est alors nécessaire. Un choix possible est de remplacer une valeur désirée du noeud j, d_n^j par $d_n^j \cdot \frac{|\, d_n^j \,|}{|\, 1+d_n^j \,|}$. Cependant, cette transformation ne respecte pas la topologie de l'espace originel. La transformation que nous avons choisie consiste à centrer et à réduire les valeurs d'apprentissage d'un noeud de sortie j :

$$d_n^j \rightarrow \frac{d_n^j - Moy_j}{\alpha\ Etc_j}, \qquad (5.14)$$

où Moy_j est l'estimation de la moyenne des valeurs d'apprentissage du noeud de sortie j, Etc_j représente l'estimation de l'écart type. Le choix de α doit être optimisé sachant que :

— Si α est trop petit, un grand nombre des données d'apprentissage en sortie se trouvent hors de l'intervalle (i.e., -1 à +1) couvert par le noeud j. L'erreur de reconstruction augmente. Le noeud fonctionne le plus souvent dans la zone de saturation.

— Si α est trop grand, un grand nombre des données d'apprentissage en sortie sont proches de l'origine. Le noeud j fonctionne dans la zone linéaire de la sigmoïde. L'erreur de reconstruction augmente.

Un compromis est donc nécessaire. Au cours de nos expériences, le paramètre α a été choisi égal à 2. Dans le cas d'une répartition gaussienne, 95% des exemples sont dans la zone couverte par la sigmoïde.

5.1.3.3. Initialisation linéaire des poids

Le principe est d'utiliser les coefficients de la matrice de régression multiple RM (cf. formule 5.02) pour initialiser la première couche des poids. Ceux des autres couches sont choisis pour ne pas apporter de modifications à la réponse du réseau. Pour un réseau quelconque à plusieurs couches cachées l'algorithme est le suivant :

— Soit N_1 le nombre de noeuds de la première couche et N_2 le nombre de noeuds de la deuxième couche.

  —Si $N_1 = N_2$, les poids de la première couche sont initialisés avec les coefficients de la matrice de régression multiple.

  —Si $N_1 > N_2$, les poids de la première couche sont initialisés avec les N_2 premières composantes principales.

  —Si $N_1 < N_2$, les poids excédentaires sont initialisés à zéro.

— Les couches supplémentaires ne doivent pas modifier la réponse du réseau. Les poids des couches supplémentaires sont initialisés à $1/f'(0)$ (i.e., 1/2 pour un sigmoïde entre -1 et +1) pour les liaisons entre noeuds de même indice, et mis à zéro pour les autres.

— Une faible valeur aléatoire est additionnée à chaque poids pour permettre au réseau de converger.

Cet algorithme accélère la convergence dans un rapport de 10. L'erreur de reconstruction (cf . formule 5.15) est inférieure à celle de la régression multiple (cf. figure 5.4). Cela prouve que l'apprentissage par les méthodes linéaires ne constitue pas un minimum local de la surface d'énergie de l'erreur (ce qui empêcherait le perceptron multi-couches de converger).

L'erreur de reconstruction Err est la moyenne des erreurs de reconstruction pour chacun des N_{noeud} noeuds de sortie et de chacune des N formes de sortie de l'ensemble d'apprentissage. Elle est calculée par la formule suivante :

$$\mathbf{Err} = \left(\sum_{n=1}^{N} || \mathbf{d_n} - \mathbf{y_n} ||^2 \right) / \left(\mathbf{N_{noeud}} \cdot \mathbf{N} \right) \qquad \mathbf{(5.15)}$$

Puisque les sorties sont à écart-type constant, on peut calculer un Rapport Signal à Bruit (RSB) en décibels par la formule suivante :

$$\mathbf{RSB_{dB} = -10\ log_{10}(Err\ /\ \alpha^2)} \qquad \mathbf{(5.16)}$$

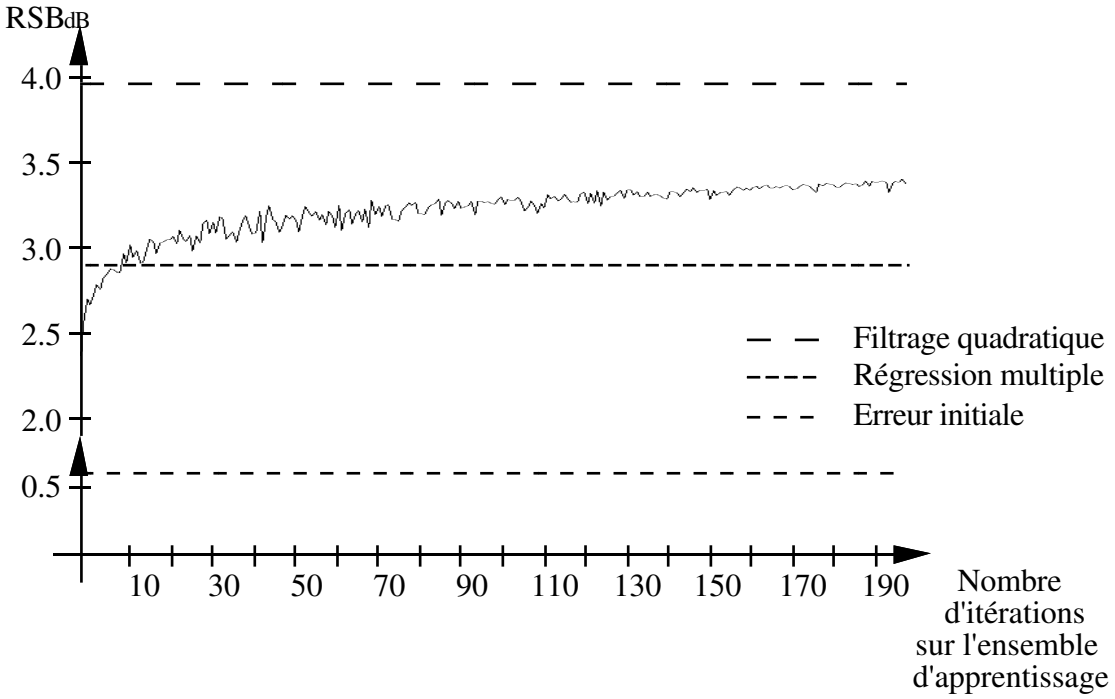


Figure 5.4. Rapport Signal à Bruit en fonction du nombre d'itérations pour une initialisation linéaire

5.1.3.4. Séparation des noeuds saturés

La présence des "plateaux" est l'un des principaux problèmes pour la convergence des perceptrons multi-couches ayant des fonctions de transfert f(x) bornées (e.g., la sigmoïde). Les "plateaux" de la surface de l'erreur sont des zones où l'erreur de sortie de certains noeuds est grande, tandis que le gradient est presque nul. Les "plateaux" sont provoqués par la présence de noeuds à fonction de transfert saturée. Lorsque les poids des connections arrivant à un noeud sont trop élevés, la somme pondérée des entrées est constamment dans la zone saturée de la fonction non-linéaire. La dérivée de la fonction de transfert devient très petite. Le gradient devient nul et se rétropropage sur les gradients des noeuds inférieurs. La convergence du réseau est au mieux ralentie et au pire bloquée.

Une solution à ce problème est de faire décroître les poids exponentiellement à chaque itération (i.e., $w_{ij}(n+1) = \alpha\, w_{ij}(n)$, avec $0 < \alpha < 1$). Cette solution présente deux inconvénients :

— Le minimum global cherché n'est plus le même.

— Si les poids deviennent trop petits, le réseau converge vers la solution stable où tous les poids sont nuls.

La solution proposée consiste à détecter les noeuds fonctionnant en zone saturée, et à remplacer chacun d'entre eux par deux ou plusieurs noeuds. Les noeuds créés devront fonctionner en mode normal (i.e., non saturé). La réponse du réseau ne devra pas trop changer. L'algorithme décrivant le dédoublement des noeuds saturés est le suivant :

— A chaque itération (i.e., passage des N exemples), le noeud le plus saturé parmi P est détecté. Dans le cas où la fonction de transfert f(x) est impaire, le noeud choisi est celui dont V_{sat} est maximal, avec

$$V_{sat} = \sum_{i=1}^{P} | \, Sortie_i \, | \; / \; N. \qquad (5.17)$$

— Le noeud détecté est remplacé par deux noeuds dont les poids d'entrée w_{ij} sont ceux du précédent noeud multipliés par β , avec

$$\beta = \frac{f^{-1}(V_{sat} / 2)}{f^{-1}(V_{sat})} \qquad (5.18)$$

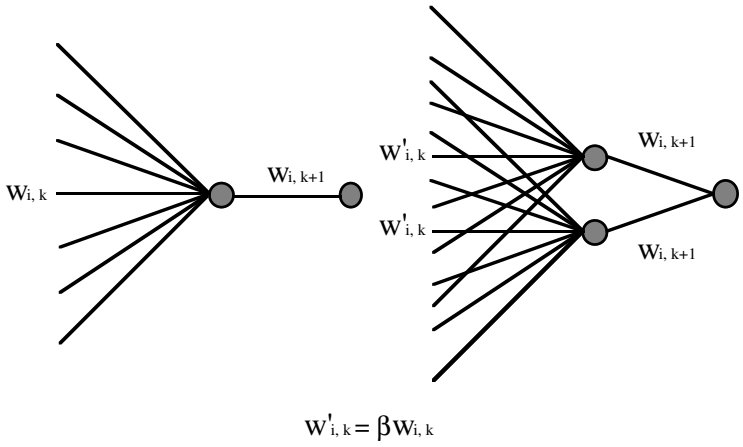


Figure 5.5. Séparation du noeud le plus saturé

Cette procédure de division des noeuds saturés ne doit être utilisée qu'un nombre limité de fois, sinon on risquerait de perdre toute généralisation de la fonction de transformation apprise.

5.1.3.5. L'implémentation de l'algorithme

Cet algorithme a été utilisé sur de nombreuses architectures matérielles. La qualité d'une implémentation se mesure au nombre moyen d'Apprentissages par Connexion et par Seconde (ACS). Chaque apprentissage comprend une propagation des données d'entrée et une rétropropagation de l'erreur de sortie. Le programme que nous avons écrit en langage C [Montacié et Barral 89] permet 80000 ACS sur une DEC 3100. Les caractéristiques de notre application, pour le calcul d'une fonction de transformation spectrale, sont les suivantes :

— Le réseau comprend une couche cachée. Chaque couche comprend à l'initialisation 8 noeuds (ce qui correspond à la taille des vecteurs spectraux utilisés). Le nombre de connexions est de 128 (i.e., réseau entièrement interconnecté).

— La base d'apprentissage comprend 200 à 500 couples de vecteurs.

— Le nombre d'itérations sur la base d'apprentissage est d'environ cinquante.

Le nombre d'apprentissages de connexions est de 128x500x50. Le temps d'apprentissage est donc de 320000/80000 soit 4 secondes. Pour des simulations, ce temps est très satisfaisant. Des architectures spécialisées sont nécessaires pour obtenir des vitesses d'apprentissage plus élevées. Deux réalisations peuvent être citées :

— La première fonctionne sur une machine à transputers [Petrowski et al. 89]. Il s'agit de processeurs spécialisés pour des calculs en parallèle. Ils sont spécialement adaptés aux algorithmes connexionistes. La vitesse obtenue est de 3 millions d'ACS pour 16 transputers.

— La deuxième réalisation consiste en une machine IBM spécialisée comportant 566 processeurs [Witbrock et Zagha 89]. Sa vitesse est de 900 millions d'ACS et devrait monter bientôt à plus de 1 milliard.

5.2. ALIGNEMENT DES SIGNAUX DE PAROLE

L'alignement temporel des signaux de parole se fait par comparaison dynamique des trames spectrales correspondant à chaque signal. Cette technique est très utilisée et

découle du principe de Bellmann [Bellmann 57]. Elle est a été introduite pour la première fois en traitement de la parole par T.-K. Vintsjuk [Vintsjuk 68]. Nous en donnerons ici le principe. Nous discuterons ensuite les limitations de l'algorithme de comparaison dynamique et des réponses que nous y avons apportées. Il s'agit de l'utilisation de cibles de décomposition temporelle au lieu de paramètres spectraux, de la détermination de la qualité d'un alignement dynamique et de l'introduction d'un algorithme itératif pour améliorer cet alignement.

5.2.1. Les algorithmes de comparaison dynamique

Le but de la comparaison dynamique est de mettre en correspondance temporellement des mots de durées différentes par une transformation non-linéaire afin d'obtenir une coïncidence optimale.

Soit $\{x_i\}$ (i = 1 ,.., r) et $\{y_j\}$ (j = 1 ,.., t). deux suites de paramètres correspondant respectivement à l'analyse d'un même mot (ou d'une même phrase) du locuteur de référence et d'un locuteur de test.

Soit une mesure de dissimilarité $\mathcal{D}$. On peut définir une matrice A (r x t) telle que $A(i, j) = \mathcal{D}(x_i,y_j)$.

L'ajustement temporel entre les $\{x_i\}$ et les $\{y_j\}$ est obtenu en définissant deux transformations m et n telles que m(k) = i et n(k) = j. Celles-ci représentent le chemin $\{C_k\}$ (k = 1 ,.., L) dans le plan r-t (cf. Figure 5.6).

$$\mathbf{C_k = (m(k),\ n(k)\ )} \qquad \mathbf{(5.21)}$$

L'ajustement temporel optimal est représenté par le chemin C_{opt} :

$$\mathbf{C_{opt} = \underset{C_k}{Argmin} \frac{\sum_{k=1}^{L} D(C_k)\ a_k}{\sum_{k=1}^{L} a_k}} \qquad \mathbf{(5.22)}$$

La distance entre les $\{x_i\}$ et les $\{y_j\}$ au sens de la comparaison dynamique est définie par l'équation suivante :

$$D(\{x_i\}, \{y_j\}) = \frac{\sum_{k=1}^{L_{opt}} D(C_{opt})\, a_k}{\sum_{k=1}^{L_{opt}} a_k}, \qquad (5.23)$$

où $a_k = a(k)$ est une fonction de pondération à valeurs positives.

Cette fonction de pondération permet de normaliser la distance $D(\{y_j\}, \{x_i\})$ pour que celle-ci soit indépendante des longueurs r et t ainsi que de la longueur du chemin optimal L_{opt}. Généralement la fonction de pondération a(k) est choisie de telle sorte que $\sum_{k=1}^{s} a(k)$ soit indépendant du chemin C choisi.

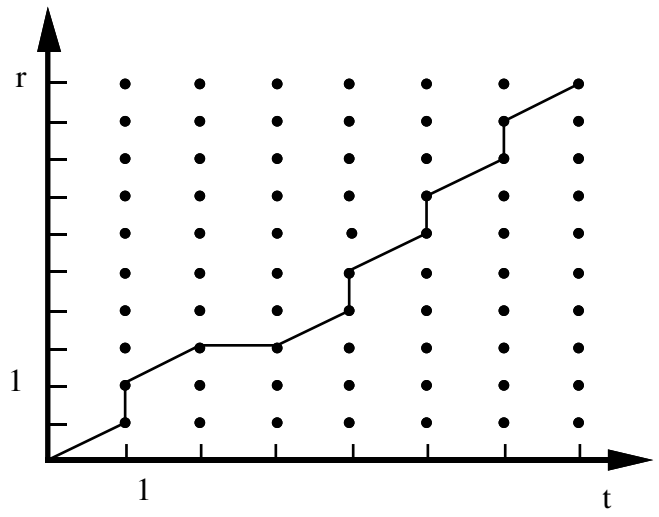


Figure 5.6. Ajustement temporel

La réalité physique impose aux fonctions m(k) et n(k) (représentant le chemin C) d'être croissantes. Cela permet d'ailleurs d'utiliser un algorithme récursif de programmation dynamique pour déterminer le chemin optimal C_{opt} correspondant à la distance entre les deux mots $D(\{y_j\}, \{x_i\})$. Des contraintes supplémentaires sont ajoutées pour éliminer des chemins n'ayant pas de sens physique. Elles permettent aussi de contrôler les distorsions temporelles minimales et maximales admises entre deux mots. On pourra trouver dans [Gauvain 82] la liste des contraintes les plus employées.

5.2.2. Les limitations de la comparaison dynamique

Cette procédure peut s'effectuer directement sur les paramètres centisecondes ou sur les cibles spectrales issues de la décomposition temporelle. La première méthode a été utilisée dans nos expériences préliminaires pour la reconnaissance de mots isolés [Montacié et al. 89] ; elle s'est révélé inadéquate dans le cadre du décodage acoustico-phonétique. La comparaison dynamique n'effectue pas un alignement entre deux signaux son par son. Il a été montré [Vicard 87] que des problèmes intervenaient sur les voyelles, où la variabilité de leur durée dépasse les contraintes généralement admises. La mise en correspondance de zones fortement transitoires très variables, comme les plosives, introduit aussi un biais dans l'estimation de la fonction d'alignement. Le recalage s'effectue alors sur les cibles de la décomposition temporelle. Il a permis d'obtenir des résultats plus satisfaisants que ce soit pour les erreurs de reconstruction des spectres après transformation ou pour les taux de reconnaissance après adaptation.

5.2.3. Calcul d'un indice de qualité de l'alignement dynamique

La validité de cet indice repose sur l'hypothèse que la corrélation entre les vecteurs alignés de deux locuteurs est une mesure de la qualité de l'alignement dynamique. Plus les vecteurs alignés sont corrélés, plus l'alignement est optimal. Inversement, si après un alignement dynamique, les vecteurs ne sont pas corrélés, l'alignement réalisé ne correspond à aucune réalité physique (e.g., alignement de chaque allophone).

L'indice est calculé en employant à la fois des techniques d'alignement dynamique et d'adaptation par régression multiple. L'indice de qualité est estimé à partir de la matrice de régression multiple A obtenue sur les paramètres alignés, centrés et réduits. Il est donné par la comparaison entre la matrice A et la matrice identité I. La matrice identité représente la matrice d'adaptation de l'alignement optimal (i.e., correspondant à des élocutions identiques). La comparaison entre les deux matrices A et I est mesurée par deux valeurs : le cosinus de l'angle α entre les deux matrices (en considérant la matrice comme un vecteur) et la distance pondérée λ entre les deux matrices.
Le cosinus de l'angle α entre les matrices A et I est calculé par la formule suivante :

$$\cos \alpha = \frac{\mathbf{A} \cdot \mathbf{I}}{\| \mathbf{A} \| \, \| \mathbf{I} \|} \qquad (5.24)$$

La distance pondérée λ entre les matrices A et I est calculée par la formule suivante :

$$\lambda = \frac{||\mathbf{A} - \mathbf{I}||}{||\mathbf{I}||} \qquad (5.25)$$

L'indice de qualité Q(A) est calculé comme $\mathbf{Q(A) = (1 - \lambda)\ cos\ \alpha}$

$$\mathbf{Q(A)} = (1 - \frac{||\mathbf{A} - \mathbf{I}||}{||\mathbf{I}||}) \frac{\mathbf{A} \cdot \mathbf{I}}{||\mathbf{I}||} \qquad (5.26)$$

Cet indice est égal à 1 pour deux nuages identiques (aux distorsions temporelles, centre de gravité et à la dynamique des coefficients près). Il tend vers zéro pour des vecteurs non-corrélés, l'alignement dynamique calculé ne correspond alors à rien, la fonction de transformation spectrale estimée est inexacte.

Cet indice a permis de déterminer la paramétrisation la plus adaptée à l'alignement dynamique d'élocutions de différents locuteurs (cf. Tableau 5.1, 5.2 et 5.3). Les résultats obtenus sont cohérents avec les performances du système de reconnaissance adapté, quelle que soit la méthode d'adaptation employée (i.e., régression multiple, filtrage polynomial, perceptrons multi-couches).

5.2.4. Algorithme d'alignement dynamique itératif

Cette technique permet d'améliorer l'alignement dynamique. La matrice de régression multiple A est utilisée pour calculer un alignement dynamique entre les vecteurs adaptés du locuteur de référence et ceux du nouveau locuteur. L'algorithme est résumé ci-dessous:

Etape 0: Les vecteurs de paramètres représentant tous les mots du 1er locuteur (locuteur de référence) sont centrés et réduits.
Les vecteurs de paramètres représentant tous les mots du 2ème locuteur (nouveau locuteur) sont centrés et réduits.
i est initialisé à 0.
La matrice d'adaptation A_i est initialisée par la matrice identité.

Etape i: Pour chaque mot, un alignement dynamique est effectué entre les vecteurs de paramètres du 1er locuteur adaptés par la matrice $A = \prod_{j=0}^{i} A_i$ et les vecteurs de paramètres du 2ème locuteur .

La matrice de régression multiple A_{i+1} est calculée sur l'ensemble des vecteurs alignés. L'indice de qualité $Q(A_{i+1})$ de l'alignement est calculé.

Si $\| Q(A_i) - Q(A_{i+1}) \| > \varepsilon$ donné et $i <$ maxI donné, i est incrémenté. Dans les autres cas l'algorithme s'arrête.

Cet algorithme n'est utilisable que pour un alignement initial de bonne qualité (i.e., $Q(A_1)$ proche de 1). Il diverge le plus souvent quand $Q(A_1)$ est faible (e.g., inférieur à 0.4) et il n'est utilisable que dans le cadre de l'apprentissage d'une fonction de transformation pour l'adaptation au locuteur. Il permet l'apprentissage d'une meilleure fonction de transformation entre ces couples de paramètres appariées.

0.82	0.176	0.016	0.069	-0.044	0.015	0.016	0.011	-0.002	0.062	-0.021	0.064
0.447	**0.38**	0.154	0.096	-0.005	0.002	-0.039	-0.019	0.048	0.160	-0.036	0.209
0.353	0.033	**0.35**	0.108	0.172	0.012	0.041	-0.068	0.111	-0.091	0.041	-0.061
0.037	0.361	0.062	**0.35**	0.068	0.085	-0.002	0.052	0.063	0.162	-0.076	0.237
-0.003	-0.017	0.153	0.109	**0.29**	0.157	0.110	0.109	-0.019	-0.134	-0.022	-0.229
-0.022	0.103	0.004	0.199	0.164	**0.29**	0.148	0.116	-0.045	0.073	-0.070	-0.080
-0.088	0.032	-0.056	0.173	0.052	0.174	**0.23**	0.173	0.029	0.017	0.070	-0.028
-0.143	-0.053	-0.016	0.061	0.048	0.131	0.042	**0.25**	0.065	0.152	0.049	-0.001
-0.101	0.074	-0.082	-0.052	-0.131	0.008	0.078	0.108	**0.16**	0.017	0.172	0.113
-0.016	-0.028	0.025	-0.131	-0.046	-0.053	0.003	-0.025	0.048	**0.18**	0.162	0.195
-0.069	0.009	-0.035	0.038	-0.184	-0.041	-0.032	-0.034	0.047	0.062	**0.20**	0.299
-0.166	0.190	-0.233	0.123	-0.134	-0.043	-0.059	0.018	-0.060	0.255	-0.080	**0.38**

Tableau 5.1. Matrice d'alignement[1] pour une paramétrisation LAR sur 12 coefficients d'un dictionnaire de logatome

$\cos \alpha_{LAR} = 0.61$

$1 - \lambda_{LAR} = 0.36$

$Q_{LAR}(A) = 0.22$

[1]matrice de régression multiple obtenue sur l'ensemble des paramètres alignés, centrés et réduits de deux locuteurs FB (masculin) et MJC (féminin)

0.14	0.027	0.024	0.015	0.027	0.033	0.002	0.030
-0.005	**0.10**	0.000	-0.012	-0.015	0.021	0.042	0.007
0.013	0.017	**0.11**	0.014	0.025	0.005	0.036	0.000
0.001	-0.023	0.019	**0.13**	-0.007	-0.014	-0.025	0.009
0.032	-0.001	0.013	0.009	**0.09**	0.001	-0.026	0.018
0.009	0.012	-0.021	-0.005	-0.002	**0.11**	0.015	0.005
-0.011	0.011	-0.000	-0.026	0.023	-0.010	**0.12**	-0.006
0.010	0.012	0.028	-0.007	-0.013	0.022	0.018	**0.09**

Tableau 5.2. Matrice d'alignement pour une paramétrisation WMFCC sur 8 coefficients d'un dictionnaire de logatome

$\cos \alpha_{WMFCC} = 0.91$

$1 - \lambda_{WMFCC} = 0.11$

$Q_{WMFCC}(A) = 0.10$

0.91	-0.019	0.016	0.006	0.028	0.017	0.019	-0.013
0.007	**0.71**	0.222	0.047	-0.020	0.031	-0.009	-0.076
-0.058	0.068	**0.82**	0.128	-0.015	-0.096	0.060	-0.011
0.043	0.014	-0.250	**0.50**	0.312	0.082	0.007	0.257
-0.093	0.167	-0.153	-0.341	**0.50**	0.147	-0.049	-0.029
-0.245	0.010	-0.016	0.124	-0.153	**0.34**	0.242	0.242
-0.034	-0.242	-0.146	0.057	-0.062	-0.230	**0.27**	0.273
-0.072	-0.150	-0.152	-0.018	0.347	0.054	-0.184	**0.41**

Tableau 5.3. Matrice d'alignement pour une paramétrisation WMFCC avec lissage cepstral itératif sur 8 coefficients d'un dictionnaire de logatome

$\cos \alpha_{WMFCC1} = 0.79$

$1 - \lambda_{WMFCC1} = 0.60$

$Q_{WMFCC1}(A) = 0.47$

On peut remarquer que l'indice de qualité de l'alignement dynamique est beaucoup plus élevé pour les WMFCC avec lissage cepstral itératif que pour les WMFCC classiques: le lissage cepstral itératif devrait donc permettre une meilleure adaptation entre les deux locuteurs.

5.3. CHOIX DU VOCABULAIRE D'ADAPTATION

Le vocabulaire d'adaptation est un vocabulaire restreint prononcé par un locuteur inconnu du système de reconnaissance. Il doit permettre d'obtenir des performances de reconnaissance similaires à celles d'un locuteur qui prononce le vocabulaire complet d'apprentissage.

Le vocabulaire d'adaptation optimal est celui qui permet d'obtenir les performances maximales.

Le vocabulaire caractérisant un locuteur connu (i.e., vocabulaire de référence) pour une application donnée est le plus souvent long à enregistrer et à étiqueter. Il est donc difficile de le prendre comme vocabulaire d'adaptation, bien qu'il soit proche du vocabulaire optimal. Nous avons admis que le vocabulaire d'adaptation devait avoir la même répartition dans l'espace des paramètres spectraux que le vocabulaire de référence tout entier.

Pour des raisons de temps de calcul, la comparaison des deux répartitions est faite en utilisant des techniques de quantification vectorielle [Gray 84]. Une arithmétique entière est utilisée au lieu d'une arithmétique flottante.

La procédure de comparaison est la suivante :

Par une méthode de quantification vectorielle (e.g., boules optimisées), on cherche les NV vecteurs spectraux représentant au mieux l'ensemble des paramètres des mots du vocabulaire de référence. Chaque mot est alors représenté par une suite d'étiquettes codant le vecteur de référence le plus proche du vecteur spectral initial. La répartition discrète des paramètres spectraux d'un mot est alors un vecteur de dimension n à valeurs entières contenant le nombre d'apparitions de la $i^{ème}$ étiquette. Une opération d'addition peut être définie sur ce vecteur de répartition. Le vecteur de répartition de l'ensemble de deux mots est la somme des deux vecteurs de répartition. Le vecteur de répartition globale S de tous les mots est la somme de tous les vecteurs de répartition des mots. La mesure de comparaison entre les deux vecteurs de répartition s_1 et s_2 est définie par :

$$\chi^2 (s_1,s_2) = \sum_{i=1}^{NV} \frac{(s_1[i] - A.s_2[i])^2}{A.s_2[i]}, \qquad (5.27)$$

où A est un facteur de normalisation valant $\sum_{i=1}^{NV} s_1[i] / \sum_{i=1}^{NV} s_2[i]$

La loi de probabilité de cette mesure de comparaison, appelée aussi χ^2 [Pearson], ne dépend que de la taille des vecteurs de répartition (i.e., nombre de degrés de liberté). Le calcul n'est cependant valable que si $A.s_2[i]$ est au moins égal à 10. Dans le cas contraire on regroupe les composantes des vecteurs les plus faibles. Si le χ^2 calculé est dans l'intervalle de confiance à 95% pour le nombre de degrés de liberté (cf. Tableau 5.4) on considère que la différence entre les vecteurs s_1 et s_2 n'est pas significative.

d.d.l.	1	2	3	4	5	6	7	8	9	10	11
Seuil	**3.8**	**6.0**	**7.8**	**9.5**	**11.1**	**12.6**	**14.1**	**15.5**	**16.9**	**18.3**	**19.7**
d.d.l.	12	13	14	15	16	17	18	19	20	21	22
Seuil	**21.0**	**22.4**	**23.7**	**25.0**	**26.3**	**27.6**	**28.9**	**30.1**	**31.4**	**32.7**	**33.9**
d.d.l.	23	24	25	26	27	28	29	30	31	32	33
Seuil	**35.2**	**36.4**	**37.7**	**38.9**	**40.1**	**41.3**	**42.6**	**43.8**	**45.0**	**46.2**	**47.4**

Tableau 5.4. Intervalle de confiance à 95% du χ^2 en fonction du nombre de degrés de liberté (d.d.l.)

Le sous-vocabulaire optimal $\mathbf{M_{opt}}$ de M mots parmi les P mots du vocabulaire tout entier est le suivant :

$$\mathbf{M_{opt} = \underset{\{M_i\}}{Argmin}\ \chi^2 \left(\sum_{i=1}^{NV} s_{Mi}\ ,\ S\right), avec\ M_i \in [1..P]} \qquad (5.28)$$

Malgré la réduction de temps calcul apportée par l'utilisation de la quantification vectorielle, le problème reste Non Polynomial réductible (i.e., dont le temps de calcul est proportionnel à l'exponentielle de la taille du vocabulaire). Le risque d'une explosion combinatoire apparaît pour un vocabulaire dépassant une vingtaine de mots. Nous avons alors défini une stratégie de choix sous-optimale, dont le temps de calcul est proportionnel au carré de la taille du vocabulaire, à l'aide de la procédure itérative suivante :

Etape 0:

Soit deux listes de mots $L_{+,0}$ et $L_{-,0}$,

on initialise $L_{+,0}$ à Ø et $L_{-,0}$ avec le vocabulaire tout entier

Etape k:

a) on cherche le mot $V_{+,k}$ du complémentaire de $L_{+,k}$ par rapport à S qui minimise $\chi^2\ (L_{+,k} + V_{+,k}\ ,\ S)$,

$L_{+,k+1} = L_{+,k} + V_{+,k}$

b) on cherche le mot $V_{-,k}$ du complémentaire de $L_{-,k}$ par rapport à S qui minimise $\chi^2 (L_{-,k} - V_{-,k} , S)$,

$L_{-,k+1} = L_{-,k} - V_{-,k}$

c) on incrémente k tant que $L_{-,k+1}$ est non vide.

A la fin de cet algorithme, deux ensembles de listes de sous-vocabulaires proches de la répartition globale sont obtenus. Pour un nombre M de mots utilisés, il suffit de prendre la liste $L_{+,M}$ ou $L_{-,M}$ qui minimise la distance par rapport à S. Cette stratégie, sous-optimale comparée à la stratégie optimale par recherche combinatoire, ne présente pas de différences notables.

5.3.1. Expériences préliminaires sur une base de mots isolés

Des expériences préliminaires ont eu lieu sur la base TEXAS de mots isolés [Lea 82]. Cette base comprend 26 élocutions de 20 mots prononcés par 16 locuteurs (seulement 14 locuteurs ont été utilisés). La quantification vectorielle a eu lieu sur 32 vecteurs codes. Le vocabulaire d'adaptation dans l'expérience initiale [cf. Annexe VIII] [Montacié et al. 89] [Tubach et al. 90] était composé de cinq mots ("no", "one", "two", "five", "six"). Nous avons refait l'expérience en utilisant le vocabulaire d'adaptation optimal tel que nous l'avons défini plus haut.

Numéro	1	2	3	4	5	6	7	8	9	10
Mot	Yes	No	Erase	Rubout	Repeat	Go	Enter	Help	Stop	Start
Numéro	11	12	13	14	15	16	17	18	19	20
Mot	One	Two	Three	Four	Five	Six	Seven	Eight	Nine	Zero

Tableau 5.5. Vocabulaire de la base de données TEXAS

Le système de reconnaissance employé est un système de reconnaissance de mots connectés par programmation dynamique. La première réalisation sert pour l'adaptation, les 25 autres constituent le test. Une référence de silence est construite en prenant les trois premières trames de la première réalisation du premier mot. La syntaxe est de type silence-mot-silence.

Locuteur	degrés de liberté	χ^2 Minimal	Vocabulaire	Ancien χ^2
CJP	16	25.8	2 4 8 13 16	60.0
ALK	14	15.3	3 5 6 11 15	67.2
GNL	16	14.7	3 8 9 11 13	93.4
JWS	17	13.5	9 11 17 18 20	83.0
MSW	12	9.4	3 8 9 11 20	42.0
RLD	13	15.5	2 3 8 9 13	23.0
SJN	13	6.2	5 6 7 10 17	47.6
WMF	12	14.2	1 5 6 7 17	61.0
DFG	17	23.1	3 4 10 11 12	37.6
GRD	14	9.7	2 10 11 13 16	29.4
KAB	14	6.7	1 4 8 11 20	57.4
REH	12	8.1	2 5 7 10 15	32.2
SAS	15	10.9	1 2 10 11 13	40.6
TBS	13	8.8	4 8 13 16 19	51.7

Tableau 5.6. Vocabulaire optimal d'adaptation de cinq mots

On peut remarquer que les mots d'adaptation ne sont pas les mêmes pour tous les locuteurs. Il n'y a pas non plus de mots plus représentatifs que d'autres.

degrés de liberté	χ^2	Vocabulaire
7	3.4	8
7	13.7	8 13
10	11.0	1 4 20
13	6.7	2 7 9 18
16	14.7	3 8 9 11 13
18	25.9	3 8 9 13 17 20

Tableau 5.7. Vocabulaires optimaux d'adaptation (locuteur GNL)

Pour un même locuteur, on peut remarquer que le sous-vocabulaire optimal de taille n est souvent contenu, en partie, dans les sous-vocabulaires optimaux de taille supérieure à n. Cela confirme la validité de notre stratégie sous-optimale de recherche du vocabulaire d'adaptation.

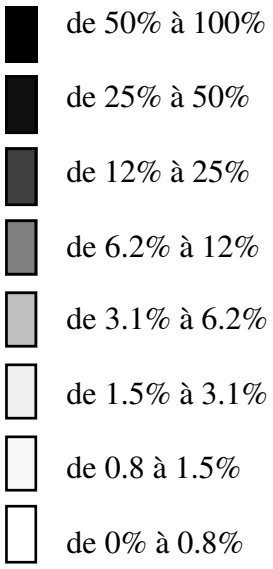


Figure 5.7. Teintes de gris associées à la fréquence d'une composante

La fréquence P(s[i]) de la composante i d'un vecteur de répartition s de taille N est définie par :

$$P(s[i]) = s[i] \,/\, \left(\sum_{j=1}^{NV} s[j] \right) \qquad (5.29)$$

Nous avons représenté sous forme de teintes de gris, les vecteurs de répartition associés à chacun des vingt mots pour le locuteur. Comme la quantification vectorielle a eu lieu sur NV=32 vecteurs codes, la fréquence de chaque composante est en moyenne de 1/32, soit environ 3.4% ce qui correspond au milieu de la table des teintes de gris.

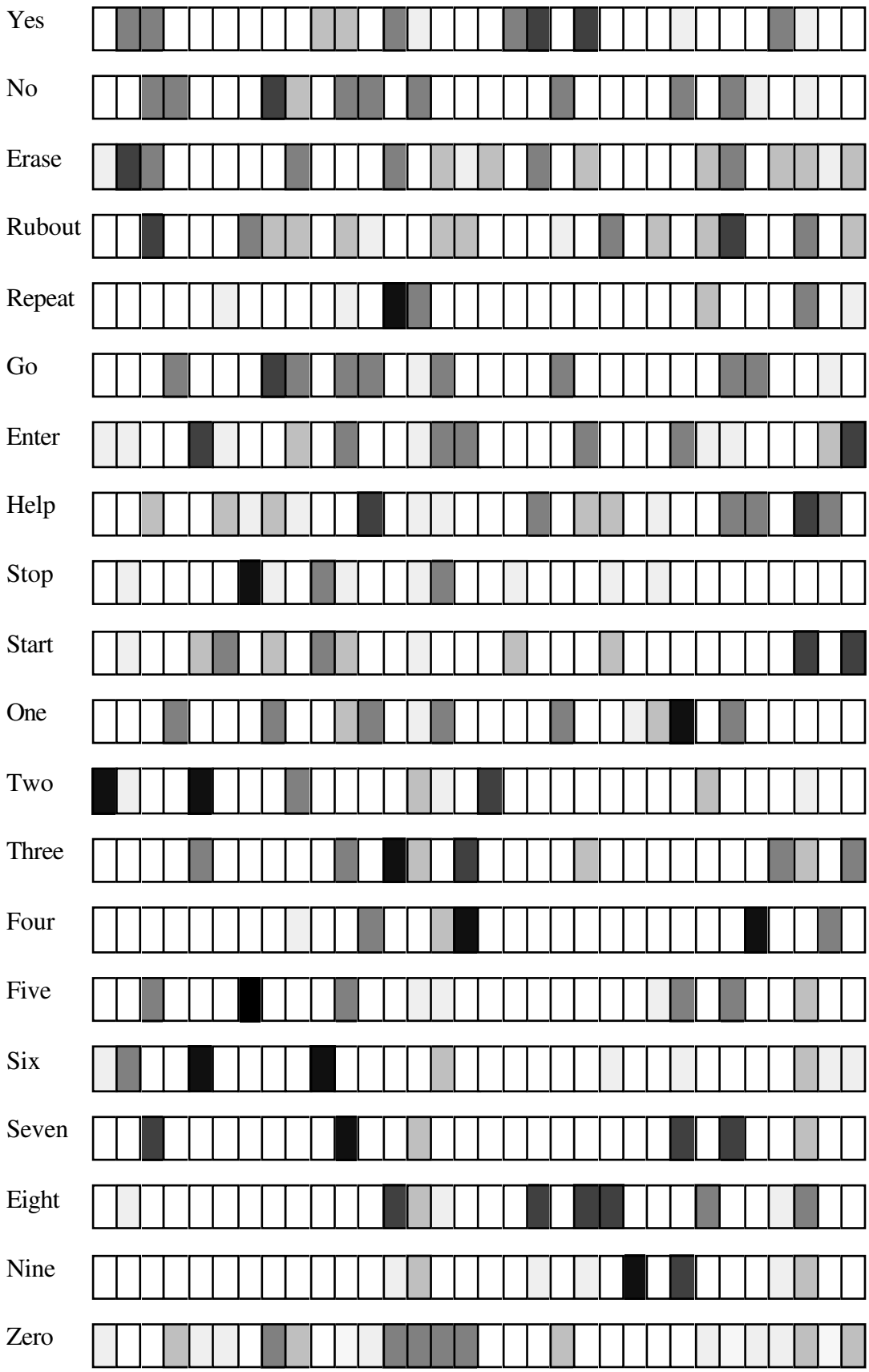


Figure 5.8. Représentation des mots du vocabulaire (locuteur GNL)

On peut constater sur la figure 5.8. que des mots au contenu phonétique proche se ressemblent comme "no" et "go" , mais on ne retrouve pas le son (f) contenu dans les mots "five" " et "four".

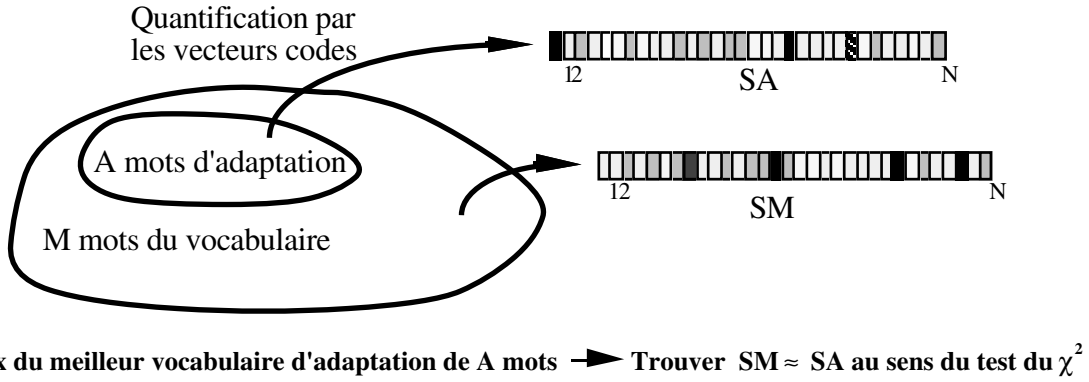


Figure 5.9. Synoptique du choix du vocabulaire d'adaptation

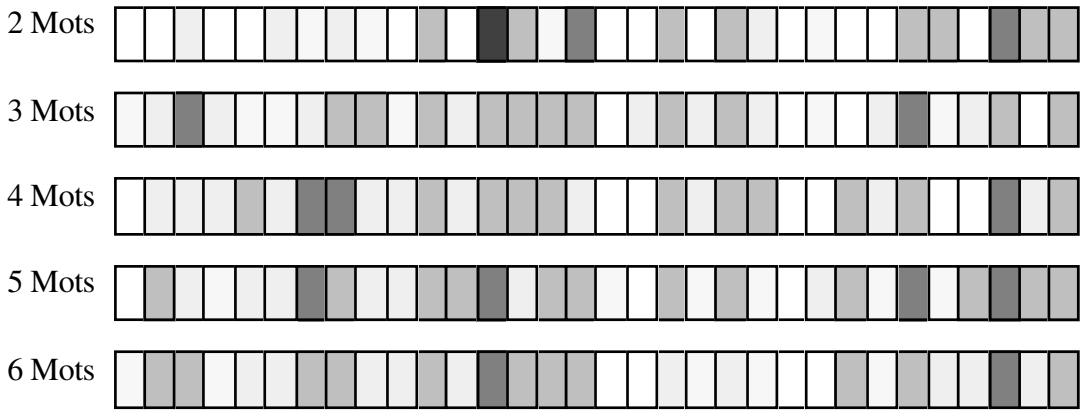


Figure 5.10. Représentation du meilleur sous-vocabulaire représentatif (locuteur GNL)

Les mots choisis sont ceux indiqués dans le tableau 5.7. On peut remarquer que plus le sous-vocabulaire est important, plus sa représentation tend vers la représentation du vocabulaire tout entier (cf. Figure 5.11.)

Figure 5.11. Représentation du vocabulaire tout entier (locuteur GNL)

mot	réussite 1ère expérience	réussite 2ème expérience	reconnaissance 1ère expérience	reconnaissance 2ème expérience	écart-type réduit
1	151	395	89.6%	95.0%	10.6
2	803	484	75.8%	68.8%	9.0
3	52	494	88.3%	98.0%	19.7
4	290	305	90.3%	90.6%	0.6
5	122	458	88.4%	95.8%	14.3
6	469	852	36.0%	44.5%	10.7
7	108	527	85.3%	94.5%	17.2
8	200	1145	68.2%	88.9%	27.9
9	331	390	88.2%	89.5%	2.2
10	66	692	82.9%	96.7%	24.1
11	364	114	96.1%	90.6%	11.6
12	673	74	96.2%	83.1%	23.2
13	98	1040	75.1%	95.8%	30.7
14	335	462	86.2%	89.0%	4.5
15	1148	282	86.4%	67.3%	24.3
16	91	38	99.1%	97.9%	4.7
17	239	97	97.3%	94.2%	7.8
18	386	832	67.7%	77.5%	13.0
19	543	587	74.4%	75.4%	1.3
20	183	513	86.9%	94.2%	12.7
tout	**6652**	**9781**	**82.9%**	**86.4%**	**24.7**

Tableau 5.8. Comparaison des deux expériences

Les résultats dans le tableau 5.8. montrent que notre choix optimal de vocabulaire augmente le taux de reconnaissance global de plus de 3% (l'intervalle de confiance pour les 91000 reconnaissances de mots étant de 0.5%). La colonne "réussite 1ère expérience" donne le nombre de fois où la première expérience [Tubach et al. 90] reconnait le bon mot quand la deuxième expérience ne le reconnait pas (et inversement pour la colonne "réussite 2ème expérience"). Si l'écart-type réduit est supérieur à 2, les différences entre les deux expériences sont significatives. On peut remarquer que parmi les six mots que la deuxième expérience reconnaît moins bien se trouvent les cinq mots d'adaptation de la première expérience.

CONCLUSIONS

De l'étude des différents problèmes rencontrés pour l'adaptation au locuteur de notre système de décodage acoustico-phonétique, nous pouvons tirer les enseignements suivants :

— Le perceptron multi-couches permet une meilleure estimation d'une fonction de transformation que la régression linéaire multiple, avec moins de paramètres que le filtrage quadratique. Au cours d'expériences préliminaires sur la reconnaissance de mots isolés, cette technique a permis une augmentation faible, mais significative, du taux d'adaptation moyen.

— L'indice de qualité de l'alignement dynamique permet de déterminer la paramétrisation la plus adaptée pour l'adaptation au locuteur d'un système de reconnaissance.

— La procédure automatique pour le choix du vocabulaire d'adaptation, par rapport à un choix arbitraire, a permis une augmentation de plus de 3% du taux de reconnaissance au cours d'expériences préliminaires sur la reconnaissance de mots isolés.

Nous allons maintenant appliquer l'ensemble de ces techniques en évaluant notre système de décodage acoustico-phonétique.

CHAPITRE 6. EVALUATION DU SYSTÈME DE DÉCODAGE ACOUSTICO-PHONÉTIQUE

Pout mettre en valeur les améliorations apportées par notre système de décodage acoustico-phonétique par rapport à l'état de l'art du domaine, un corpus de parole de mots épelés en continu a été choisi. Les conditions d'expérimentation ont été multiples. Elles ont porté sur le choix de la paramétrisation, le type d'alignement des signaux de parole et les fonctions d'adaptation choisies.

Un système de référence, choisi parmi les meilleures techniques de reconnaissance de la parole et reposant sur les machines de Markov cachées, permet de faire une évaluation comparative.

6.1. CORPUS

6.1.1. Corpus de test

Le corpus de test est constitué de 50 noms de famille épelés en mode continu (sans pause entre des lettres consécutives) [cf. Annexe VI]. Il correspond aux corpus REC01 et REC02 de la base de données de parole BDSON du Greco [Cervantes et Sérignat 86]. C'est un corpus difficile, car les unités lexicales qui le constituent forment plusieurs séries phonétiques minimales, sources de confusions : {(a), (e), (i), (o), (y), (é)}, {(bé), (sé), (dé), (xé), (pé), (té), (vé), (é)}, {(èf), (èl), (èm), (èn), (èr), (ès)}, {(a), (ac)}, {(xé), (xi)}, {(i), (xi)}, {(a), (ka)}, {(ka), (ky)}, {(ky), (y)}. Ce corpus contient 513 sons (pour environ 300 lettres). Deux locuteurs ont été étudiés : l'un de sexe masculin (FB), l'autre de sexe féminin(MJC). Les corpus ont été enregistrés à Télécom-Paris.

Nous n'avons pas utilisé les bases de données BDSON pourtant pluri-locuteurs. En effet, l'apprentissage proposé (ALA0x) constitué de quatre prononciations de l'alphabet en mode isolé ne nous a pas paru représentatif, dans l'objectif d'une reconnaissance efficace. Par ailleurs, l'enchaînement de lettres épelées entraîne de nombreux effets de coarticulation, que les méthodes globales (e.g., comparaison dynamique et machines de Markov cachées) ne parviennent pas à résoudre de façon satisfaisante.

6.1.2. Corpus d'apprentissage

L'apprentissage est conduit sur un ensemble de 269 logatomes [cf. Annexe V] porteurs de tous les diphones possibles dans une suite de lettres prononcées continûment. Outre les contextes (non utilisés pour l'apprentissage), ce corpus contient 513 sons et a été prononcé une fois par chacun des deux locuteurs (l'un féminin, l'autre masculin).

Chaque son est représenté 15 ou 30 fois (environ), selon qu'il apparaît seulement en début ou en fin de certaines lettres, ou bien dans les deux positions. Les cibles de transitions sont conservées et sont pour ce corpus au nombre d'une centaine.

6.2. CALCUL DU TAUX DE RECONNAISSANCE

6.2.1. Alignement de la chaîne phonétique décodée

L'évaluation précise du décodage acoustico-phonétique est difficile. La syntaxe n'est pas contrainte. Le nombre maximum de sons attendus est libre lui aussi (nous avons cependant dû le limiter à 20 sons attendus pour le système de référence). De nombreuses différences peuvent apparaître entre la succession des événements décodés et celle des sons attendus. Il peut y avoir des insertions, des substitutions ou des sons non trouvés. L'algorithme de Wagner et Fisher [Wagner et Fisher 74] [cf. Annexe VII] est utilisé pour trouver l'alignement optimal entre les deux chaînes.

La figure 6.1. donne un exemple du décodage fourni par notre système de reconnaissance. Il s'agit du mot "DURAND" ((déyèraèndé)). Sur dix sons attendus, seulement six sons ont été trouvés ((é), (y) , (a), (è), (n), (é)), deux sons ont été substitués ((d) / (b), (è) / (é)) et deux sons ont été omis ((r), (d)), un son inséré ((l)).

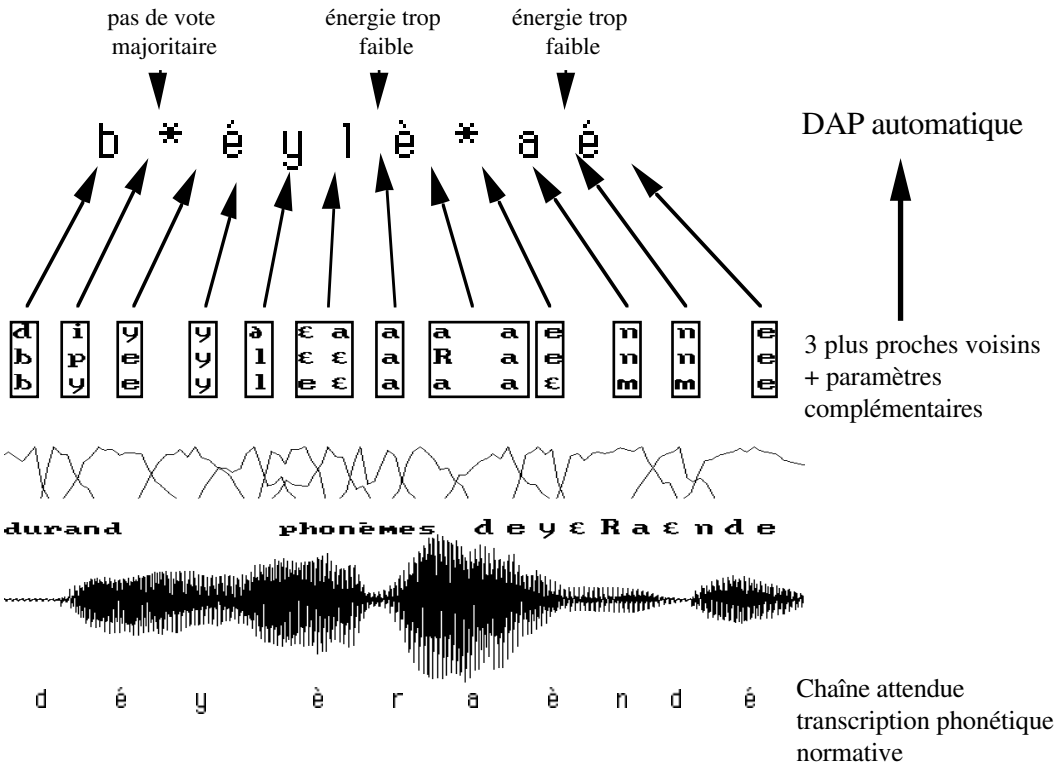


Figure 6.1. Décodage acoustico-phonétique du mot "DURAND" épelé en continu

6.3 SYSTÈME DE RÉFÉRENCE COMPARATIF

Une méthode d'évaluation est de se servir de systèmes de référence. Il s'agit d'un système de reconnaissance qui doit refléter l'état de l'art, que ce soit au niveau des algorithmes employés ou du taux de reconnaissance. Il doit offrir une panoplie de représentations du signal de parole et de distances interspectrales. Il doit être facilement implantable pour que de nombreux laboratoires l'adoptent.

6.3.1 Description

Le système de référence utilisé est une carte de reconnaissance de parole à base de machines de Markov connectées. Elle fonctionne en temps réel grâce à un processeur de signal TMS320C25. L'analyse du signal est de type MFCC. La syntaxe choisie ainsi que le type de topologie des machines de Markov sont entièrement programmables. Le besoin de résultats significatifs nécessite la recherche préalable de la topologie adaptée à notre corpus.

6.3.2 Topologies employées

Les tests ont été effectués en employant neuf topologies différentes suivant le modèle de Bakis (cf. Figure 1.1.). Le nombre d'états varie de de 1 à 5, le nombre de distributions de 1 à 12. L'ensemble des résultats du système de référence et la description complète de la syntaxe et des modèles employés se trouvent dans l'Annexe II et III.

Modèles	Nombre d'états	Nombre de transitions	Nombre de distributions
A	1	1	1
B	2	3	2
C	2	3	3
D	3	6	3
E	3	6	6
F	4	9	4
G	4	9	9
H	5	12	5
I	5	12	12

Tableau 6.1. Topologie des différents modèles employés

Le tableau 6.2. donne deux résultats du système de référence pour 12 expériences de décodage acoustico-phonétique. La première colonne correspond aux performances du modèle s'étant révélé le plus performant à l'apprentissage. La seconde colonne donne les meilleurs résultats du modèle le plus performant pour le test. Dans une application réelle seule la première colonne pourrait être obtenue.

Le taux de décodage acoustico-phonétique est en moyenne de 60% avec un taux d'insertion de 30%. L'apprentissage d'un modèle commun pour les deux locuteurs permet d'augmenter les performances de 4%.

Les faibles performances en reconnaissance croisée (test avec un autre locuteur que celui de l'apprentissage) montrent qu'une procédure d'adaptation serait utile pour le système de référence comparatif.

Corpus d'apprentissage	Corpus de test	Meilleur modèle pour l'apprentissage				Meilleur modèle pour le test			
		Trouvés	Substitués	Omis	Insérés	Trouvés	Substitués	Omis	Insérés
SL[a]	SL[a]	63%	31%	6%	21%	-	-	-	-
SL[a]	SL[c]	**60%**	**34%**	**6%**	**33%**	**60%**	**34%**	**6%**	**33%**
FB[a]	FB[a]	75%	20%	5%	16%	-	-	-	-
FB[a]	FB[c]	**55%**	**35%**	**10%**	**35%**	**57%**	**36%**	**7%**	**29%**
FB[b]	FB[b]	80%	19%	1%	8%	-	-	-	-
FB[b]	FB[c]	**58%**	**35%**	**7%**	**23%**	**58%**	**35%**	**7%**	**23%**
MJC[b]	MJC[b]	79%	19%	2%	2%	-	-	-	-
MJC[b]	MJC[c]	**55%**	**39%**	**6%**	**19%**	**63%**	**33%**	**4%**	**40%**
FB[b] MJC[b]	FB[c]	**59%**	**34%**	**7%**	**18%**	**63%**	**30%**	**7%**	**18%**
FB[b] MJC[b]	MJC[c]	**62%**	**33%**	**5%**	**26%**	**66%**	**31%**	**3%**	**39%**
MJC[b]	FB[c]	**30%**	**43%**	**28%**	**6%**	**34%**	**55%**	**11%**	**22%**
FB[b]	MJC[c]	**22%**	**53%**	**24%**	**9%**	**33%**	**59%**	**8%**	**32%**

Tableau 6.2. Résultats du système de référence, 1er choix, sur 513 sons
[a] corpus de lettres isolées
[b] corpus de logatomes
[c] corpus de noms épelés en continu

6.4. RÉSULTATS ET DISCUSSION

Le Tableau 6.3. donne les scores de reconnaissance mono-locuteur. Toutes les différences entre les étiquettes phonétiques attendues et produites, sans aucun regroupement des sons en macro-classes ont été comptées comme confusions. L'intervalle de confiance à 95% des scores de reconnaissance est de 3 % [Montacié et al. 87].

6.4.1. Choix de la paramétrisation

Bien que les LAR aient un comportement plus réaliste au sens de l' interpolation linéaire simple des cibles spectrales et permettent une re-synthèse du signal [Bimbot 88], les WMFCC permettent comparativement une reconnaissance globale de mots isolés de qualité bien supérieure [Cordeau 88].

Ces résultats se confirment dans nos expériences de décodage par décomposition temporelle. L'amélioration est d'environ 15 % (pour 8 WMFCC vs 12 LAR, et une métrique euclidienne). L'introduction du lissage cepstral itératif permet, par ailleurs, d'améliorer l'adaptation d'un dictionnaire de cibles de référence à un locuteur différent.

Les expériences de reconnaissance avec la distance d'Itakura et Saito symétrique n'ont pas eu lieu car le temps calcul demandé est pour l'instant trop important. Cela est dommage car les résultats préliminaires de la décomposition temporelle avec cette distance donnent des résultats tout-à-fait encourageants (cf. paragraphe 2.6. et Fig 2.8.).

Locuteur	Paramétrisation	Trouvés	Substitués	Omis	Insérés
FB(M)	LAR	50%	39%	11%	14%
FB(M)	WMFCC	73%	19%	8%	16%
FB(M)	WMFCC lissés	74%	19%	7%	17%
MJC(F)	LAR	58%	32%	10%	23%
MJC(F)	WMFCC	64%	29%	7%	27%
MJC(F)	WMFCC lissés	67%	26%	7%	27%

Tableau 6.3. Reconnaissance phonétique, 1er choix, sur 513 sons
(M) → Locuteur de sexe masculin
(F) → Locuteur de sexe féminin

Les résultats de l'évaluation comparative utilisant des machines de Markov [Annexe II] donnent des résultats inférieurs de 16% pour le locuteur masculin et 4% pour le locuteur féminin pour la meilleure topologie de test. Dans le cas où le modèle de test choisi est celui qui donne de meilleurs résultats à l'apprentissage, la différence est de 12% pour le locuteur féminin, et de 16% pour le locuteur masculin.

6.4.2. Différences Homme-Femme et confusions

Les scores de reconnaissance montrent de meilleurs résultats sur les voix masculines que sur les voix féminines, ce qui paraît être plutôt un artefact de la paramétrisation spectrale employée (malgré l'introduction du lissage cepstral itératif) qu'un biais introduit par la décomposition temporelle proprement dite.

Les confusions les plus fréquentes sont : (m) / (n), (e) / (l), (é) / (è), (b) / (d), (f) / (s). (l) et (r) sont plutôt bien reconnus au prix d'un taux d'insertion

sensiblement plus élevé de ces sons. Les explosions de (p), durée de l'ordre de 20 ms, sont extrêmement mal détectées par la décomposition temporelle, ce qui a pour conséquence un score très bas pour ce son.

6.4.3. Adaptation au locuteur

Le vocabulaire d'adaptation est composé de 50 logatomes sur les 269 logatomes du vocabulaire d'apprentissage. Ces logatomes ont été choisis par la procédure de choix du vocabulaire d'adaptation optimal définie dans le chapitre précédent. Dans le tableau 6.3. ne sont présentés que les résultats avec des coefficients cepstraux en échelle Mel et lissage cepstral itératif. Les résultats de DAP obtenus avec les coefficients cepstraux classiques sont inférieurs de 5% à 10% suivant la méthode d'adaptation utilisée. Les coefficients LAR ne donnent pas de résultats de DAP supérieur à 40%.

Locuteur de référence	Locuteur de test	Type de transformation	Trouvés	Substitués	Omis	Insérés
MJC	FB	sans	42%	47%	11%	35%
MJC	FB	régression multiple	50%	43%	7%	25%
MJC	FB	filtrage quadratique réduit	40%	49%	11%	26%
MJC	FB	perceptron multi-couches	**54%**	41%	5%	26%
FB	MJC	sans	60%	34%	6%	54%
FB	MJC	régression multiple	**70%**	**25%**	**5%**	**38%**
FB	MJC	filtrage quadratique réduit	69%	26%	5%	39%
FB	MJC	perceptron multi-couches	68%	26%	6%	38%

Tableau 6.4. Adaptation à de nouveaux locuteurs

Le tableau 6.4. donne les scores d'adaptation en fonction de la technique d'adaptation employée. Le lissage cepstral itératif permet une meilleure reconnaissance dans le cas difficile d'une adaptation homme-femme. Le filtrage quadratique réduit ne donne pas les résultats attendus en raison du faible nombre de couples d'apprentissage. Les réseaux connexionistes semblent plus performants en moyenne, mais la taille réduite du corpus de test permet difficilement de conclure sur le meilleur type de transformations pour l'adaptation au locuteur de notre système de décodage acoustico-phonétique.

CONCLUSIONS

Les résultats obtenus ont validé l'ensemble de l'approche utilisée par notre processus de décodage. Le taux de décodage acoustico-phonétique est supérieur de 10% à celui du système de référence choisi, c'est-à dire à base de machines de Markov connectées.

Nous allons essayer d'estimer la part de nos différents apports (parmétrisation, modèle de la trajectoire spectrale) dans cette amélioration :

— L'utilisation des MFCC avec un lissage cepstral itératif apporte un gain de 2% en moyenne par rapport au MFCC classiques utilisés par le système de référence. La différence apparaît surtout pour le locuteur féminin.

— Il est plus difficile de calculer l'apport exact des cibles spectrales par rapport aux paramètres spectraux originaux. On peut cependant remarquer que, sur la base d'apprentissage (cf. tableau 4.1), le taux d'autocohérence des cibles spectrales est supérieur de 5% à celui des paramètres spectraux originaux.

Les procédures d'adaptation utilisées se sont révélées efficaces, avec une augmentation du taux de décodage acoustico-phonétique de plus de 10% par rapport à ce que l'on obtient en l'absence d'adaptation, pour un locuteur de test différent de celui de référence.

CONCLUSIONS ET PERSPECTIVES

Nous reviendrons dans cette conclusion sur l'essentiel de notre recherche avant d'exposer ses développements futurs.

— Cette thèse a conduit à la mise en oeuvre d'un **ensemble de méthodes** pour obtenir un **décodage acoustico-phonétique performant** de la voix continue. Chacune des méthodes utilisées a été testée indépendamment pour mettre en valeur son apport.

La **décomposition temporelle généralisée** est un outil efficace et robuste pour décrire l'évolution spectrale du signal de parole. Les cibles spectrales permettent de retrouver une partie de la structure acoustico-phonétique sous-jacente comme le montrent nos expériences de reconnaissance. La possibilité de choisir une **métrique quelconque** améliore de manière indiscutable la flexibilité de cet outil.

L'implémentation et le développement d'un algorithme de **lissage cepstral itératif**, pour le calcul de nos paramètres d'analyse, améliore l'adaptation de notre système de décodage acoustico-phonétique à des locuteurs des 2 sexes en minimisant l'influence de la fréquence fondamentale. L'introduction d'un **indice de qualité de l'alignement dynamique** a permis de prévoir cette amélioration de l'adaptation.

L'utilisation du **perceptron multi-couches** pour le calcul de **transformations spectrales non-linéaires** permet une augmentation faible mais significative du taux d'adaptation moyen. L'**initialisation linéaire du perceptron multi-couches** et la **séparation des noeuds saturés** ont permis d'améliorer la stratégie d'apprentissage et de modifier dynamiquement l'architecture du réseau. A cette occasion, les relations entre les perceptrons multi-couches et les techniques classiques d'analyse linéaire comme la régression multiple ont été mises en évidence.

La recherche du vocabulaire d'adaptation par l'utilisation conjointe de la quantification vectorielle et d'un test statistique a montré que le **vocabulaire d'adaptation optimal** (du moins pour les petits vocabulaires) dépendait du locuteur et ne pouvait être prédit a priori.

Le taux de décodage acoustico-phonétique obtenu par notre système de reconnaissance, **70% au premier choix**, est supérieur à celui d'un **système de reconnaissance de référence** basé sur les **machines de Markov cachées (60% au premier choix)**

— Nos efforts vont porter désormais sur la prise en compte de **connaissances complémentaires** pour enrichir notre système de décodage. Par exemple, la décomposition temporelle modélise difficilement les **transitions brutales** dues à la variation du mode de production au niveau de la source. Cette difficulté est la plupart du temps causée par les insuffisances de la paramétrisation comme le montre l'**indice d'instabilité pour les WMFCC**. La localisation de telles zones permettra d'éviter des décompositions temporelles non pertinentes, compte-tenu de l'inadéquation du modèle à cet endroit. Par ailleurs, l'opérateur de reconnaissance de notre système de décodage acoustico-phonétique n'utilise que les cibles spectrales. La **prise en compte des fonctions d'interpolation** pour la reconnaissance devrait améliorer la puissance de cet opérateur. Nos efforts doivent aussi s'orienter vers la recherche d'une **métrique appropriée à la description des trajectoires spectrales**.

Enfin, il ne fait guère de doute que nos méthodes doivent être testées sur des **corpus beaucoup plus larges**, faisant varier les situations qu'il s'agisse des **conditions d'élocution** ou de **bruit**. C'est alors seulement, que l'on pourra accroître la précision de nos évaluations des gains apportés par le décodage acoustico-phonétique de la parole continue par décomposition temporelle généralisée et transformations spectrales non-linéaires.

RÉFÉRENCES BIBLIOGRAPHIQUES

[Ahlbom et al. 87] G. AHLBOM, F. BIMBOT et G. CHOLLET : **Modeling Spectral Speech Transitions Using Temporal Decomposition Techniques**.IEEE ICASSP, pp. 13-16, 1987

[Akaike 71] H. AKAIKE : **Information Theory and an Extension of the Maximum Likehood Principle.** 2nd Int. Symp. on Informatic Theory, Tsakhadsor, Arménie, URSS, 1971.

[André-Obrecht et Delyon 89] R. ANDRE-OBRECHT et B. DEYLON : **Etude des segments transitoires à l'aide de modèles AR évolutifs et de critère d'Akaike**. GRETSI, pp. 501-503, Juans-Les-Pins, France, 1989.

[André-Obretch 88] R. ANDRE-OBRECHT : **A New Statistical Approach for the Automatic Segmentation of Continuous Speech** : IEEE Trans. ASSP, Vol. 36, pp. 29-40, 1988.

[Artières et al. 91] T. ARTIERES, Y. BENNANI, P. GALLINARI et C. MONTACIE : **Connectionist and Conventional Models for Free-Text Talker Identofication Tasks**. Neuro-Nimes '91, Nimes, France, 1991.

[Atal 83] B.-S. ATAL : **Efficient Coding of LPC Parameters by Temporal Decomposition.** IEEE ICASSP, pp. 81-84, 1983.

[Averbuch et al. 87] A. AVERBUCH, L. BAHL, P. BROWN, G. DAGGETT, S. DAS, K. DAVIES, S. DE GENNERO, P. DE SOUZA, E. EPSTEIN, D. FRALEIGH, F. JELINEK, B. LEWIS, R. MERCER, J. MOORHEAD, A. NADAS, D. NAHAMOO, M. PICHENEY, G. SHICHMAN, P. SPINELLI, D. VAN COMPERNOLLE et H. WILKENS : **Experiments with the TANGORA 20,000 Word Speech Recognizer.** IEEE-ICASSP., pp. 701-704, Dallas, USA, April 1987.

[Baker 75] J.-K. BAKER : **Stochastic Modeling for Automatic Speech Understanding Speech Recognition**. R. Reddy ed., New York, Academic Press, 1975.

[Basseville 89] M. BASSEVILLE : **Distance Measure for Signal Processing and Pattern Recognition.** Signal Processing. Vol 18, n°4, 1989.

[Baum 72] L.-E. BAUM : **An Inequality and Associated Maximization Technique in Statistical Estimation of Probabilistic Functions of Markov Process.** Inequalities, Vol. 3, pp. 1-8, 1972.

[Baum 90] E.-B. BAUM : **When are K-Nearest Neighbours and Back Propagation Accurate for Feasible Sized Sets of Examples?**. EURASIP Workshop, Sesimbra, Portugal, Février 1990.

[Bellmann 57] R. BELLMANN : **Dynamic Programming.** Princeton University Press, 1957.

[Bennani 88] Y. BENNANI : **Décodage acoustico-phonétique par réseaux connexionnistes.** Technical Report 88D016, ENST, 1988.

[Benzécri 70] J.-P. BENZECRI : **Représentation euclidienne d'un ensemble muni de masses et de distances.** Polycopié, Institut de Statistique de l'Université de Paris 6, Paris, 1970.

[Bimbot 88] F. BIMBOT : **Synthèse de la parole : des segments aux règles, avec utilisation de la décomposition temporelle.** Thèse de doctorat (PhD), ENST, Décembre 1988.

[Bimbot et al. 87] F. BIMBOT, G. AHLBOM et G. CHOLLET : **From Segmental synthesis to Acoustic Rules Using Temporal Decomposition.** ICPhS., Vol. 5, pp. 31-34, Août 1987.

[Bimbot et al. 88] F. BIMBOT, G. CHOLLET, P. DELEGLISE et C. MONTACIE : **Temporal Décomposition and Acoustic-Phonetic Decoding of Speech.** IEEE-ICASSP., pp. 445-448, New York, USA, April 1988.

[Bonneau et Gauvain 87] H. BONNEAU et J.-L. GAUVAIN : **Vector Quantization for Speaker Adaptation.** IEEE-ICASSP., pp. 1434-1437, Dallas, USA, 1987

[Bouliane et al. 86] G. BOULIANE, J.-P. TUBACH, Y. GRENIER et G. CHOLLET : **Recognition of Non Stationary Speech Segments using Autoregressive Time-Dependent Models.** ICPR, pp. 57-59, Paris, France, 1986.

[Bourjot et al. 88] C. BOURJOT, A. BOYER, G. PERENNOU, J.-P. TUBACH et N. VIGOUROUX : **Analyse de deux enquêtes sur l'évaluation des systèmes de reconnaissance à grand vocabulaire et du décodage acoustico-phonétique de la parole continue.** XVIIèmeJEP., pp. 30-35, Nancy, 1988.

[Bourlard et Wellekens 88] H. BOURLARD et C.J. WELLEKENS : **Links between Markov Models and Multi-Layer Perceptrons.** Technical Report 88M263, Philips, Septembre 1988.

[Bourlard et Wellekens 89] H. BOURLARD et C.J. WELLEKENS : **Speech Pattern Discrimination and Multilayer-Perceptrons**. Computer, Speech and Language, Vol. 3, 1989.

[Bulot et Méloni 88] R. BULOT et H. MELONI : **Le décodage acoustico-phonétique au GIA : identification ascendante des voyelles.** XVIIèmeJEP., pp. 151-155, Nancy, 1988.

[Caelen et al. 88] J. CAELEN, H. TATTEGRAIN, H. MELONI, R. BULOT, G. MERCIER et A. BONNEAU : **Une base de règles pour le décodage acoustico-phonétique: le cas des occlusives sourdes.** Actes du séminaire "acoustico-phonétique"., pp 51-63. Nancy, 1988.

[Calliope 89] Collectif CALLIOPE : **La parole et son traitement automatique.** Masson, 1989.

[Caraty 87] M.-J. CARATY : **Contribution au décodage acoustico-phonétique : études de distance inter-spectres et reconnaissance de cycles vocaliques**. Thèse de doctorat (PhD), Université Paris VI, 1987.

[Carbonell et al. 84] N. CARBONELL, D. FOHR, J.P. HATON, F. LONCHAMP et J.-M. PIERREL : **An Expert System for the Automatic Reading of French Spectrograms.** IEEE-ICASSP, 42-8, San Diego, California, USA., 1984.

[Cervantes et Sérignat 86] O. CERVANTES et J.-F. SERIGNAT : **Définition et réalisation d'une base de données des sons du français.** XVèmeJEP., pp. 213-216, Aix en Provence, 1986.

[Cheng et O'Shaugnessy 90] Y.-M. CHENG et D. O'SHAUGHNESSY : **Efficacité de la prédiction linéaire de vecteurs dans le codage de la parole à très bas débit.** XVIIIèmeJEP., pp. 296-300, Montréal, 1990.

[Chevalier et al. 85] M.-C. CHEVALIER, G. CHOLLET et Y. GRENIER : **Speech Analysis and Restitution using Time-Dependant Autoregressive Models.** IEEE-ICASSP., pp. 501-504, Tampa, Floride, USA, 1985.

[Chollet et al. 86] G. CHOLLET, Y. GRENIER et S.-M. MARCUS : **Temporal Decomposition and Non-Stationary Modeling of Speech.** Eurasip, pp. 365-368, 1986.

[Choquet 73] G. CHOQUET : **Cours d'analyse Tome II, Topologie.** Masson, Paris, 1973.

[Choukri 87] K. CHOUKRI : **Several Approaches to Speaker Adaptation in Automatic Speech Recognition Systems.** Thèse de doctorat, ENST, Paris, France, 1987.

[Class et al. 90] F. CLASS, A. KALTENMEIER, P. REGEL et K. TROTTLER : **Fast Speaker Adaptation for Speech Recognition Systems.** IEEE-ICASSP, pp. 133-136, Albuquerque, New-Mexico, USA., 1990.

[Cordeau 88] J.-P. CORDEAU : **Analyse de quelques métriques pour la reconnaissance automatique des mots connectés**. Technical report ENST. 88D017, Paris, France, 1988.

[Cover et Hart 67] T. COVER et P.-E. HART : **Nearest Neighbour Pattern Classification.** IEEE Trans. Inf. Theory., 1967.

[Delannoy 80] C. DELANNOY : **Un algorithme rapide de recherche des plus proches voisins.** RAIRO Informatique, n°14, mars 1980.

[Deléglise 91] P. DELEGLISE : **Une architecture logicielle pour le décodage acoustico-phonétique, Applications à la détection d'événements phonétiques.** Thèse d'état, à paraître

[Devijver et Kittler 82] P.-A. DEVIJVER et J. KITTLER : **Pattern Recognition, a Statistical Approach.** Prentice Hall, Inglewodd, N.J., USA., 1982.

[De Lima 89] A. DE LIMA : **Modèles non-stationnaires à deux niveaux**, Thèse de doctorat , ENST, Paris, France, 1989.

[De Mori et al. 83] R. DE MORI, A. GIORDANA et P. LAFACE : **Phonetic Feature Hypothesization in Continuous Speech.** IEEE-ICASSP, pp. 316-319, Boston, Massachusetts, USA., 1983.

[Dijk et Marcus 89] A.-M.-L. Van DIJK-KAPPERS et S.-M. MARCUS : **Temporal Decomposition of Speech.** Speech communication., Vol. 8, n°2, 1989.

[Dubuisson 90] B. DUBUISSON : **Diagnostic et reconnaissance des formes.** Traité des Nouvelles Technologies., Ed. Hermès, 1990.

[Duvaut 89] P. DUVAUT : **Le filtrage de Wiener linéaire-quadratique à horizon fini. Application à la prédiction.** Traitement du signal, Vol. 6, n°3, 1989.

[Eckart et Young 36] C. ECKART et G. YOUNG : **The Approximation of one Matrix by another of Lower Rank**. Psycometrika, Vol. 1, n° 3, September 1936.

[Fisher et al. 86] W.-M. FISHER, G.-R. DODDINGTON et K.-M. GOUDIE-MARSHALL : **The DARPA Speech Recognition Research Database : Specifications and Status.** Speech Recognition Workshop, 1986.

[Fisher et al. 87] W.-M. FISHER, V. ZUE, J. BERNSTEIN et D. PALETT : **An Acoustic-Phonetic Data Base.** 113th Meeting of Acoustic Society of America, Mai 1987.

[François et Fohr 90] D. FRANCOIS et D. FOHR : **Première évaluation d'APHODEX, système expert pour le décodage acoustico-phonétique de la parole continue.** XVIIIèmeJEP., pp. 191-195, Montréal, 1990.

[Franzini et al. 90] M. FRANZINI, K.-F. LEE et A. WAIBEL : **Connectionist Viterbi Training : A new Hybrid Method for continuous Speech Recognition**. IEEE-ICASSP., pp. 425-428, Albuquerque, New Mexico, USA., 1990.

[Frieman et al. 75] J.-H. FRIEMAN, F. BASKETT et L.-J. SHUTEK : **An Algorithm for Finding the Nearest Neighbours.** IEEE Trans on Computers, 1975.

[Funahashi 89] K.-I. FUNAHASHI : **On the approximate realization of continuous mappings by neural network**. Neural Networks, Vol 2, pp. 193-192, Mars 1989.

[Furui 80] FURUI : **A Training Procedure for Isolated Word Recognition.** IEEE Trans ASSP. Vol. 28, n° 2, pp. 129-136, Avril 1980.

[Gaillat 81] G. GAILLAT : **Méthodes statistiques en reconnaissance des formes**. Cours de l'ENSTA, Paris, 1981.

[Galas et Rodet 90] T. GALAS et X. RODET : **An Improved Cepstral Method for Source-Filter System Deconvolution : Application to Musical Sounds Signals**. ICMC, Glasgow, Septembre 1990.

[Gauvain 82] J.-L. GAUVAIN : **Reconnaissance de mots enchaînés et détection de mots dans la parole continue**. Thèse de 3ème cycle, Paris-Sud, 1982.

[Gauvain et Mariani 82] J.-L. GAUVAIN et J.-J. MARIANI : **A Method for Connected Word Recognition and Word Spotting on a Microprocessor.** IEEE ICASSP, pp. 891-894, 1982.

[Goatcher et Mason 86] GOATCHER et MASON : **An Adaptative Approch to a Speaker-Independent Isolated System with Short Training.** International Conference on Speech Input/Output Techniques and Applications, Londres, pp. 67-70.

[Gottsried et Weisman 73] B.S. GOTTSRIED et J. WEISMAN : **Introduction to Optimisation Series.** Prentice Hall, 1973.

[Gray 84] R.-M. GRAY : **Vector Quantization**. IEEE ASSP MAGAZINE,Vol.1, n°2. April 1984.

[Gray et al. 80] R.-M. GRAY, A. BUZO, A.-H. GRAY et Y. MATSUYAMA : **Distorsion Measures for Speech Processing.** IEEE Trans. ASSP, Vol. 28, pp. 367-376, 1980.

[Grenier 77] Y. GRENIER : **Identification du locuteur et adaptation au locuteur d'un système de reconnaissance phonémique.** Thèse de docteur-ingénieur, ENST, PARIS, France, 1977.

[Grenier 80] Y. GRENIER : **Utilisation de la prédiction linéaire en reconnaissance et adaptation au locuteur**. XIèmeJEP., pp. 163-171, Strasbourg, 1980.

[Grenier 84] Y. GRENIER : **Modélisation de signaux non-stationnaires**. Thèse d'état, Université de Paris-Sud, France, 1984.

[Halle 83] P. HALLE : **Techniques cepstrales amélorées pour l'extraction d'enveloppe spectrale et la détection du pitch.** Séminaire "Analyse de la parole", GRECO "Communication Parlée", ENST, 1983.

[Hart 68] P.-E. HART : **The Condensed Nearest Neighbour Rule**. IEEE Trans IT., Vol. 14, 1968.

[Hataoka et Waibel 89] N. HATAOKA et A.-H. WAIBEL : **Speaker-Independant Phoneme Recognition on TIMIT Database Using Integrated Time-Delay Neural Networks (TDNNs).** Technical Report CS-89-190, CMU, Pittsburgh, 1989.

[Haton et Pierrel 78] J.-P. HATON et J.M. PIERREL : **Data Structure and Organization of MIRTYLLE II System.** IJCPR, Kyoto, Japan, Novembre 1978.

[Hess 83] W. HESS : **Pitch Determination of Speech Signal : Algorithms and Devices.** Springer-Verlag, 1983.

[Hornik 89] HORNIK : **Multilayer Feedworward Networks are Universal Approximators**. Neural Network, Vol. 5, 1989.

[Imai et Abe 81] S. IMAI et Y. ABE : **Spectral Envelope Extraction by Improved Cepstral Method.** Electron. and commun., In Japan, Vol 62A, n°4, pp. 10-17, 1979.

[Itakura 75] F. ITAKURA : **Minimum Prediction Residual Principle Applied to Speech Recognition.** IEEE Trans. ASSP, Vol. 23, pp. 67-72, 1975.

[Jelinek et al. 75] F. JELINEK, L.-R. BAHL et R.-L. MERCER : **Design of a Linguistic Statistical Decoder for the Recognition of Continuous Speech**. IEEE Trans IT., Vol. 21., 1975.

[Jouvet 88] D. JOUVET : **Reconnaissance de mots connectés indépendamment du locuteur par des méthodes statistiques**, Thèse de doctorat, ENST, Paris, France, 1988.

[Klatt 77] D.H. KLATT : **Review of ARPA speech understanding project.** JASA, Vol. 62, pp. 1345-1356, 1977.

[Klatt 82] D.H. KLATT : **Prediction of Perceived Phonetic Distance from Critical-Band Spectra : a First Step.**, IEEE-ICASSP., pp. 1278-1281, Paris, France, 1982.

[Kohonen 84] T. KOHONEN : **Self-Organization and Associative Memory.** Springer, 1984.

[Kubala et al. 88] F. KUBALA, Y. CHOW, A. DERR, M. FENG, O. KIMBALL, J. MAKHOUL, P. PRICE, J. ROHLICEK, S. ROUCOS, R. SCHWARTZ et J. VANDEGRIFT : **Continuous Speech Recognition Results of the BYBLOS System on the Darpa 1000-Word Resource Management Database.** IEEE-ICASSP, pp. 291-294, New York, USA., 1988.

[Lapedes et Farber 87] A. LAPEDES et R. FARBER : **Non-Linear Signal Processing using Neural Network : Prediction and System Modeling.** Technical Report., Los Alamos Laboratory, New Mexico, USA., Juillet 1987.

[Lea 82] W. LEA : **Available Speech Databases for Evaluating Speech Recognizers.** Workshop on Standardization for Speech I-O Technology, Washington, 1982.

[Lee 88] K.-F. LEE : **The Development of the SPHINX System,** Kluwers Academic Publishers, Norwell, Massachussets, USA., 1988.

[Le Cun 85] Y. LE CUN : **Une procédure d'apprentissage pour réseau à seuil asymétrique** : COGNITIVA, Cesta-Afcet, pp. 599-604, Paris., 1985.

[Le Cun 87] Y. LE CUN : **Modèles connexionnistes de l'apprentissage**, Thèse de doctorat , Université Paris VI, Paris, France, 1987.

[Lee et Hon 89] K.-F. LEE et H.-W. HON : **Speaker-Independant Phone Recognition using Hidden Markov Models.** IEEE Trans ASSP, Vol. 37 n°11, pp. 1641-1649, 1989.

[Lelong-Ferrand 63] J. LELONG-FERRAND : **Géométrie différentielle.** Masson, Paris, 1963.

[Liénard et al. 77] J.S. LIENARD, J.-J. MARIANI et G. RENARD : **Intelligibilité de phrases phonétiquement altérées : application à la transmission de la parole.** 9ème ICA, Madrid, 1977.

[Liporace 75] L.-A. LIPORACE : **Linear Estimation of Non-Stationary Signals**. JASA, Vol. 58, n°6, pp. 1288-1295, 1975.

[Lippman 87] R.-P. LIPPMAN : **Computing with Neural Nets**. IEEE ASSP MAGAZINE, Vol. 4, n°2, 1987.

[Lowerre 76] B.-T. LOWERRE : **The Harpy Speech Recognition System.** Ph.D thesis, Dept of Computer Science, Carnegie-Mellon University, Pittsburg, USA, 1976.

[Mailles 80] J.-P. MAILLES : **Analyse des tableaux de proximités**. Cahier n° 33, Institut de Statistique, Université Pierre et Marie Curie, France, 1980.

[Mansour et Juang 89] D. MANSOUR et B.-H. JUANG : **A Family Measure Based upon Projection Operation for Robust Speech Recognition.** IEEE Trans ASSP, Vol. 37 n°11, pp. 1659-1670, 1989.

[Marcus et Lieshout 84] S.-M. MARCUS et R.-A.-J.-M. Van LIESHOUT : **Temporal Decomposition of Speech.** IPO annual progress report 19., pp. 25-31, 1984.

[Mariani 82] J.-J. MARIANI : **ESOPE : Une système de compréhension de la parole continue.** Thèse d'état, université Pierre et Marie Curie ,France, 1982.

[Mariani 89] J.-J. MARIANI : **Recents Advances in Speech Processing.** IEEE-ICASSP, pp. 429-440, Glasgow, Scotland., 1989.

[Markel et Gray 75] J.-D. MARKEL et J.-D. GRAY : **Linear Prediction of Speech**, Springer Verlag, New-York, 1975.

[Marteau 88] P.-F. MARTEAU : **Cibles et trajectoires acoustiques : application à la segmentation et à l'étiquetage automatique du signal de parole**. Thèse de doctorat, ICP Grenoble, France, 1988.

[Mc Dermott et Katagiri 89] E. Mc DERMOTT et S. KATAGIRI : **Shift-Invariant, Multi-Category Phoneme Recognition Using Kohonen's LVQ2**. IEEE-ICASSP, pp.81-84, 1989.

[Méloni 82] H. MELONI : **Etude et réalisation d'un système de reconnaissance automatique de la parole continue.** Thèse d'état, université d'Aix-Marseille III, Février 1982.

[Méloni et Bulot 87] H. MELONI et R. BULOT : **Signal Processing and Pattern Recognition in Prolog II for Acoustic-Phonetic Decoding of Continuous Speech.** Proc RFIA. pp 21-36, 1987.

[Mercier 69] G. MERCIER : **Reconnaissance des formes, approximation des fonctions de décision et application à la reconnaissance des phonèmes.** Thèse de 3ème cycle, Rennes, Juin 1969.

[Mercier et al. 88] G.MERCIER, A. COZANNET et J. VAISSIERE.: **Recognition of Speaker-Dependant Continuous Speech with KEAL-NEVEZH**. NATO ASI Series. Vol F46 pp 459-463., 1988.

[Miclet 84] L. MICLET : **Méthodes structurelles pour la reconnaissance des formes.** Eyrolles, Paris, 1984.

[Miclet et Nehme 79] L. MICLET et A. NEHME : **Expériences en reconnaissance de la parole par prédiction linéaire.** AFCET-RFIA, Toulouse, 1979.

[Mokeddem 85] A. MOKEDDEM : **Analyse de la parole : reconnaissance multilocuteur de mots isolés pour les systèmes miniaturisés**, Thèse de doctorat , Université de Neuchâtel, Suisse, 1985.

[Monné 88] J. MONNE : **Description des logiciels d'apprentissage et d'exploitation sur PC développé pour la carte RDP50.** CNET LAA/TSS/RCP, Lannion, Novembre 1988.

[Montacié et al. 89] C. MONTACIE, K. CHOUKRI et G. CHOLLET : **Speech Recognition Using Temporal Décomposition and Multi-Layer Feed-Forward Automata.** IEEE-ICASSP.,pp. 409-412 1989.

[Montacié et Barral 89] C. MONTACIE et H. BARRAL : **Réseaux neuromimétiques**. Rapport de contrat EDF, ENST, 1989.

[Montacié et Chollet 87] C. MONTACIE et G. CHOLLET : **Systèmes de référence pour l'évaluation d'applications et la caractérisation de bases de données en reconnaissance automatique de la parole.** XVIèmeJEP., pp. 323-326, Hammamet, 1987.

[Niles et Silverman 89] L. NILES et H.-F. SILVERMAN : **How Limited Training Data can allow a Neural Network to Out-Perform an 'Optimal' Statistical Classifier.** IEEE-ICASSP. pp. 17-20, 1989.

[Niranjan et Fallside 87] M. NIRANJAN et F. FALLSIDE : **On Modelling the Dynamics of Speech Patterns.** ECSP,Vol. 1, pp. 71-74, Edinburgh, 1987.

[Nocerino et al. 85] N. NOCERINO, F.-K. SOONG, L.R. RABINER et D.-H. KLATT : **Comparative Study of Several Distorsion Measures for Speech Recognition.** IEEE, Speech Communication, Vol. 4, pp. 25-28, 1985.

[Noll 67] A.-M. NOLL : **Cepstrum Pitch Determination.** IEEE Trans AU. Vol. 16, pp. 239-309, 1967.

[Olson et Belar 56] H.-F. OLSON et H. BELAR : **Phonetic Typewriter.** JASA, Vol. 28, pp. 1072-1081, 1956.

[Pau 71] L.-F. PAU : **Méthodes statistiques de réduction et de reconnaissance des formes - Normalisations des paramètres phonétiques - Application à la reconnaissance de la voix.** Thèse de docteur-ingénieur, Université de Paris-Sud, France, 1971.

[Petitot-Cocorda 84] J. PETITOT-COCORDA : **Les catastrophes de la parole**. Maloine, Paris, 1984.

[Petrowski et al. 89] A. PETROWSKI, G. DREYFUS et C. GIRAULT : **Implantation de réseaux de neurones formels sur une architecture multiprocesseurs.** Technical report MASI. n° 284, Université Paris VI, France, 1989.

[Picone 90] J. PICONE : **Continuous Speech Recognition Using Hidden Markov Models**.IEEE ASSP MAGAZINE, Vol. 7, n° 3, pp. 26-40, Juillet 1990.

[Poirier 90] F. POIRIER : **Reconnaissance multilocuteur de voyelles par un réseau connexioniste auto-organisateur.** XVIIIèmeJEP., pp. 196-200, Montréal, 1990.

[Potage 86] J. POTAGE : **La reconnaissance de la parole**. Cours de l'Ecole Supérieure d'Electricité, 1986.

[Rabiner et Juang 86] L.-R. RABINER et B.-H. JUANG : **An Introduction to Hidden Markov Models**.IEEE ASSP MAGAZINE, Janvier 1986.

[Reddy et al. 77] D.R. REDDY et al.: **Speech Understanding Systems Final Report.** Dept of Computer Science, Carnegie-Mellon University, Pittsburg, USA, 1977.

[Shikano et al. 86] K. SHIKANO, K.-F. LEE et R. REDDY : **Speaker Adaptation Through Vector Quantization.** IEEE-ICASSP, 49-5., Tokyo, Japan, 1986.

[Stern et al. 86] P.-E. STERN, M. ESKENAZI et D. MEMMI : **An Expert System for Speech Spectrogram Reading.** IEEE-ICASSP, 23-1., Tokyo, Japan, 1986.

[Sugiyama 90] M. SUGIYAMA : **Spectral Interpolation Using Distorsion Geodesic Lines.** ICSLP, pp. 477-480, Kobe, Japan, 1990.

[Taillanter 88] S. TAILLANTER : **Constitution d'un dictionnaire de cibles et utilisation pour l'inférence en décodage acoustico-phonétique**. Technical Report ENST. 88H003. Paris, France, 1988.

[Tohkura 86] Y. TOHKURA : **A Weighted Cepstral Measure for Speech Recognition** IEEE-ICASSP, pp. 761-764., Tokyo, Japan, 1986.

[Tubach et al. 90] J.-P. TUBACH, G. CHOLLET, K. CHOUKRI, C. MONTACIE, C. MOKBEL et H. VALBRET : **Adaptation au locuteur de systèmes de reconnaissance. Régression linéaire multiple et perceptrons multi-couches.** Numéro spécial "reconnaissance de la parole", revue Traitement du Signal, Vol. 7, n°4, pp. 285-292, 1990.

[Vicens 69] P. VICENS : **Aspects of Speech Recognition by Computer.** thèse de doctorat (PhD), Université de Standford, USA, 1969.

[Vintsjuk 68] T.-K. VINTSJUK : **Speech Discrimination by Dynamic Programming** . Traduit dans Kibernetica. Vol. 4, pp. 81-88, 1968.

[Viterbi 67] A.J. VITERBI : **Errors Bounds for Convolutional Codes and an Asymptotically Optimum Decoding Algorithm.** IEEE Trans on Information Theory, Vol. 13 n°2, pp. 260-269, Avril 1967.

[Wagner et Fisher 74] R.-A. WAGNER et M.-J. FISHER : **The String to String Correction Problem.** JACM., 1974.

[Waibel et al. 89] A. WAIBEL, T. HANAZAWA, G. HINTON, K. SHIKANO et K. LANG : **Phoneme Recognition Using Time-Delay Networks.** IEEE Trans ASSP, 1989.

[Waibel et Lee 89] A. WAIBEL et K.-F. LEE : **Readings in Speech Recognition.** Morgan Kaufman Publishers, Inc. San Mateo, California, 1989.

[Whittle 63] P. WHITTLE : **On the Fitting of Multivariate Autoregression and the Approximate Canonical Factorization of a Spectral Density Matrix**. Biometrika, Vol. 50, pp. 129-134, 1963.

[Widrow et Hopf 60] B. WIDROW et M.-E. HOPF : **Adaptive Switching Circuits.** IRE WESCON Conv. record, part. 4, pp. 96-104, 1960.

[Wiswanathan et Makhoul 75] R. WISWANATHAN et J. MAKHOUL : **Quantization Properties of Transmission Parameters in Linear Predictive Systems.** IEEE Trans ASSP. Vol. 23, pp. 309-321, 1975.

[Witbrock et Zagha 89] M. WITBROCK et M. ZAGHA : **An Implementation of Back-Propagation Learning on GF11, a large SIMD Parallel Computer**. Technical Report CS-89-208, CMU, Pittsburgh, 1989.

[Woods et al. 76] W. WOODS, M. BATES, G. BROWN, B. BRUCE, C. COOK, J. KLOVSTAD, J. MAKHOUL, B. NASH-WEBBER, R. SCHWARTZ, J. WOLF et V. ZUE : **Speech Understanding System: Final Technical Report, Bolk Beranek and Newman, Inc**. Technical Report 3428, Cambridge, 1976.

[Zue et Lamel 86] V.-W. ZUE et L.-F. LAMEL : **An Expert Spectrogram Reader : A Knowledge-Based Approach to Speech Recognition.** IEEE-ICASSP, pp. 1197-1200., Tokyo, Japan, 1986.

[Zuk 84] E.-A. ZUK : **An Investigation of Temporal Decomposition of Speech Parameters for Automatic Segmentation of Speech.** IPO annual progress report. n° 459., 1984.

[Zwicker 77] E. ZWICKER : **Masking Period Patterns and Hearing Theories in Psychophysics and Physiology of Hearing**. Ed. E.F. EVANS & J.P. WILSON, New york, 1977.

ANNEXE I. DISTANCE D'ITAKURA-SAITO (LOGARITHME DU RAPPORT DE VRAISEMBLANCE)

Comme nous l'avons vu au chapitre 3, la prédiction linéaire représente le signal de parole, sur une fenêtre $\{s_n\}$, (n = 1 ,.., T), par l'équation récurrente suivante :

$$s_n = \sum_{i=1}^{p} a_i \, s_{n-i} + \varepsilon_n \qquad \textbf{(I.01)}$$

Les ε_n sont supposés non corrélés et de moyenne nulle. Si l'on considère que les $\{\varepsilon_n\}$ sont représentés par une variable aléatoire gaussienne, la densité de cette variable aléatoire est :

$$\frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-\varepsilon_n^2}{\sigma^2}\right) \qquad \textbf{(I.02)}$$

Itakura et Saito ont démontré que la densité de probabilité conditionnelle P des variables aléatoires s_1, s_2, .., s_T connaissant σ et les a_i peut être approchée par :

$$P(s_1, s_2, .., s_T / \sigma, a_1, a_2, .., a_p) \approx (2\pi \sigma^2) \exp\left(\frac{-\alpha}{2\sigma^2}\right) \qquad \textbf{(I.03)}$$

$$\text{avec } \alpha = \sum_{n=1}^{T} \left(s_n - \sum_{i=1}^{p} a_i \, s_{n-i}\right)^2 \qquad \textbf{(I.04)}$$

et $s_n = 0$ en dehors de l'intervalle [1, T].

On cherche alors les valeurs de σ et des a_i qui rendent maximale cette probabilité en considérant que les variables σ et a_i sont libres. En calculant les dérivées partielles de (I.03) par rapport à σ et aux a_i, on aboutit au système linéaire classique de type Toeplitz. Le maximum de la densité de probabilité P_{max} est alors donné par les valeurs maximales des a_i et de σ :

$$P_{max} = \frac{2\pi}{T} \exp(-T/2) \, \alpha^{-T/2} \qquad \textbf{(I.05)}$$

La distance d'Itakura et Saito consiste à comparer le maximum de la densité de probabilité P_{ts} d'une fenêtre du signal de test S_{ts} au maximum de la densité de probabilité P_{rf} d'une fenêtre du signal de référence S_{rf}, en considérant que σ est libre et que les a_i sont ceux du signal de référence. La distance d'Itakura et Saito est alors définie comme le logarithme d'un rapport de vraisemblance qui est le rapport des deux densités de probabilité P_{ts} et P_{rf}:

$$D_{IS}(S_{ts},S_{rf}) = \text{Log}\left(\frac{P_{max}(S_{ts1},S_{ts2},..,S_{tsT} / \sigma_{ts},a_{rf1},a_{rf2},..,a_{rfp})}{P_{max}(S_{rf1},S_{rf2},..,S_{réfT} / \sigma_{rf},a_{rf1},a_{rf2},..,a_{rfp})}\right) \quad \textbf{(I.06)}$$

$$= \text{Log}\ (\alpha_{ts} / \alpha_{rf}), \quad \textbf{(I.07)}$$

avec $\alpha_{rf} = \beta$ (cf. formule 3.05).

Cocker et Boll ont donné de la distance d'Itakura-Saito une interprétation spectrale. Si l'on considère les fonctions de transfert F_{ts} et F_{rf} des filtres inverses calculés pour les signaux des fenêtres d'analyse S_{ts} et S_{rf}, on a :

$$\alpha_{ts} / \alpha_{rf} = 1 + \frac{T_{ech}}{2\pi} \int_{-\pi/T_{ech}}^{+\pi/T_{ech}} \frac{| F_{ts}(\omega) - F_{rf}(\omega) |^2}{| F_{rf}(\omega) |^2}\ d\omega, \quad \textbf{(I.08)}$$

avec T_{ech} la période d'échantillonnage

Le numérateur montre bien que la distance dépend de l'écart entre les deux spectres. On remarque que cet écart est pondéré par $1/| F_{rf}(\omega) |^2$. Donc les écarts entre les deux spectres, pour les valeurs de ω avoisinant la fréquence centrale d'une masse spectrale de la fenêtre d'analyse à reconnaître S_{rf}, seront pondérés plus fortement que les écarts pour des valeurs de ω choisies entre les masses spectrales.

Cette mesure de dissimilarité est donc spécifiquement adaptée à la comparaison de masses spectrales.

ANNEXE II. RÉSULTATS COMPARATIFS DU SYSTÈME DE RÉFÉRENCE

Désignation des Modèles	Nombre d'états	Nombre de transitions	Nombre de distributions	Nom du modèle (RDP50)
A	1	1	1	M1_1_1
B	2	3	2	M2_3_2
C	2	3	3	M2_3_3
D	3	6	3	M3_6_3
E	3	6	6	M3_6_6
F	4	9	4	M4_9_4
G	4	9	9	M4_9_9
H	5	12	5	M5_12_5
I	5	12	12	M5_12_12

Tableau II.1. Topologie des différents modèles employés

Modèles	Trouvés	Substitués	Omis	Insérés
A	48%	44%	9%	35%
B	**58%**	**35%**	**7%**	**23%**
C	52%	35%	13%	13%
D	50%	38%	12%	18%
E	51%	35%	14%	14%
F	51%	31%	18%	9%
G	49%	31%	20%	8%
H	41%	33%	26%	4%
I	32%	34%	35%	3%

Tableau II.2. Reconnaissance des Mots épelés FB (Apprentissage Diphones FB)

Modèles	Trouvés	Substitués	Omis	Insérés
A	42%	49%	9%	39%
B	61%	35%	4%	32%
C	**63%**	**33%**	**4%**	**40%**
D	47%	45%	8%	31%
E	52%	38%	10%	29%
F	62%	33%	5%	23%
G	55%	39%	6%	19%
H	49%	37%	13%	11%
I	47%	45%	8%	20%

Tableau II.3. Reconnaissance des Mots épelés MJC (Apprentissage Diphones MJC)

Modèles	Trouvés	Substitués	Omis	Insérés
A	62%	31%	6%	7%
B	**80%**	**19%**	**1%**	**8%**
C	73%	25%	2%	7%
D	77%	22%	1%	7%
E	77%	20%	3%	17%
F	76%	20%	4%	3%
G	**80%**	**19%**	**2%**	**3%**
H	74%	22%	4%	4%
I	77%	19%	3%	3%

Tableau II.4. Reconnaissance des Diphones FB (Apprentissage Diphones FB)

Modèles	Trouvés	Substitués	Omis	Insérés
A	54%	41%	5%	14%
B	64%	34%	2%	13%
C	70%	28%	2%	18%
D	61%	37%	2%	11%
E	74%	22%	4%	8%
F	70%	28%	3%	1%
G	**79%**	**19%**	**2%**	**2%**
H	67%	30%	4%	2%
I	73%	26%	1%	11%

Tableau II.5. Reconnaissance des Diphones MJC (Apprentissage Diphones MJC)

Modèles	Trouvés	Substitués	Omis	Insérés
A	48%	44%	8%	9%
B	53%	38%	9%	8%
C	50%	42%	8%	10%
D	**63%**	**31%**	**6%**	**21%**
E	46%	46%	7%	12%
F	54%	42%	4%	10%
G	43%	54%	3%	14%
H	42%	48%	9%	2%
I	42%	52%	5%	6%

Tableau II.6. Reconnaissance des Lettres SL (Apprentissage Lettres SL)

Modèles	Trouvés	Substitués	Omis	Insérés
A	52%	44%	4%	11%
B	63%	36%	2%	15%
C	**75%**	**20%**	**5%**	**16%**
D	72%	23%	5%	11%
E	68%	30%	2%	14%
F	64%	28%	7%	7%
G	57%	41%	1%	22%
H	66%	25%	9%	8%
I	64%	30%	6%	9%

Tableau II.7. Reconnaissance des Lettres FB (Apprentissage Lettres FB)

Modèles	Trouvés	Substitués	Omis	Insérés
A	28%	60%	12%	14%
B	**63%**	**30%**	**7%**	**18%**
C	52%	27%	21%	10%
D	47%	40%	13%	10%
E	51%	35%	13%	16%
F	43%	42%	15%	7%
G	59%	34%	7%	18%
H	40%	35%	24%	2%
I	43%	41%	16%	5%

Tableau II.8. Reconnaissance des Mots épelés FB (Apprentissage Diphones FB et MJC)

Modèles	Trouvés	Substitués	Omis	Insérés
A	38%	54%	8%	37%
B	**66%**	**31%**	**3%**	**39%**
C	58%	31%	12%	27%
D	47%	46%	7%	26%
E	51%	43%	7%	33%
F	49%	44%	7%	20%
G	62%	33%	5%	26%
H	45%	41%	14%	14%
I	46%	46%	8%	17%

Tableau II.9. Reconnaissance des Mots épelés MJC (Apprentissage Diphones FB et MJC)

Modèles	Trouvés	Substitués	Omis	Insérés
A	22%	42%	36%	6%
B	23%	56%	21%	9%
C	**34%**	**55%**	**11%**	**22%**
D	26%	58%	16%	12%
E	28%	61%	11%	16%
F	**34%**	**46%**	**20%**	**11%**
G	30%	43%	28%	6%
H	26%	39%	35%	3%
I	28%	60%	12%	16%

Tableau II.10. Reconnaissance des Mots épelés FB (Apprentissage Diphones MJC)

Modèles	Trouvés	Substitués	Omis	Insérés
A	25%	58%	18%	14%
B	**33%**	**59%**	**8%**	**32%**
C	24%	66%	10%	17%
D	31%	54%	15%	20%
E	21%	66%	13%	12%
F	30%	54%	16%	10%
G	22%	53%	24%	9%
H	24%	53%	24%	8%
I	16%	61%	23%	3%

Tableau II.11. Reconnaissance des Mots épelés MJC (Apprentissage Diphones FB)

ANNEXE III. SYNTAXE ET MODÈLES DU SYSTÈME DE RÉFÉRENCE

Cette annexe donne les "listings" correspondant aux différentes modélisations utilisées par notre système de référence. Dans les listings, les mots clés retenus par le programme de compilation des machines de Markov cachées sont indiqués en gras. Les commentaires sont mis en italique.

Pour toute méthode d'estimation, il est nécessaire d'avoir un ensemble d'apprentissage suffisant. Le calcul de chaque distribution consiste en l'estimation d'un vecteur moyenne et d'un écart-type. Quelques dizaines de vecteurs d'analyse suffisent pour cela. La taille du corpus d'apprentissage est d'environ dix mille trames par locuteur et cela suffit pour apprendre une dizaine de distributions par son.

```
!
! Modèle de Markov pour la reconnaissance de séquence de mots épelés
! on utilise un modèle par son
!=========================================================
==============
! Description de la syntaxe:
Syntaxe {
  Productions {
      SEQ_1 = SILDEB . Ph;
      SEQ_2 = SILDEB . Ph . Ph;
      SEQ_3 = SILDEB . Ph . Ph . Ph;
      SEQ_4 = SILDEB . Ph . Ph . Ph .Ph;
      SEQ_5 = SILDEB . Ph . Ph . Ph . Ph . Ph;
      SEQ_6 = SILDEB . Ph . Ph . Ph . Ph . Ph . Ph;
      SEQ_7 = SILDEB . Ph . Ph . Ph . Ph . Ph . Ph . Ph;
      SEQ_8 = SILDEB . Ph . Ph . Ph . Ph . Ph . Ph . Ph . Ph;
      SEQ_9 = SILDEB . Ph . Ph . Ph . Ph . Ph . Ph . Ph . Ph . Ph;
      SEQ_10 = SILDEB . Ph10;
      SEQ_11 = SILDEB . Ph . Ph10;
      SEQ_12 = SILDEB . Ph . Ph . Ph10;
      SEQ_13 = SILDEB . Ph . Ph . Ph . Ph10;
      SEQ_14 = SILDEB . Ph . Ph . Ph .Ph . Ph10;
      SEQ_15 = SILDEB . Ph . Ph . Ph . Ph . Ph . Ph10;
      SEQ_16 = SILDEB . Ph . Ph . Ph . Ph . Ph . Ph . Ph10;
```

```
        SEQ_17 = SILDEB . Ph . Ph . Ph . Ph . Ph . Ph . Ph . Ph10;
        SEQ_18 = SILDEB . Ph . Ph . Ph . Ph . Ph . Ph . Ph . Ph . Ph10;
        SEQ_19 = SILDEB . Ph . Ph . Ph . Ph . Ph . Ph . Ph . Ph . Ph . Ph10;
        SEQ_20 = SILDEB . Ph . Ph . Ph . Ph . Ph . Ph . Ph . Ph . Ph .Ph . Ph10;
        Ph10 = Ph . Ph . Ph . Ph . Ph . Ph . Ph . Ph . Ph . Ph;
        Ph = (Voyl + Cons) . (SILFIN + ());
        Voyl = a + et + e + ai + i + o + y + u;
      Cons = b + s + d + f + j + h + k + l + m + n + p + r + t + v  + g + z;
   }
   Substitutions  Mots {
      SILDEB  = "" ;     ! "" représente le nom vide
      SILFIN  = "" ;
        }
}
!
!============================================================
==============
!  Description acoustique
!
Acoustique {
   Modele MSIL {
      ! Modèle pour les silences : 3 états, 6 transitions, 3 fonctions
      Nombre Etats = 3 ;
      Nombre  Transitions = 6 ;
      Nombre  Fonctions = 3 ;
      Transitions {   ! description des transitions
            <1,1,1> ; <1,2,1> ; <1,3,2> ;
            <2,2,2> ; <2,3,3> ;
            <3,3,3> ;
      }
   }
   Modele M1_1_1 {
      Nombre Etats = 1 ;
      Nombre  Transitions = 1 ;
      Nombre  Fonctions = 1 ;
```

```
   Transitions {   ! description des transitions
         <1,1,1> ;
   }
}
Modele M2_3_2 {
   Nombre Etats = 2 ;
   Nombre Transitions = 3 ;
   Nombre Fonctions = 2 ;
   Transitions {   ! description des transitions
         <1,1,1> ; <1,2,2> ;
         <2,2,2> ;
   }
}
Modele M2_3_3 {
   Nombre Etats = 2 ;
   Nombre Transitions = 3 ;
   Nombre Fonctions = 3 ;
   Transitions {   ! description des transitions
         <1,1,1> ; <1,2,2> ;
         <2,2,3> ;
   }
}
Modele M3_6_3 {
   Nombre Etats = 3 ;
   Nombre Transitions = 6 ;
   Nombre Fonctions = 3 ;
   Transitions {   ! description des transitions
         <1,1,1> ; <1,2,2> ; <1,3,3> ;
         <2,2,2> ; <2,3,3> ;
         <3,3,3> ;
   }
}
Modele M3_6_6 {
   Nombre Etats = 3 ;
   Nombre Transitions = 6 ;
```

```
   Nombre Fonctions = 6 ;
   Transitions {   ! description des transitions
         <1,1,1> ; <1,2,2> ; <1,3,3> ;
         <2,2,4> ; <2,3,5> ;
         <3,3,6> ;
   }
}
Modele M4_9_4 {
   Nombre Etats = 4 ;
   Nombre Transitions = 9 ;
   Nombre Fonctions = 4 ;
   Transitions {   ! description des transitions
         <1,1,1> ; <1,2,2> ; <1,3,3> ;
         <2,2,2> ; <2,3,3> ; <2,4,4> ;
         <3,3,3> ; <3,4,4> ;
         <4,4,4> ;
   }
}
Modele M4_9_9 {
   Nombre Etats = 4 ;
   Nombre Transitions = 9 ;
   Nombre Fonctions = 9 ;
   Transitions {   ! description des transitions
         <1,1,1> ; <1,2,2> ; <1,3,3> ;
         <2,2,4> ; <2,3,5> ; <2,4,6> ;
         <3,3,7> ; <3,4,8> ;
         <4,4,9> ;
   }
}
Modele M5_12_5 {
   Nombre Etats = 5 ;
   Nombre Transitions = 12 ;
   Nombre Fonctions = 5 ;
   Transitions {   ! description des transitions
         <1,1,1> ; <1,2,2> ; <1,3,3> ;
```

```
            <2,2,2> ; <2,3,3> ; <2,4,4> ;
            <3,3,3> ; <3,4,4> ; <3,5,5> ;
            <4,4,4> ; <4,5,5> ;
            <5,5,5> ;
        }
    }
    Modele M5_12_12 {
        Nombre Etats = 5 ;
        Nombre Transitions = 12 ;
        Nombre Fonctions = 12 ;
        Transitions {   ! description des transitions
            <1,1,1> ; <1,2,2> ; <1,3,3> ;
            <2,2,4> ; <2,3,5> ; <2,4,6> ;
            <3,3,7> ; <3,4,8> ; <3,5,9> ;
            <4,4,10> ; <4,5,11> ;
            <5,5,12> ;
        }
    }
    ! On associe chaque unité à un type de modèle
    ! Les unités SILDEB et SILFIN sont associées au modèle MSIL
    ! Les unités des phonémes sont associées au modèle M*_*_*
    Modelisations {
        MSIL pour (SILDEB,SILFIN) ;
        M*_*_* pour (a,et,e,ai,i,o,y,u) ;
        M*_*_* pour (b,s,d,f,j,h,k,l,m,n,p,r,t,v,g,z) ;
    }
}
```

ANNEXE IV. CALCUL DES VECTEURS ET DES VALEURS PROPRES D'UNE MATRICE

Soit A la matrice à étudier, on suppose que ses vecteurs propres $\{u_i\}$ (i = 1 ,.., n), sont linéairement indépendants. Les valeurs propres associées $\{\lambda_i\}$ (i = 1 ,.., n), sont rangées par modules décroissants (i.e., $|\lambda_i| > |\lambda_j|$ si i > j).

MÉTHODE DE LA PUISSANCE

Les vecteurs propres étant linéairement indépendants, tout vecteur x peut s'écrire :

$$x = x_1\, u_1 + x_2\, u_2 + .. + x_n\, u_n.$$

ce qui conduit à :

$$A\, x = x_1\, A\, u_1 + x_2\, A\, u_2 + .. + x_n\, A\, u_n.$$

Soit :

$$A\, X = \lambda_1\, x_1\, u_1 + \lambda_2\, x_2\, u_2 + .. + \lambda_n\, x_n\, u_n.$$

Si on multiplie par A à gauche le résultat un nombre de fois égal à k-1 (et en supposant $x_1 \neq 0$)

$$A^k\, x = \lambda_1^k\, x_1\, u_1 + \lambda_2^k\, x_2\, u_2 + .. + \lambda_n^k\, x_n\, u_n.$$

$$A^k\, x = \lambda_1^k\, [\, x_1\, u_1 + (\frac{\lambda_2}{\lambda_1})^k\, x_2\, u_2 + .. + (\frac{\lambda_n}{\lambda_1})^k\, x_n\, u_n].$$

Comme $|\lambda_1| > |\lambda_i|$, il en résulte que $(\frac{\lambda_i}{\lambda_1})^k \to 0$ quand $k \to \infty$.

Donc :

$$\mathbf{A^k\, X \approx \lambda_1^k\, x_1\, u_1} \qquad \textbf{(IV.01)}$$

dès que k est assez grand.

Quand k tend vers l'infini, A^k x tend vers une direction fixe : celle du vecteur propre u_1 associé à la valeur propre de A de plus grand module λ_1. Le calcul de celle-ci est donné par la formule suivante :

$$\mathbf{\lambda_1 \approx \frac{||\, A^{k+1}\, X\, ||}{||\, A^k\, x\, ||}} \qquad \textbf{(IV.02)}$$

dès que k est assez grand.

MÉTHODE DE DÉFLATION

Pour déterminer les autres vecteurs propres $\{u_i\}$ (i = 2 ,.., n), une méthode dite de "déflation" est utilisée. Il s'agit de calculer à partir de la matrice A, du vecteur propre u_1 et de la valeur propre associée λ_1, une matrice B ayant n-1 vecteurs propres $\{u_i\}$ (i = 2 ,.., n). La méthode de Hotelling est utilisée pour cela.

Soit le vecteur propre u_1 associé de la matrice A associé à la valeur propre λ_1.
Le vecteur ligne défini par $v_1 A = \lambda_1 v_1$ est appelé vecteur propre ligne de A associé à la valeur propre λ_1.
On montre que :

$$\begin{cases} v_1 u_i = 0 & i \neq 1 \\ v_1 u_1 \neq 0 \end{cases}$$

Comme v_1 et u_1 sont définis à une constante multiplicative près, on les choisit pour avoir:

$$\begin{cases} v_1 u_i = 0 & i \neq 1 \\ v_1 u_1 = 1 \end{cases}$$

Dans ce cas

$$\mathbf{B = A - \lambda_1 u_1 v_1} \qquad \mathbf{(IV.03)}$$

La recherche de v_1 se fait aussi par la méthode de la puissance itérée en calculant :

$$\mathbf{y\ A^k \approx \lambda_1^k x_1\ v_1} \qquad \mathbf{(IV.04)}$$

UTILISATION POUR LA DÉCOMPOSITION TEMPORELLE

Cette méthode a été utilisée pour deux raisons :

— Sa simplicité (i.e., calculs de produits matriciels) permet de l'implanter rapidement sur des processeurs de traitement de signaux (e.g., TMS320C30) dans le but d'une utilisation temps réel.

— Son optimisation (i.e., nombre d'itérations), pour le calcul des fonctions d'interpolation de la décomposition temporelle, est simple. Les vecteurs propres

successifs cherchés diffèrent peu les uns des autres. L'algorithme de recherche est le suivant :

Etape 0: Soit à l'instant $t = t_0$, recherche de la meilleure fonction d'interpolation sur $[t, t_0+T]$. Il s'agit de calculer les p premiers vecteurs propres $\{u_i\}$ d'une matrice A (TxT). Les vecteurs servant à l'initialisation de la méthode de la puissance sont choisis de manière aléatoire. On passe à l'étape t.

Etape t: Les vecteurs $\{u_i\}$ calculés à l'étape précédente permettent de calculer un nouveau segment d'analyse. Il s'agit de [t+1,t+1+T] (recherche d'une nouvelle fonction d'interpolation) ou de [t+a,t+b+T] (itération du fenêtrage adaptatif) avec a et b.

a[ch]e	ac	a[da]ta	adata	a[ka]ta	akata
[ai]sta	èsta	a[dai]ta	adèta	a[kai]ta	akèta
[b]a	ba	a[db]ata	adbta	a[kb]ata	akbata
[d]a	da	a[deu]ta	adeta	a[kd]ata	akdata
[eu]ta	eta	a[di]ta	adita	a[keu]ta	aketa
[i]ta	ita	a[dj]ata	adxata	a[ki]ta	akita
[j]a	xa	a[dk]ata	adkata	a[kj]ata	akxata
[k]a	ka	a[do]ta	adota	a[ko]ta	akota
[o]ta	ota	a[doublevé]ta	adublvéta	a[kp]ata	akpata
[p]a	pa	a[dp]ata	adpata	a[ks]ata	aksata
[s]a	sa	a[ds]ata	adsata	a[kt]ata	aktata
[t]a	ta	a[dt]ata	adxata	a[ku]ta	akyta
[u]ta	yta	a[du]ta	adyta	a[kv]ata	akvata
[v]a	va	a[dv]ata	advata	a[kz]ata	akzata
[z]a	za	a[dz]ata	adzata	a[la]ta	alata
a[bé]ta	abéta	a[fa]ta	afata	a[lai]ta	alèta
a[cha]ta	acata	a[fai]ta	afèta	a[lb]ata	albata
a[chai]ta	acèta	a[fb]ata	afbata	a[ld]ata	aldata
a[chb]ata	acbata	a[fd]ata	afdata	a[leu]ta	aleta
a[chd]ata	acdata	a[feu]ta	afeta	a[li]ta	alita
a[cheu]ta	aceta	a[fi]ta	afita	a[lj]ata	alxata
a[chi]ta	acita	a[fj]ata	afxata	a[lk]ata	alkata
a[chj]ata	acxata	a[fk]ata	afkata	a[lo]ta	alota
a[chk]ata	ackata	a[fo]ta	afota	a[lp]ata	alpata
a[chota	acota	a[fp]ata	afpata	a[ls]ata	alsata
a[chp]ata	acpata	a[fs]ata	afsata	a[lt]ata	altata
a[chs]ata	acsata	a[ft]ata	aftata	a[lu]ta	alyta
a[cht]ata	actata	a[fu]ta	afyta	a[lv]ata	alvata
a[chu]ta	acyta	a[fv]ata	afvata	a[lz]ata	alzata
a[chv]ata	acvata	a[fz]ata	afzata	a[ma]ta	amata
a[chz]ata	aczata	a[ji]ta	axita	a[mai]ta	amèta
a[d-d]ata	addata	a[jé]ta	axéta	a[mb]ata	ambata
a[d]e	ad	a[k-k]ata	akkata	a[md]ata	amdata

a[meu]ta	ameta	a[rk]ata	arkata	at[aif]ta	atèfta
a[mi]ta	amita	a[ro]ta	arota	at[ail]a	atèla
a[mj]ata	amxata	a[rp]ata	arpata	at[aim]ta	atèmta
a[mk]ata	amkata	a[rs]ata	arsata	at[aine]ta	atènta
a[mo]ta	amota	a[rt]ata	artata	at[air]ta	ataata
a[mp]ata	ampata	a[ru]ta	aryta	at[ais]ta	atèsta
a[ms]ata	amsata	a[rv]ata	arvata	at[aj]a	ataxa
a[mt]ata	amtata	a[rz]ata	arzata	at[ak]a	ataka
a[mu]ta	amyta	a[sb]ata	asbata	at[ap]a	atapa
a[mv]ata	amvata	a[sd]ata	asdata	at[at]a	atata
a[mz]ata	amzata	a[sj]ata	asxata	at[av]a	atava
a[n]e	an	a[sk]ata	askata	at[az]a	ataza
a[na]ta	anata	a[so]ta	asota	at[eu-a]ta	ateata
a[nai]ta	anèta	a[sp]ata	aspata	at[eu-ai]ta	ateèta
a[nb]ata	anbata	a[st]ata	astata	at[eu-eu]ta	ateeta
a[nd]ata	andata	a[sv]ata	asvata	at[eu-i]ta	ateita
a[neu]ta	aneta	a[sz]ata	aszata	at[eu-o]ta	ateota
a[ni]ta	anita	a[té]ta	atèta	at[eu-u]ta	ateyta
a[nj]ata	anxata	a[vé]ta	avèta	at[eu]	ate
a[nk]ata	ankata	a[zaid]ta	azèdta	at[eub]a	ateba
a[no]ta	anota	as[sa]ta	asata	at[eud]a	ateda
a[np]ata	anpata	as[sai]ta	asèta	at[euj]a	atexa
a[ns]sata	ansata	as[seu]ta	aseta	at[euk]a	ateka
a[nt]ata	antata	as[si]ta	asita	at[eup]a	atepa
a[nu]ta	anyta	as[su]ta	asyta	at[eus]sa	atesa
a[nv]ata	anvata	at[a-a]ta	ataata	at[eut]a	ateta
a[nz]ata	anzata	at[a-ai]ta	ataèta	at[euv]a	ateva
a[pai]ta	apèta	at[a-eu]ta	ataeta	at[euz]a	ateza
a[ra]ta	arata	at[a-i]ta	ataita	at[i-a]ta	atiata
a[rai]ta	arèta	at[a-o]ta	ataota	at[i-ai]ta	atièta
a[rb]ata	arbata	at[a-u]ta	atayta	at[i-eu]ta	atieta
a[rd]ata	ardata	at[a]	ata	at[i-i]ta	atiita
a[reu]ta	areta	at[ab]a	ataba	at[i-o]ta	atiota
a[ri]ta	arita	at[ach]ta	atacta	at[i-u]ta	atiyta
a[rj]ata	arxata	at[ad]a	atada	at[i]	ati

at[ib]a	atiba	at[os]sa	atosa	at[é-eu]ta	atéeta
at[id]a	atida	at[ot]a	atota	at[é-i]ta	atéita
at[igraik]a	atigrèkta	at[ov]a	atova	at[é-o]ta	atéota
at[ij]a	atixa	at[oz]a	atoza	at[é-u]ta	atéyta
at[ik]a	atika	at[u-a]ta	atyata	at[é]	até
at[ip]a	atipa	at[u-ai]ta	atyèta	at[éb]a	atéba
at[is]sa	atièa	at[u-eu]ta	atyeta	at[éd]a	atéda
at[it]a	atita	at[u-i]ta	atyita	at[éj]a	atéxa
at[iv]a	ativa	at[u-o]ta	atyota	at[ék]a	aték a
at[iz]a	atiza	at[u-u]ta	atyyta	at[ép]a	atépa
at[o-a]ta	atoata	at[u]	aty	at[és]sa	atésa
at[o-ai]ta	atoèta	at[ub]a	atyba	at[ét]a	até ta
at[o-eu]ta	atoeta	at[ud]a	atyda	at[év]a	atéva
at[o-i]ta	atoita	at[uj]a	atyxa	at[éz]a	atéza
at[o-o]ta	atoota	at[uk]a	atyka	ata[s-s]a	atéssa
at[o-u]ta	atoyta	at[up]a	atypa	ta[f]	taf
at[o]	ato	at[us]sa	atysa	ta[k]	tak
at[ob]a	atoba	at[ut]a	atyta	ta[l]	tal
at[od]a	atoda	at[uv]a	atyva	ta[m]e	tam
at[oj]a	atoxa	at[uz]a	atyza	ta[r]e	tar
at[ok]a	atoka	at[é-a]ta	atéata	ta[s]se	tas
at[op]a	atopa	at[é-ai]ta	atéèta		

ANNEXE VI. VOCABULAIRE DE TEST

LUROT	èlyèroté	NICOLAS	èniséoèlaès
BRETAUX	béèretéayiks	NELSON	èneèlèsoèn
PECQUEUX	péesékyyeyiks	MURAT	èmyèraté
MORIN	èmoèrièn	NOGRET	ènoxéèreté
NAVARETTE	ènavéaèretétée	DUVAL	déyvéaèl
BRAIN	béèraièn	SALAT	èsaèlaté
BLANC	béèlaènsé	JACQUES	xiasékyyeès
BROSSARD	béèroèsèsaèrdé	RICHARD	èriséacaèrdé
RIVALANT	èrivéaèlaènté	SAVIO	èsavéio
REBEYROL	èrebéeigrèkèroèl	PUJOL	péyxioèl
HUSSON	acyèsèsoèn	GENDRE	xéeèndéère
FOUCAULT	èfoyèsayèlté	PORCHEZ	péoèrséacezèd
SOUCHET	èsoyséaceté	MACHAUT	èmaséacayté
VINCENT	véièns éeènté	LEGOS	èlexéoès
LABROCA	èlabéèroséa	GOUBIN	xéoybéièn
SUAREZ	èsyaèrezèd	MARTIN	èmaèrtéièn
DUPOY	déypéoigrèk	PONS	péoènès
ROBERT	èrobéeèrté		
MICHON	èmiséacoèn		
MICHEL	èmiséaceèl		
PINSON	péièn èsoèn		
SANCHEZ	èsaènséacezèd		
PETIT	péetéité		
DURAND	déyèraèndé		
ROUZEAU	èroyzèdeay		
BOURDIN	béoyèrdéièn		
JOBARD	xiobéaèrdé		
REUTER	èretéytéeèr		
QUENTIN	kyyeèntéièn		
CHOMET	séacoèmeté		
ROUSSEL	èroyèsèseèl		
PIOT	péioté		
LABONDE	èlabèoèndée		

ANNEXE VII. ALGORITHME DE WAGNER ET FISHER

L'algorithme de Wagner et Fisher est utilisé pour trouver l'alignement optimal entre X la chaîne phonétique trouvée de longueur n et Y la chaîne phonétique attendue de longueur m. Cet algorithme recherche la transformation de coût minimal qui permet de passer de X à Y. Cette transformation est une combinaison de trois opérations élémentaires, portant sur les sons, chacunes affectées d'un coût élémentaire λ :

— Remplacement du son [a] remplacé par le phonème [b] avec un coût $\lambda(a, b)$
— Insertion du son [b] avec un coût $\lambda(\emptyset, b)$
— Omission du son [b] avec un coût $\lambda(b, \emptyset)$

Le coût de la transformation est défini comme la somme des coûts élémentaires. L'algorithme d'alignement est le suivant :

Etape 0: Initialisation à zéro de deux matrices (m+1) x (n+1) A et B.

Etape 1: Pour i := 1 à n
A(i, 0) prend la valeur $A(i-1, 0) + \lambda(X_i, \emptyset)$.

Etape 2: Pour j := 1 à m
A(0, j) prend la valeur $A(0, j-1) + \lambda(\emptyset, Y_j)$.

Etape 3: Pour i := 1 à n et pour j := 1 à m
m_1 prend la valeur $A(i-1, j-1) + \lambda(X_i, Y_j)$.
m_2 prend la valeur $A(i-1, j) + \lambda(X_i, \emptyset)$.
m_3 prend la valeur $A(i, j-1) + \lambda(\emptyset, Y_j)$.

B(i, j)prend la valeur $\underset{k}{\text{Argmin}}\ (m_k)$

A(i, j) prend la valeur k si B(i, j) prend la valeur m_k.

Etape 4: i est initialisé à n, j est initialisé à m.
$Pho_X := X_i$
$Pho_Y := Y_j$
Suivant la valeur de B(i, j) trois cas se présentent :

— B(i, j) est égal à 1. Si Pho_X = Pho_Y, le son Pho_Y est bien celui qui est attendu. Dans l'autre cas il y a eu une substitution; i et j sont décrémentés.

— B(i, j) est égal à 2. Le son Pho_X a été omis;
i est décrémenté.

— B(i, j) est égal à 3. Le son Pho_Y a été inséré;
j est décrémenté.

Si i et j sont égaux à zéro l'algorithme s'arrête sinon on retourne à l'étape 4

La fonction de coût λ est choisie pour que l'algorithme d'alignement soit le plus robuste possible. Sa valeur va de 0 à 3 :

— $\lambda(X_i, Y_j) = 0$ si X_i est égal à Y_j.

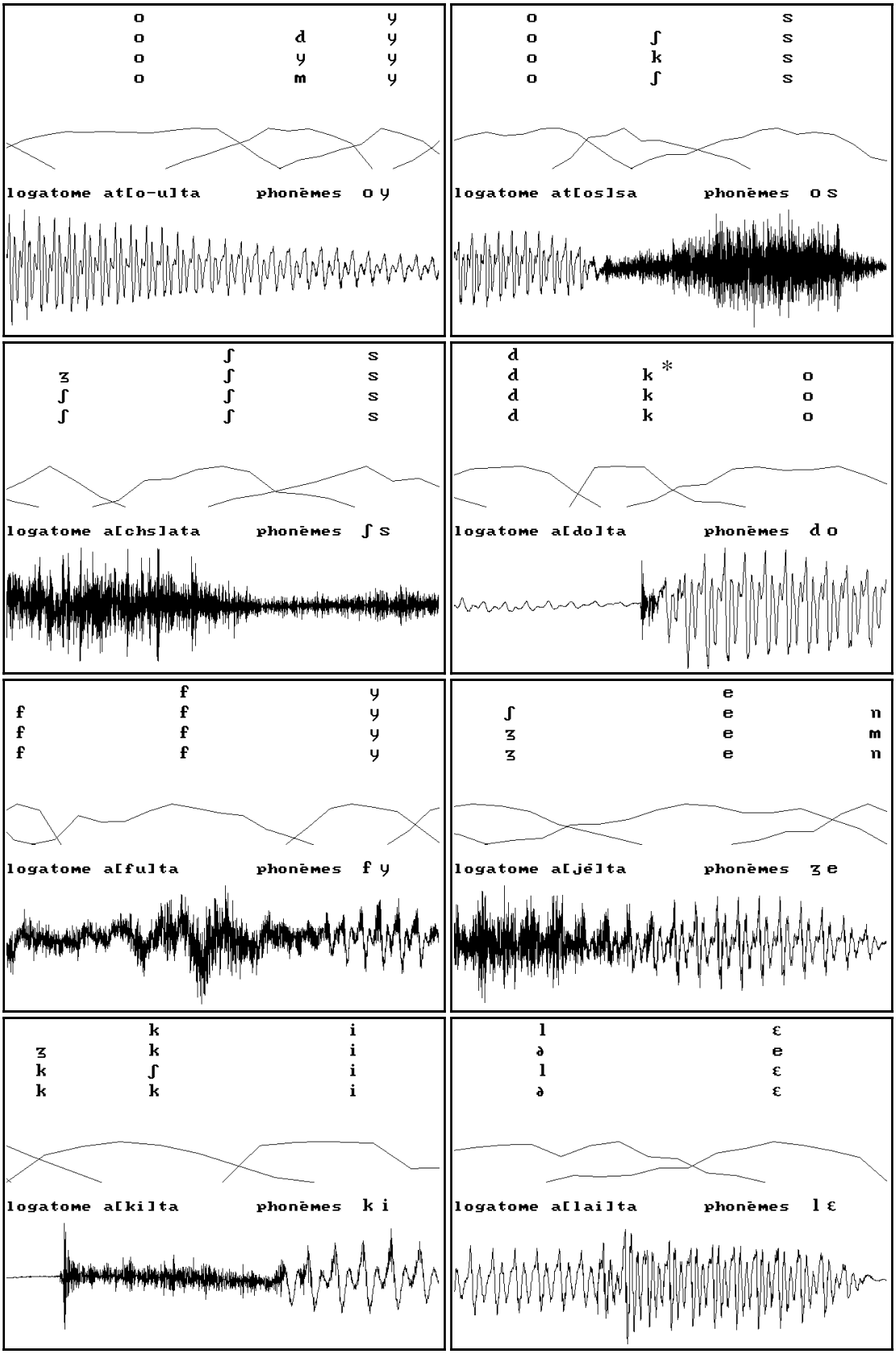
— $\lambda(X_i, Y_j) = 1$ si X_i et Y_j.sont sur la table des confusions fréquentes (cf. Tableau 4.3.)

— $\lambda(X_i, Y_j) = 2$ dans le cas d'une insertion ou d'une omission (i.e., $X_i = 0$ ou $Y_j = 0$).

— $\lambda(X_i, Y_j) = 3$ dans les autres cas.

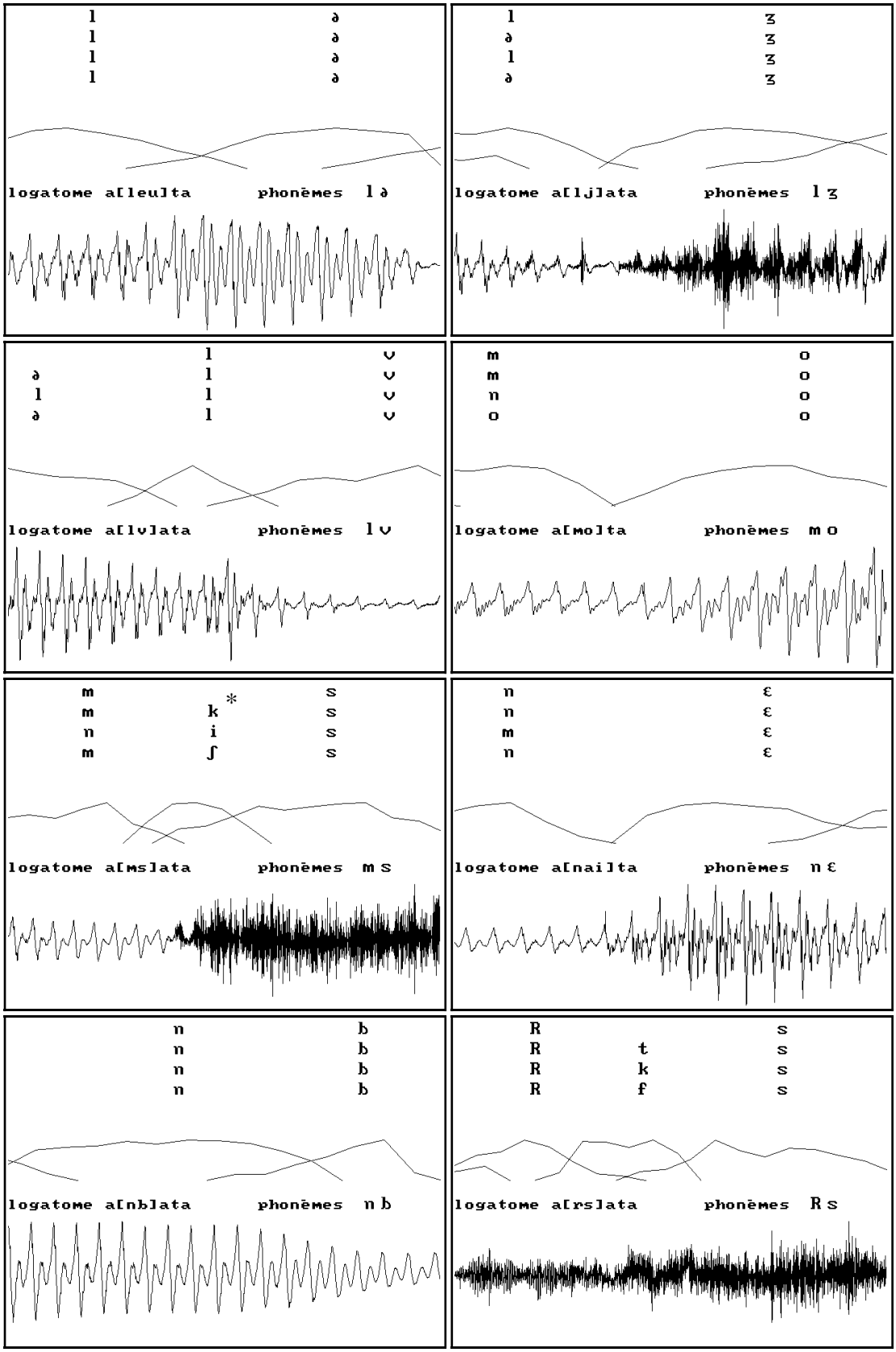
ANNEXE VIII. ARTICLE ICASSP

ANNEXE IX. EXEMPLES DE DÉCOMPOSITION TEMPORELLE

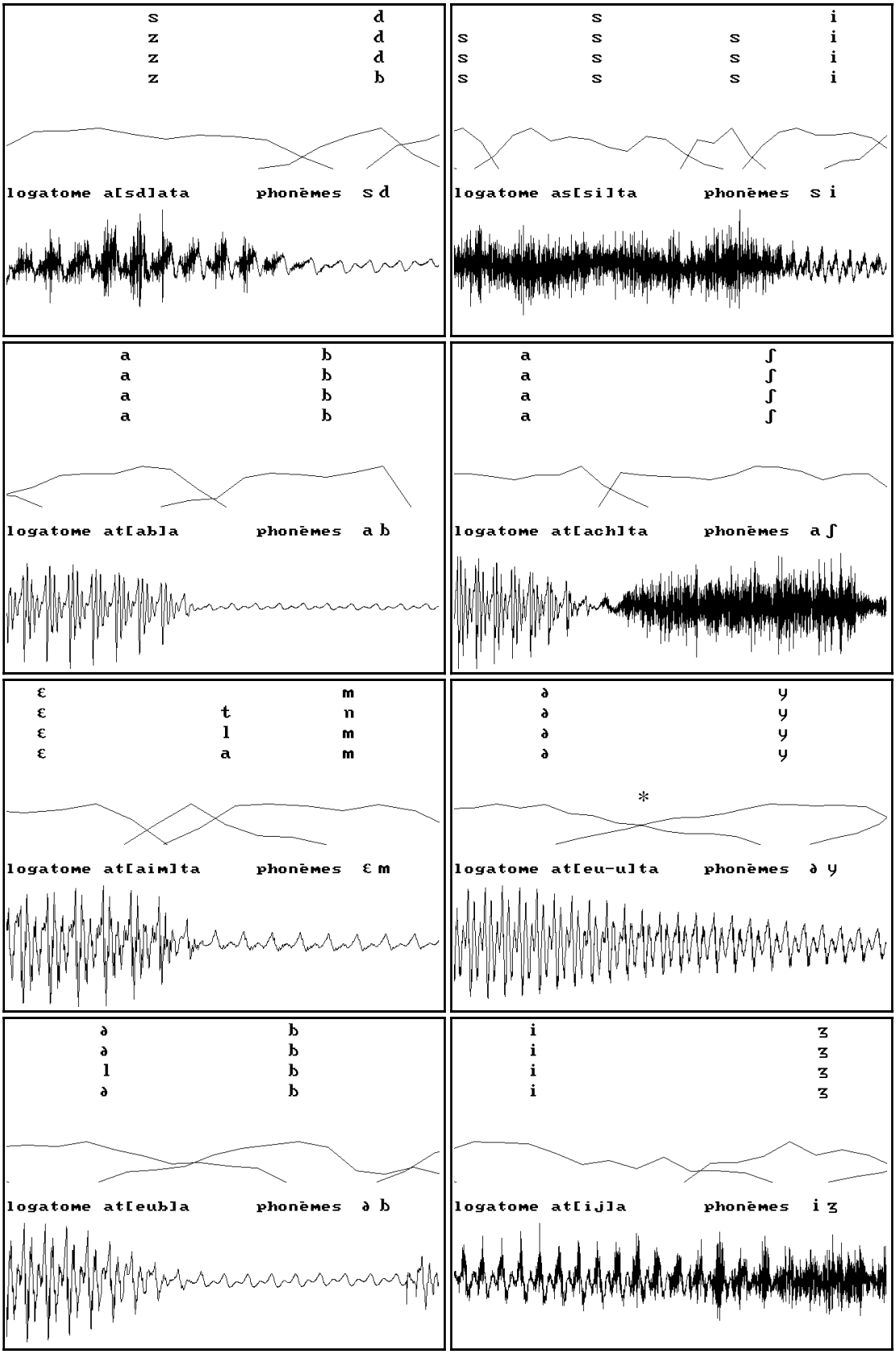


* Explosion du (d)

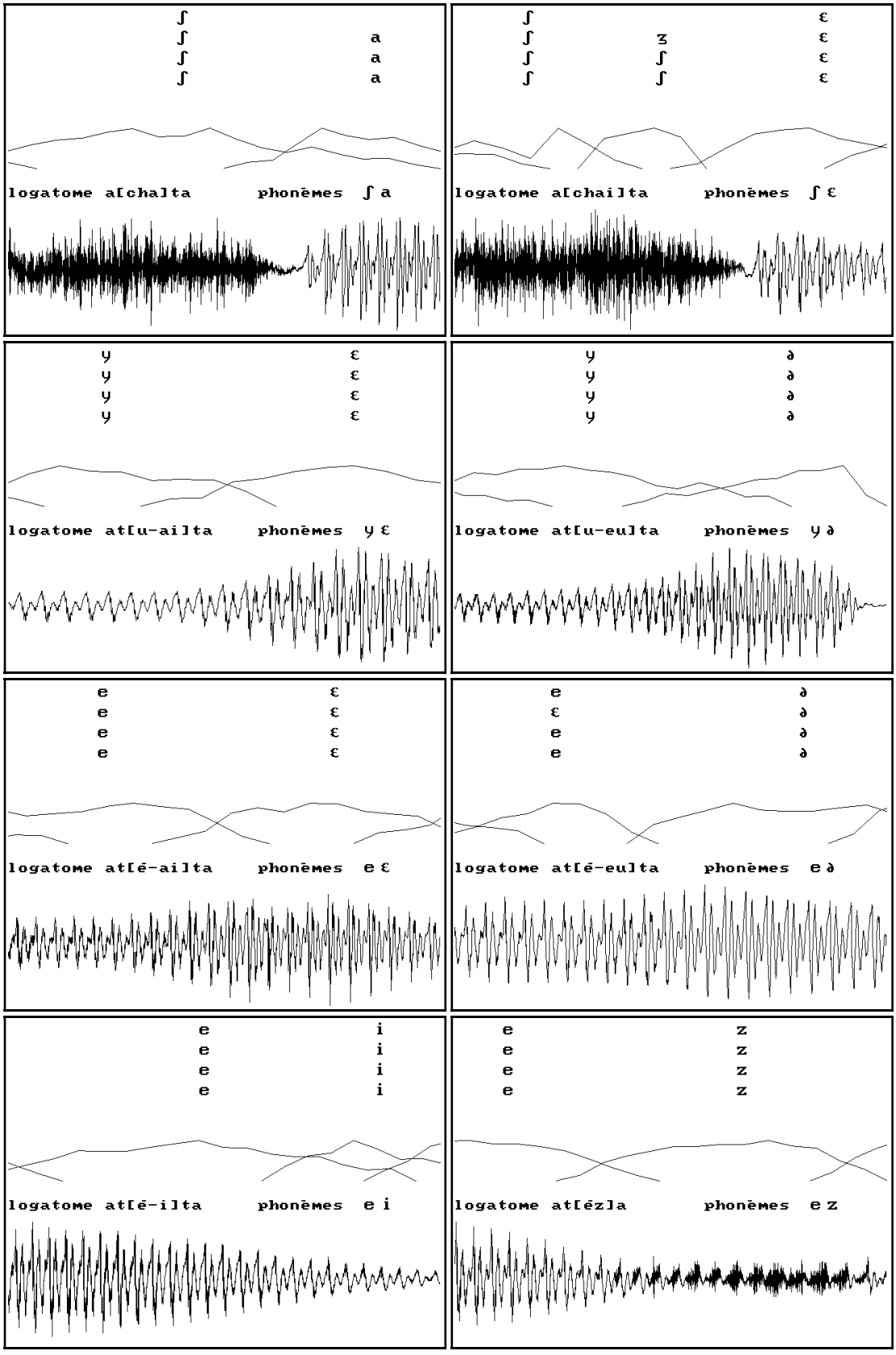
Le contenu des figures est décrit page 67.



* Cible de transition



* Transition lente (ey)


ʃ ʃ ʃ ʃ
a a a
logatome a[cha]ta phonèmes ʃ a
ʃ ʃ ʃ ʃ
ʒ ʃ ʃ
ɛ ɛ ɛ ɛ
logatome a[chai]ta phonèmes ʃ ɛ
ɥ ɥ ɥ ɥ
ɛ ɛ ɛ ɛ
logatome at[u-ai]ta phonèmes ɥ ɛ
ɥ ɥ ɥ ɥ
ə ə ə ə
logatome at[u-eu]ta phonèmes ɥ ə
e e e e
ɛ ɛ ɛ ɛ
logatome at[é-ai]ta phonèmes e ɛ
e e e e
ə ə ə ə
logatome at[é-eu]ta phonèmes e ə
e e e e
i i i i
logatome at[é-i]ta phonèmes e i
e e e e
z z z z
logatome at[éz]a phonèmes e z

ANNEXE X. ALPHABET PHONÉTIQUE

	Code API	Mot clé
Voyelles orales		
	(a)	patte
	(i)	lit
	(é)	été
	(è)	lait
	(e)	peur
	(y)	but
	(o)	pot
	(u)	sou
Occlusives sourdes		
	(p)	pou
	(t)	tout
	(k)	cou
Occlusives sonores		
	(b)	bout
	(d)	doux
	(g)	goût
Fricatives sourdes		
	(f)	faon
	(s)	sang
	(c)	champ
Fricatives sonores		
	(v)	vent
	(z)	zan
	(x)	gens
Liquides		
	(r)	rang
	(l)	lent
Nasales		
	(m)	mot
	(n)	nos

TABLE DES FIGURES

Figure A.1. Méthode de reconnaissance globale [Potage 86].....2
Figure A.2. Méthode avec reconnaissance de phonèmes [Potage 86].....2
Figure 1.1. Modèle de Bakis 17
Figure 1.2. Exemple de décomposition temporelle sur le segment (ia) 18
Figure 2.1. différentes fonctions caractéristiques 33
Figure 2.2. Fonction d'interpolation pour deux intervalles 34
Figure 2.3. Itérations du fenêtrage adaptatif 39
Figure 2.4. Chronogramme du triphone (èro) tiré de l'épellation du mot "LUROT" 45
Figure 2.5. Trajectoire dans le plan principal d'inertie pour une paramétrisation MFCC, associée à une distance euclidienne 46
Figure 2.6. Trajectoire dans le plan principal d'inertie pour une paramétrisation LAR, associée à une distance euclidienne..... 47
Figure 2.7. Trajectoire dans le plan principal d'inertie pour une paramétrisation avec des coefficients de corrélation, associée à une distance euclidienne..... 47
Figure 2.8. Trajectoire dans le plan principal d'inertie pour une paramétrisation avec des coefficients de corrélation, associée à une mesure de dissimilarité d'Itakura et Saito 48
Figure 3.1. Spectre de la voyelle (a) avec une Echelle fréquentielle linéaire 53
Figure 3.2. Spectre de la voyelle (a) et Banc de filtres avec une Echelle fréquentielle Mel 53
Figure 3.3. Cepstre de la voyelle (a) et gabarit du lissage cepstral..... 56
Figure 3.4. Lissage cepstral itératif pour un spectre de la voyelle (a) pour une voix masculine 58
Figure 3.5. Lissage cepstral itératif pour un spectre de la voyelle (a) pour une voix féminine..... 58
Figure 3.6. Lissage cepstral itératif pour un spectre de la voyelle (a) pour une voix masculine avec une Echelle fréquentielle Mel 59
Figure 3.7. Chronogrammes de l'indice d'instabilité du mot épelé en continu "PECQUEUX" pour deux paramétrisations, WMFCC et WMFCC avec lissage cepstral itératif 61
Figure 4.1. Synoptique du processus de décodage acoustico-phonétique 63

Figure 4.2. Illustration de l'étiquetage semi-automatique 67
Figure 5.1. Synoptique de la transformation 76
Figure 5.2. Perceptron multi-couches 79
Figure 5.3. Noeud du perceptron multi-couches 79
Figure 5.4. Rapport Signal à Bruit en fonction du nombre d'itérations
pour une initialisation linéaire 84
Figure 5.5. Séparation du noeud le plus saturé 85
Figure 5.6. Ajustement temporel 88
Figure 5.7. Teintes de gris associées à la fréquence d'une composante 97
Figure 5.8. Représentation des mots du vocabulaire (locuteur GNL) 98
Figure 5.9. Synoptique du choix du vocabulaire d'adaptation 99
Figure 5.10. Représentation du meilleur sous-vocabulaire représentatif
(locuteur GNL) 99
Figure 5.11. Représentation du vocabulaire tout entier (locuteur GNL) 99
Figure 6.1. Décodage acoustico-phonétique du mot "DURAND" 105

TABLE DES TABLEAUX

Tableau 1.1. Bases de fonctions orthogonales 15
Tableau 2.1. Nombre de fonctions d'interpolation et moyenne de leur durée
en fonction de la paramétrisation et la mesure de dissimilarité..... 45
Tableau 3.1. Echelle de fréquence Mel 52
Tableau 4.1. Comparaison de l'autocohérence et des confusions 68
Tableau 4.2. Informations complémentaires (WMFCC locuteur FB) 73
Tableau 4.3. Exemples de la discrimination obtenue par l'ensemble
des paramètres complémentaires 74
Tableau 5.1. Matrice d'alignement pour une paramétrisation LAR
sur 12 coefficients d'un dictionnaire de logatome..... 91
Tableau 5.2. Matrice d'alignement pour une paramétrisation WMFCC
sur 8 coefficients d'un dictionnaire de logatome 92
Tableau 5.3. Matrice d'alignement pour une paramétrisation WMFCC
avec lissage cepstral itératif sur 8 coefficients
d'un dictionnaire de logatome 92
Tableau 5.4. Intervalle de confiance à 95% du χ^2 94
Tableau 5.5. Vocabulaire de la base de données TEXAS 95
Tableau 5.6. Vocabulaire optimal d'adaptation de cinq mots 96
Tableau 5.7. Vocabulaires optimaux d'adaptation (locuteur GNL) 96
Tableau 5.8. Comparaison des deux expériences 100
Tableau 6.1. Topologie des différents modèles employés 106
Tableau 6.2. Résultats du système comparatif, 1er choix, sur 513 sons 107
Tableau 6.3. Reconnaissance phonétique, 1er choix, sur 513 sons 108
Tableau 6.4. Adaptation à de nouveaux locuteurs..... 109
Tableau II.1. Topologie des différents modèles employés 131
Tableau II.2. Apprentissage (Diphones FB)
et Reconnaissance (Mots épelés FB) 131
Tableau II.3. Apprentissage (Diphones MJC)
et Reconnaissance (Mots épelés MJC) 131
Tableau II.4. Apprentissage (Diphones FB)
et Reconnaissance (Diphones FB) 132
Tableau II.5. Apprentissage (Diphones MJC)
et Reconnaissance (Diphones MJC) 132

Tableau II.6. Apprentissage (Lettres SL)
et Reconnaissance (Lettres SL)..... 132
Tableau II.7. Apprentissage (Lettres FB)
et Reconnaissance (Lettres FB) 133
Tableau II.8. Apprentissage (Diphones FB et MJC)
et Reconnaissance (Mots épelés FB) 133
Tableau II.9. Apprentissage (Diphones FB et MJC)
et Reconnaissance (Mots épelés MJC) 133
Tableau II.10. Apprentissage (Diphones MJC)
et Reconnaissance (Mots épelés FB) 134
Tableau II.11. Apprentissage (Diphones FB)
et Reconnaissance (Mots épelés MJC) 134

TABLE DES MATIÈRES

Résumé I
Sommaire III
Notations utilisées V
Introduction 1
A. La reconnaissance de la parole continue 2
A.1. Etat de l'art en reconnaissance de la parole continue 3
A.3. Les bases de données 6
B. Conclusions et Plan 7
B.1. Plan 8
Chapitre 1. Les modèles de l'évolution spectrale 11
1.1. La prédiction linéaire vectorielle 11
1.2. Le modèle a cibles spectrales 13
1.3. Le modèle évolutif 14
1.4. Les machines de Markov cachées 15
1.4.1. Le modèle de Markov 16
1.4.2. L'apprentissage d'un modèle 17
1.5. La décomposition temporelle classique 17
1.5.1. Principes de résolution [Atal 83] 19
Conclusions 22
Chapitre 2. La décomposition temporelle généralisée 23
2.1. les opérateurs de reconnaissance 23
2.2.1. Les géodésiques 25
2.2.2. Une approximation de l'interpolation linéaire généralisée 29
2.2.3. L'analyse factorielle 30
2.2.4. Conclusions et adaptation de la décomposition temporelle à une mesure de dissimilarité quelconque 32
2.3. Estimation d'une fonction d'interpolation significative sur un intervalle 36
2.3.1. Réduction de l'espace spectral 36
2.3.2. Choix des intervalles de calcul 37
2.3.2.1. Méthodes classiques 37
2.3.3.2. Fenêtrage adaptatif 38
2.4. Constitution d'un ensemble cohérent

de fonctions d'interpolation..... 39
2.5. Le calcul des cibles spectrales 41
2.5.1. Projection Factorielle inverse 41
2.5.1.1. Méthode d'optimisation 43
2.6. Expériences préliminaires sur la technique de décomposition temporelle généralisée..... 44
Conclusions..... 48
Chapitre 3. Paramétrisation spectrale adaptée au décodage acoustico-phonétique et à la décomposition temporelle généralisée : Implantation d'un lissage cepstral itératif 49
3.1. La prédiction linéaire..... 49
3.1.1. La paramétrisation LAR (Log Area Ratio)..... 50
3.1.2. La distance d'Itakura et Saito 51
3.2. La paramétrisation WMFCC..... 52
3.2.1. Calcul des WMFCC..... 52
3.2.2 Lissage cepstral itératif..... 54
3.2.2.1. Algorithme de détection et calcul de la fréquence fondamentale..... 54
3.2.2.2. Lissage cepstral classique..... 56
3.2.2.3. Lissage cepstral itératif 57
3.3. Calcul d'un indice d'instabilité de la paramétrisation 60
Conclusions..... 61
Chapitre 4. Le processus de décodage acoustico-phonétique..... 63
4.1. Choix du classificateur. 64
4.1.1. La règle de décision bayésienne 64
4.1.2. La règle du perceptron multi-couches..... 64
4.1.3. La règle des K plus proches voisins 65
4.2. Constitution du dictionnaire de références 65
4.2.1. Corpus d'apprentissage..... 66
4.2.2. L'étiquetage du dictionnaire de références..... 66
4.2.3. Cibles de transition..... 69
4.3. Informations complémentaires sur les cibles..... 69
4.3.1. L'énergie, le nombre de ppz, l'instabilité spectrale, la durée..... 70
4.3.2. La pondération des paramètres..... 71

4.3.3. La discrimination des caractéristiques complémentaires 72
Conclusions 74
Chapitre 5. Adaptation au locuteur par transformations non-linéaires des cibles spectrales 75
5.1. Estimation d'une fonction de transformation non-linéaire 75
5.1.1. La régression linéaire multiple 76
5.1.2. Le filtrage polynomial 77
5.1.3. Le perceptron multi-couches 78
5.1.3.1. La rétropropagation du gradient 80
5.1.3.2. Préparation des données d'apprentissage 82
5.1.3.3. Initialisation linéaire des poids 83
5.1.3.4. Séparation des noeuds saturés 84
5.1.3.5. L'implémentation de l'algorithme 86
5.2. Alignement des signaux de parole 86
5.2.1. Les algorithmes de comparaison dynamique 87
5.2.2. Les limitations de la comparaison dynamique 89
5.2.3. Calcul d'un indice de qualité de l'alignement dynamique 89
5.2.4. Algorithme d'alignement dynamique itératif 90
5.3. Choix du vocabulaire d'adaptation 93
5.3.1. Expériences préliminaires sur une base de mots isolés 95
Conclusions 101
Chapitre 6. Evaluation du système de décodage acoustico-phonétique 103
6.1. Corpus 103
6.1.1. Corpus de test 103
6.1.2. Corpus d'apprentissage 104
6.2. Calcul du taux de reconnaissance 104
6.2.1. Alignement de la chaîne phonétique décodée 104
6.3 Système de référence comparatif 105
6.3.1 Description 105
6.3.2 Topologies employées 106
6.4. Résultats et discussion 107
6.4.1. Choix de la paramétrisation 107
6.4.2. Différences Homme-Femme et confusions 108
6.4.3. Adaptation au locuteur 109
Conclusions 110

Conclusions et Perspectives..... 111
Références Bibliographiques..... 113
Annexe I. Distance d'Itakura-Saito..... 129
Annexe II. Résultats comparatifs du système de référence..... 131
Annexe III. Syntaxe et modèles du système de référence 135
Annexe IV. Calcul des vecteurs et des valeurs propres d'une matrice 141
Annexe V. Vocabulaire d'apprentissage..... 145
Annexe VI. Vocabulaire de test..... 149
Annexe VII. Algorithme de Wagner et Fisher 151
Annexe VIII. Article ICASSP 153
Annexe IX. Exemples de décomposition temporelle 157
Annexe X. Alphabet phonétique..... 161
Table des Figures..... 163
Table des Tableaux..... 165
Table des Matières..... 167