Université Henri Poincaré – Nancy I

Département de Formation Doctorale en Informatique — École Doctorale IAE + M

Reconnaissance automatique de la parole continue en environnement bruité : application à des modèles stochastiques de trajectoires

THÈSE

présentée et soutenue publiquement le 4 septembre 1995

pour l'obtention du

Doctorat de l'Université Henri Poincaré – Nancy I

(Spécialité Informatique)

par

Olivier Siohan

Composition du jury

Président :	René Schott
Rapporteurs :	Régine André-Obrecht Jean-Pierre Martens René Schott
Examinateurs :	Philip Lockwood Christian Wellekens
Directeurs de thèse :	Yifan Gong Jean-Paul Haton

Centre de Recherche en Informatique de Nancy
INRIA–Lorraine

Remerciements

Je souhaite remercier René Schott, professeur à l'université Henri Poincaré, de me faire l'honneur et le plaisir de présider ce jury. Qu'il trouve ici l'expression de ma profonde gratitude pour l'intérêt manifesté à l'égard de ce travail.

Je suis très reconnaissant à Régine André-Obrecht, chargée de recherches au CNRS, d'avoir accepté d'être rapporteur de cette thèse. Je souhaite lui apporter mes plus vifs remerciements pour l'attention portée à ce mémoire.

Je suis très honoré que Jean-Pierre Martens, professeur à l'université de Gent, Belgique, ait accepté de juger ce travail. Je suis très sensible à l'attention qu'il a consacré à la lecture du manuscrit, qui n'était pas rédigé dans sa langue maternelle.

Je remercie sincèrement Philip Lockwood, responsable du département traitement de parole à MATRA Communication, d'avoir accepté de participer à ce jury.

Je suis sensible à l'attention dont Christian Wellekens, professeur à EURECOM, a fait preuve à la lecture de ce travail. Je le remercie d'avoir accepté de siéger à ce jury.

Jean-Paul Haton, professeur à l'université Henri Poincaré, et Yifan Gong, chargé de recherches CNRS, ont tous deux dirigés mes travaux. Je suis particulièrement reconnaissant à Jean-Paul Haton de m'avoir accueilli au sein de son équipe, et de m'avoir apporté un environnement de travail de grande qualité. Je le remercie pour l'intérêt constant porté à mes recherches et pour la qualité de ses remarques et suggestions, apportées tout au long de ce travail. Cette thèse n'eût pu avoir lieu sans les conseils et l'encadrement efficace d'Yifan Gong. J'ai été très impressionné par la précision de ses analyses et par ses capacités à percevoir et expliquer les problèmes de reconnaissance de parole. Je le remercie pour sa disponibilité, ses encouragements, et son enthousiasme scientifique.

Claude Lhermitte, chef du service informatique de Supélec Metz, a initié ce travail en me permettant de suivre le DEA informatique de l'université de Nancy, en parallèle avec les enseignements de Supélec. Je lui exprime ici mes sincères remerciements.

Je désire remercier également Dominique Fohr et Jean-François Mari, pour l'aide précieuse qu'ils m'ont apportée pendant toute la durée de cette thèse.

Enfin, je souhaite remercier particulièrement mes amis pour leur soutien et leurs encouragements, notamment Zouzou et Pierre, Irina, le Monstre, bien qu'il ait plusieurs fois tenté de m'empoisonner avec sa bouffe, Nicolas, ainsi qu'Isko. Merci également à Docteur TEX, pour sa classe TheseCRIN et ses conseils LaTeX 2_ε.

Résumé

Les systèmes actuels de reconnaissance automatique de la parole (RAP) sont généralement peu robustes aux variations du signal intervenant entre les conditions de test et d'apprentissage. Dans cette thèse, nous proposons et évaluons différentes approches pour améliorer la robustesse au bruit du système de reconnaissance de parole continue VINICS du CRIN-INRIA Lorraine, système fondé sur des modèles stochastiques de trajectoires de parole (STM), alternative efficace aux traditionnels modèles de Markov cachés (HMM).

Dans une première partie, nous dressons un bilan des principales approches développées ces dix dernières années dans le domaine de la RAP dans le bruit.

La seconde partie est constituée d'une étude et comparaison de trois approches. Nous développons d'une part une approche permettant d'estimer un STM hybride de parole bruitée, à partir d'un HMM de bruit et d'un STM de parole propre. D'autre part, nous proposons d'appliquer un filtrage du signal, spécifique à chaque état de chaque STM et optimisé selon un critère significatif au niveau perceptif. Ensuite, nous appliquons une méthode d'adaptation des STMs de parole propre aux variations des conditions d'environnement, calculée par régression linéaire. Ces trois approches sont comparées expérimentalement sur une tâche de reconnaissance de la parole continue, en mode dépendant du locuteur, pour un vocabulaire d'un millier de mots, en présence de différents bruits additifs réels. L'adaptation par transformation linéaire s'avère beaucoup plus efficace que les autres approches.

Enfin, dans une dernière partie, nous développons d'une part une étude expérimentale sur l'utilisation de l'analyse discriminante linéaire pour mettre en œuvre un paramétrage du signal de parole robuste au bruit. Nous mettons en évidence que l'analyse discriminante permet d'obtenir un paramétrage efficace pour la reconnaissance de la parole dans le bruit. Cependant, nos expériences montrent qu'un tel paramétrage est peu robuste aux variations du rapport signal-à-bruit, mais cette conclusion reste très dépendante de la nature du bruit. D'autre part, nous prenons en compte les variations du rythme d'élocution provoquées par l'effet Lombard, en utilisant une méthode d'adaptation des modèles de durée des phonèmes, sous le cadre général de l'apprentissage Bayésien. Cette méthode, évaluée sur une tâche de reconnaissance de mots isolés permet d'améliorer de façon significative les taux de reconnaissance.

Mots-clés: reconnaissance automatique de parole bruité, modèles stochastiques de trajectoires, combinaison de modèles stochastiques, filtrage de bruit, adaptation à l'environnement, regression linéaire, analyse linéaire discriminante, estimation Bayésienne

Abstract

Most automatic speech recognisers perform poorly when the training and testing conditions are not matched. This dissertation describes a number of algorithms that improve the noise robustness of the VINICS continuous speech recogniser, developed at CRIN-INRIA Lorraine and based on stochastic modeling of speech trajectories.

First, we give a review of recent works dealing with noisy speech recognition.

Second, we propose and compare three noisy speech recognition approaches. In the first one, a noisy speech Stochastic Trajectory Model (STM) is derived from a clean speech STM and an hidden Markov model of noise. The second approach performs an STM state-based filtering of the noisy speech parametric signal, where the estimators are derived in the log-spectrum domain. In the last method, we develop an adaptation framework based on linear regression to adapt STM means to new environments. Experiments on a continuous speech recognition task on speech corrupted by different additive noises show that the adaptation using linear regression outperforms the other approaches.

In the last part, we first study the noise robustness of speech parameters derived using linear discriminant analysis. The derived parameters are very efficient when the training and the testing signal-to-noise ratios are matched, but are very sensitive to signal-to-noise ratio variations. Then, we propose the use of Bayesian reestimation to reduce the mismatches between phone duration in Lombard and clean speech. On an isolated word speech recognition task, the adaptation of phone duration models significantly improve the recognition rate.

Keywords: noisy speech recognition, stochatic trajectory models, stochastic model combination, noise filtering, linear regression, linear discriminant analysis, Bayesian adaptation

Table des matières

Introduction 1

I Étude bibliographique 7

Introduction 9

1 Transformation de la parole 11

1 Soustraction spectrale 11
2 Annulation adaptative de bruit 14
3 Transformation d'espace 16
3.1 Transformation de *codebook* 16
3.2 Transformation analytique d'espace 18
3.3 Transformation d'espace par réseaux de neurones 19
4 Exploitation de la structure harmonique du signal de parole 20
5 Masquage de bruit 22
6 Estimation à base de modèles 23
7 Compensation de l'effet Lombard 26
8 Conclusion 26

2 Transformation des systèmes de reconnaissance 29

1 Composition/décomposition de modèles 29
2 Filtrage par état 31
3 Adaptation des modèles acoustiques 33
4 Adaptation des modèles de durée 34
5 Apprentissage discriminant 35
6 Apprentissage multiréférences 35

7 Conclusion 37

3 Paramétrages et mesures de similarité robustes **39**

1 Représentations acoustiques et distances robustes 39

2 Mesures de distorsion et paramétrages discriminants 43

3 Paramétrage à base de modèles auditifs 45

4 Suppression des variations lentes 47

5 Conclusion 50

Conclusion **51**

II Le système VINICS **53**

Introduction **55**

4 Modélisation stochastique des trajectoires de parole **61**

1 Trajectoire de parole 61

2 Définition du modèle stochastique de trajectoires 61

2.1 Introduction 61

2.2 Rééchantillonnage des trajectoires 62

2.3 Modélisation des trajectoires rééchantillonnées 63

2.4 Amélioration du modèle 64

3 Utilisation du modèle en reconnaissance 65

4 Estimation du modèle 67

4.1 Estimation de la distribution *a priori* des symboles $Pr(s)$ 67

4.2 Estimation du modèle de durée $p_{d|s}(d|s)$ 68

4.3 Estimation du modèle de trajectoires $p_{\mathbf{X}|d,s}(\mathbf{X}|d,s)$ 69

5 Récapitulatif 70

5 Recherche de la meilleure phrase **73**

1 Introduction 73

2 Probabilité d'une séquence de symboles 73

3 Algorithme de calcul de $\Theta(w)$ 75

4 Recherche de la meilleure phrase 76

Conclusion **79**

III Trois approches pour la reconnaissance de la parole bruitée 81

Introduction 83

6 Combinaison de modèles stochastiques 85
1 Introduction . . . 85
2 Caractérisation de la *pdf* d'un vecteur de parole bruitée . . . 87
3 Application à la combinaison d'un STM et d'un HMM . . . 90
3.1 Modèle de bruit à un état . . . 90
3.2 HMM de bruit . . . 91
4 Conclusion . . . 93

7 Filtrage non linéaire par états 95
1 Introduction . . . 95
2 Estimation spectrale non linéaire . . . 96
2.1 Soustraction spectrale linéaire . . . 96
2.2 Estimation spectrale non linéaire . . . 97
3 Estimation MMSE numérique . . . 98
3.1 Principe . . . 98
3.2 Application à des modèles stochastiques définis sur les cepstres . . . 99
4 Conclusion . . . 100

8 Adaptation au bruit par régression linéaire 101
1 Introduction . . . 101
2 Principe de l'adaptation . . . 102
2.1 Transformation des paramètres d'un STM . . . 103
2.2 Transformation des vecteurs de test . . . 104
2.3 Critère d'estimation des transformations . . . 105
3 Adaptation des paramètres des STMs . . . 105
3.1 STMs sans mixtures . . . 105
3.2 STMs avec mixtures . . . 108
4 Adaptation des vecteurs de test . . . 110
4.1 STMs sans mixtures . . . 110
4.2 STMs avec mixtures . . . 111
5 Réduction du nombre de paramètres des transformations . . . 112

6 Conclusion 113

9 Expériences et résultats **115**

1 Conditions expérimentales 115

1.1 Description des corpus de parole et bruit 115

1.2 Modélisation acoustique 116

1.3 Décodage des phrases 117

2 Configuration des différentes approches 117

2.1 Configuration pour la combinaison stochastique de modèles 117

2.2 Configuration pour le filtrage par états 121

2.3 Configuration pour la régression linéaire 123

2.4 Configuration pour la transformation de base 125

3 Comparaison des performances 128

4 Sensibilité de la régression linéaire aux variations du SNR 136

5 Conclusion 137

IV Prétraitement par LDA et adaptation des modèles de durée **141**

Introduction **143**

10 Robustesse du prétraitement par analyse linéaire discriminante **145**

1 Introduction 145

2 Analyse linéaire discriminante 146

3 Expériences et résultats 147

3.1 Conditions expérimentales 147

3.2 Résultats 148

4 Conclusion 157

11 Adaptation Bayésienne des modèles de durée des phonèmes **161**

1 Introduction 161

2 Estimation Bayésienne 161

2.1 Principe 161

2.2 Application à la réestimation des modèles de durée des phonèmes . 162

3 Expériences et résultats 164

3.1 Conditions expérimentales 164

3.2 Résultats 166
4 Conclusion 169

Conclusion et perspectives **171**

Bibliographie **175**

Index **197**

Annexes **199**

A Algorithme EM **201**
1 Inégalité de Jensen 201
2 Algorithme EM 202

B Transformation de base **205**
1 Introduction 205
2 Notation et principe 205
3 Adaptation par transformation de base 206
3.1 Définition d'une base 206
3.2 Définition de la transformation de base 207
4 Conclusion 208

C Tableaux des résultats de reconnaissance **209**

Table des figures

0.1 Évolution du taux de reconnaissance d'un système de RAP entraîné en milieu calme (SNR >40 dB), en fonction du rapport signal-à-bruit lors du test [Siohan *et al.*, 1995]. 3

1.1 Schéma de principe de l'annulation adaptative de bruit. 15

3.1 Évolution d'une trajectoire de parole continue dans le plan F2–F1. Les cibles situées au centre des trajectoires ne sont pas atteintes. 56

3.2 Regroupement de trajectoires par un HMM. Une unité de parole s est représentée par les 2 classes de trajectoires a-x-b et c-x-d. Le HMM reconnaît la trajectoire hachurée a-x-d comme étant caractéristique de s, bien que ne faisant pas partie des classes d'apprentissage spécifiques à s [Gong, 1994]. . . 56

3.3 Représentation d'un ensemble de trajectoires spécifiques au symbole /*m*/ dans le plan C2–C3. Les trajectoires forment des classes dans l'espace. . . . 58

3.4 Variances des coefficients cepstraux C0 à C12 d'un ensemble de vecteurs d'observations associés au symbole / *o*/ en fonction de leurs positions dans la trajectoire de parole. Pour chaque coefficient cepstral, les variances sont normalisées par rapport à leur valeur minimale. 58

4.1 N_s trajectoires représentées dans un espace de dimension 2. Chaque trajectoire est constituée de d_i vecteurs, représentés par les points sombres. Les points sont espacés d'une période d'analyse du signal. 62

4.2 Sur-, ou sous-échantillonnage des trajectoires. Les trajectoires de durée d_1 et d_2 sont rééchantillonnées en Q points. 63

4.3 Les trajectoires rééchantillonnées sont modélisées par une distribution normale. 64

4.4 Les trajectoires d'un même symbole s se répartissent en classes en fonction du contexte acoustique. 65

5.1 Courbes de probabilité $Pr(s = /Z/|\mathbf{X}_n, d)$ du symbole /Z/ en fonction du temps n et de la durée d du symbole. Phrase prononcée : «je sors», /Z s o R/. 74

5.2 Numérotation des trames de symboles consécutifs. 74

5.3 Passage du calcul de la probabilité cumulée $\Pi(k, j-1)$ à $\Pi(l, j)$, entre le symbole de rang $j-1$ et le symbole de rang j. 76

5.4 Limitation de l'espace de recherche dans le calcul des probabilités cumulées, par des contraintes sur la durée du symbole courant a_j, limitée à l'intervalle $[d_j^m, d_j^M]$. 77

6.1 Structure d'un système de reconnaissance de parole avec combinaison de modèles. 87

6.2 Combinaison d'un STM de parole propre avec une loi normale de bruit. Le modèle de parole bruitée reste un STM. 91

6.3 Combinaison d'un STM de parole propre avec un HMM de bruit. Le modèle de parole bruitée devient un STM hybride. 92

7.1 Structure d'un système de reconnaissance de parole avec filtrage non linéaire des observations bruitées lors de la reconnaissance. 96

8.1 Structure d'un système de reconnaissance de parole bruitée avec transformation par régression linéaire des paramètres des STM de parole propre. . . . 102

8.2 Structure d'un système de reconnaissance de parole bruitée avec transformation des observations de parole bruitée lors de la reconnaissance. 103

8.3 Différentes structures des matrices de transformations. (a) : indépendance des coefficients cepstraux de des états d'un STM. (b) : dépendance des coefficients cepstraux et indépendance des états d'un STM. (c) : indépendance des coefficients et dépendance des états d'un STM. 113

9.1 Estimation des densités spectrales de puissance des différents bruits. Le bruit Gaussien est représenté en traits interrompus sur chaque figure. 116

9.2 Influence du nombre d'états du HMM de bruit sur le taux de reconnaissance de 206 mots isolés. 119

9.3 Distribution du coefficient cepstral d'ordre 1 en présence de différents niveaux de bruit blanc Gaussien. La distribution est calculée à partir des 79 phrases d'apprentissage prononcées par 1 locuteur. 120

9.4 Distributions des coefficients cepstraux 0 à 5 en présence de différents niveaux de bruit blanc Gaussien. Les distributions sont calculées à partir des 79 phrases d'apprentissage prononcées par 1 locuteur, après classification automatique de l'ensemble des symboles en 2 classes : 1 classe pour les symboles de forte énergie, 1 classe pour les symboles de faible énergie. 122

9.5 Comparaison entre la combinaison stochastique de modèles et la réestimation des modèles par l'algorithme EM à partir de données de parole bruitée artificiellement générées. Bruit blanc Gaussien. Moyenne sur tous les locuteurs. 123

9.6 Tables de filtrage MMSE pour les 13 coefficients cepstraux. Symbole /R/, état 0 du modèle, bruit blanc Gaussien, SNR 0 dB. 124

9.7 Comparaison entre une adaptation par régression linéaire avec ou sans translation. Bruit blanc Gaussien. Moyenne sur tous les locuteurs. 126

9.8 Comparaison des performances de la transformation de base. Première configuration : les éléments de la base sont obtenus à partir une seule occurrence du corpus d'adaptation (Sans moyenne). Seconde configuration : les éléments de la base sont obtenus à partir d'une moyenne de différentes versions bruitées du corpus d'adaptation (Avec moyenne). Moyenne des taux de reconnaissance sur tous les locuteurs. 127

9.9 Comparaison des performances de la régression linéaire. Première configuration : le corpus d'adaptation est constitué d'une seule occurrence bruitée des phrases d'adaptation (Sans moyenne). Seconde configuration : le corpus d'adaptation est obtenus à partir d'une moyenne de différentes versions bruitées des phrases d'adaptation (Avec moyenne). Moyenne sur tous les locuteurs. 129

9.10 Comparaison des performances des différentes approches. Bruit blanc et bruit de sèche-cheveux. Moyenne sur tous les locuteurs (cf. tables C.1 et C.2). 130

9.11 Comparaison des performances des différentes approches. Bruit d'avion F16 et d'hélicoptère Lynx. Moyenne sur tous les locuteurs (cf. tables C.3 et C.4). 131

9.12 Comparaison des performances des différentes approches. Bruit d'autobus. Moyenne sur tous les locuteurs (cf. table C.5). 132

9.13 Taux de reconnaissance *vs* SNR. 3 conditions de test : pas de compensation (Modèles de parole propre), apprentissage dans les conditions de test (Apprentissage dans le bruit), adaptation au bruit par régression linéaire (Régression linéaire). Moyenne sur tous les locuteurs. 135

9.14 Évolution des taux de reconnaissance en fonction de la modification de la variance des modèles MLE bruités. Moyenne sur tous les locuteurs. 136

9.15 Sensibilité de l'adaptation par régression linéaire aux variations du rapport signal-à-bruit entre le corpus d'adaptation et le corpus de test. Moyenne sur tous les locuteurs. 138

10.1 Procédure de test de la robustesse de la LDA aux variations du rapport signal-à-bruit entre le corpus d'apprentissage et le corpus de test. 149

10.2 Comparaison de la robustesse du paramétrage LDA / MFCC aux variations du rapport signal-à-bruit entre le corpus utilisé pour déterminer la matrice de projection LDA et le corpus d'évaluation. Les courbes indiquent le taux de reconnaissance phonétique du corpus utilisé pour l'apprentissage des modèles. Moyenne sur tous les locuteurs. 150

10.3 Robustesse relative des paramétrage LDA / MFCC aux variations du rapport signal-à-bruit SNR_{test} et SNR_{ref}. Écart relatif des taux d'évaluation phonétique par rapport à la configuration où $\mathrm{SNR}_{ref} = \mathrm{SNR}_{test}$. Moyenne sur tous les locuteurs. 152

10.4 Robustesse du prétraitement par LDA aux variations du rapport signal-à-bruit entre le corpus utilisé pour déterminer la matrice de projection LDA et le corpus de test. Moyenne sur tous les locuteurs. 153

10.5 Influence de la limitation du nombre maximum de symboles par classe lors du calcul de la matrice LDA sur la robustesse aux variations du SNR entre le test et l'apprentissage. Limitation à 40 et 70 symboles par classe. Moyenne sur tous les locuteurs. 155
10.6 Influence de l'utilisation ou non des symboles de silences lors du calcul de la matrice LDA sur la robustesse aux variations du SNR entre le test et l'apprentissage. Moyenne sur tous les locuteurs. 156
10.7 Influence de la notion de classe sur la robustesse aux variations du SNR entre le test et l'apprentissage. Première configuration : classe = symbole. Deuxième configuration : classe = état. Moyenne sur tous les locuteurs. . . 158

11.1 Taux de reconnaissance *vs* quantité de données d'adaptation. Moyenne des 4 locuteurs. 167
11.2 Taux de reconnaissance *vs* quantité de données d'adaptation. Moyenne des 4 locutrices. 167

Introduction

Les performances des systèmes actuels de reconnaissance automatique de la parole (RAP) sont satisfaisantes lorsque les systèmes sont évalués sous des conditions contrôlées de laboratoire. Ainsi, sur une tâche de reconnaissance de parole continue, en mode dépendant du locuteur, pour un vocabulaire de 2 000 mots, [Gong, 1994] obtient moins de 1% d'erreur de reconnaissance. En reconnaissance de parole continue indépendante du locuteur, pour un vocabulaire de 20 000 mots, sur la tâche du *Wall Street Journal*, environ 5% d'erreurs sont obtenues [Aubert *et al.*, 1994].

Cependant, ces systèmes sont généralement peu robustes, c'est-à-dire que des variations du signal entre les conditions de test et d'apprentissage peuvent provoquer une dégradation significative des taux de reconnaissance, même si ces variations semblent minimes à l'oreille. Les principales sources de variabilité du signal, qui rendent difficile la conception de systèmes de RAP robustes, peuvent être classées selon leur provenance, qu'il s'agisse de l'environnement acoustique, de l'équipement d'acquisition du signal, ou encore du locuteur. Le signal est alors perturbé par le bruit ambiant (stationnaire ou non), les distorsions (linéaires ou non) provenant du canal de communication, et les habitudes articulatoires du locuteur. Notons que les séparations entre les différentes classes ne sont pas toujours nettes, l'environnement pouvant par exemple influencer le mode de production de parole. Le tableau 0.1 résume ces différentes sources de variabilité.

L'environnement perturbe le signal de parole sous la forme d'un bruit acoustique, que l'on suppose généralement additif. Cette hypothèse est souvent utilisée, à la fois pour sa simplicité, mais aussi car elle permet de couvrir un grand nombre de situations pratiques. Le signal enregistré est donc considéré comme la somme du signal de parole produit par le locuteur et du bruit ambiant. Les autres types de bruits, tels que les bruits électriques et bruits de quantification sont négligeables dans les applications de RAP. [Dautrich *et al.*, 1983a] sont parmi les premiers à constater la chute des performances d'un système de RAP entraîné dans des conditions calmes et testé dans le bruit : le taux d'erreur de reconnaissance du système, entraîné sur de la parole propre (SNR[1] > 40 dB) est multiplié par dix lors d'un test sur de la parole bruitée (SNR = 18 dB). Depuis, la littérature fournit pléthore d'observations analogues, et à titre d'exemple nous indiquons Figure 0.1, l'évolution des taux de reconnaissance en fonction du niveau de bruit présent lors du test, pour un système de reconnaissance de parole continue entraîné à partir de parole propre. Une telle évolution s'avère caractéristique du problème de la reconnaissance de la parole dans le bruit.

En plus des perturbations apportées par le bruit ambiant, le signal de parole est soumis à des distorsions spectrales provoquées par le canal d'acquisition. Dans le meilleur des cas, ces distorsions sont simplement linéaires, mais on rencontre également des distorsions non linéaires, beaucoup plus pénalisantes. Un changement de microphone ou de sa position, entre l'apprentissage et le test d'un système peut affecter de façon significative le spectre du signal,

[1] *Signal-to-Noise Ratio*, rapport signal-à-bruit

Environnement	– Bruit corrélé à la parole : réverbération, réflexion – Bruit non corrélé à la parole : bruit additif (stationnaire, non stationnaire)
Locuteur	– Attributs du locuteur : sexe, âge, dialecte. – Mode d'expression : soufflement, bruit des lèvres, stress, effet Lombard, rythme d'élocution, puissance sonore, fréquence fondamentale, locuteur coopératif.
Conditions d'enregistrement	– Microphone – Distance au micro – Filtrage – Matériel de transmission : distorsion, bruit, écho – Matériel d'enregistrement

TAB. 0.1 - *Principales causes de variabilité du signal de parole (d'après [Furui, 1992a])*

et dégrader ainsi les performances du système. Ainsi par exemple, [Acero et Stern, 1990] constatent que le taux de reconnaissance d'un système de RAP grand vocabulaire, valant initialement 85% s'effondre à 19% lors d'un changement de microphone entre l'apprentissage et le test.

Les variations de nature intra-locuteur sont en général moins pénalisantes que celles provoquées par le bruit ambiant ou le canal d'enregistrement. Cependant, en présence d'un bruit acoustique élevé, le locuteur modifie de façon réflexe son mode d'élocution (effet Lombard [Lombard, 1911]) pour que ses propos restent intelligibles. Cet effet provoque des distorsions importantes du signal de parole, et un système entraîné avec de la parole propre verra ses performances chuter en reconnaissance de parole Lombard [Junqua, 1989].

L'utilisateur d'une interface orale homme-machine sera naturellement réticent s'il doit parler d'une façon contrainte, si le système fonctionne mal le jour où il est enrhumé ou fatigué, ou bien encore si les performances s'effondrent en présence d'un bruit de fond inhabituel. Or, les systèmes de communication orale sont généralement utilisés dans des environnements bruités : usines, lieux publics, habitacle de voiture, d'avion, parole téléphonique, etc. Plus les systèmes de RAP seront robustes, plus le nombre de leurs applications potentielles augmentera, et l'absence de robustesse au bruit apparaît comme le principal obstacle au développement commercial de telles applications. L'amélioration de la robustesse des systèmes est donc un thème majeur de recherche, faisant appel à des connaissances pluridisciplinaires (traitement du signal, reconnaissance des formes, intelligence artificielle), essentiel pour permettre le développement d'applications en environnements réels.

Cette thèse s'insère dans ce cadre de recherche. Effectuer l'apprentissage d'un système

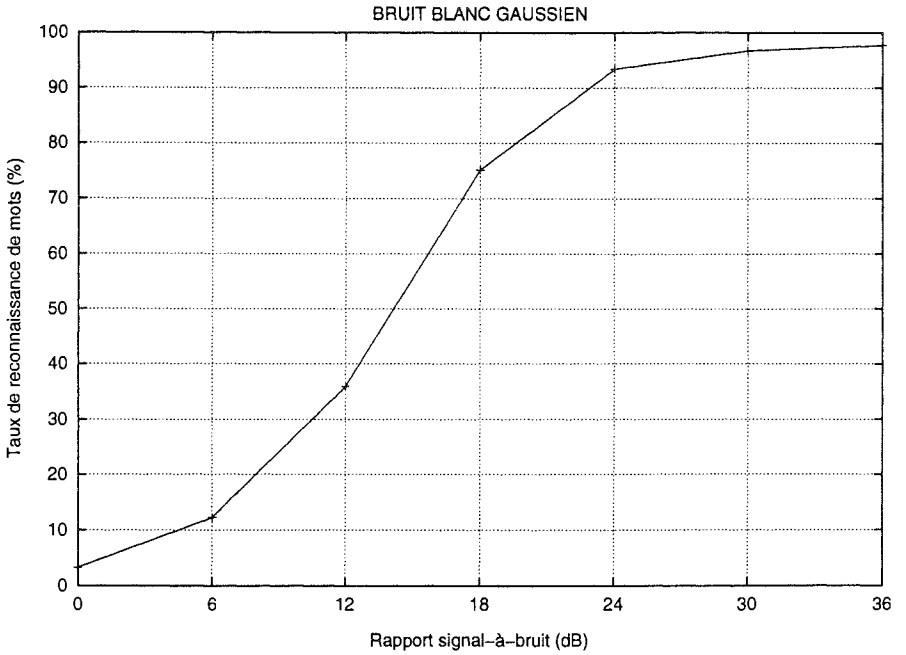

FIG. 0.1 - *Évolution du taux de reconnaissance d'un système de RAP entraîné en milieu calme (SNR >40 dB), en fonction du rapport signal-à-bruit lors du test [Siohan* et al., *1995].*

de reconnaissance de parole dans des conditions bruitées est une opération rarement envisageable en pratique. En effet, il est d'une part très coûteux d'enregistrer un corpus d'apprentissage dans le bruit. De plus, il est difficile de prévoir lors de l'apprentissage quelles seront les conditions de bruit lors de l'utilisation du système. Dans un tel cas, il devient également délicat d'autoriser une variation du SNR ou du type de bruit lors du test. Le problème à aborder est donc le suivant : étant donné un système de reconnaissance de parole continue entraîné à partir de parole propre, quelles méthodes et techniques peut-on mettre en œuvre pour améliorer la robustesse au bruit du système, c'est-à-dire pour que le système reconnaisse correctement de la parole prononcée en environnement réel, *a priori* inconnu ?

Nous présentons tout d'abord, partie I, une étude des différentes approches existantes, permettant d'améliorer la robustesse des systèmes de RAP dans le bruit. Ce travail met en évidence les problèmes spécifiques à la reconnaissance de parole dans le bruit, ainsi que les principales directions de recherches qui peuvent être développées. Trois approches majeures peuvent être distinguées. La première consiste à prétraiter le signal de parole bruitée afin de minimiser l'influence du bruit, dans le but d'utiliser un système de reconnaissance entraîné à partir de parole propre. La seconde effectue une transformation du système de reconnaissance, entraîné à partir de parole propre, afin de reconnaître directement la parole bruitée. La dernière approche se focalise sur la recherche d'un paramétrage du signal et de mesures de distorsions associées robustes aux variations d'environnement.

La partie II est consacrée à la présentation du système de reconnaissance de la parole continue VINICS, développé au CRIN-INRIA par Y. Gong [Gong et Haton, 1994; Gong, 1994]. Le système VINICS est constitué de deux parties distinctes. La première concerne la modélisation du signal acoustique de parole, et s'appuie sur les modèles stochastiques de

trajectoires de parole (STM[2]) [Gong et Haton, 1994], alternative efficace aux traditionnels modèles de Markov cachés (HMM[3]) [Rabiner, 1989]. Les méthodes de RAP dans le bruit sur lesquelles nous avons travaillé exploitent cette modélisation acoustique, ce qui justifie l'importance accordée à la présentation de ces modèles. Le décodage du signal constitue le deuxième élément de VINICS, dont l'objectif consiste à retrouver la séquence de symboles phonétiques la plus vraisemblable, étant donné les mesures de vraisemblance entre le signal et les différents modèles associés aux phonèmes.

Dans la partie III, nous proposons et évaluons trois approches pour améliorer la robustesse du système VINICS aux variations des conditions d'environnement, principalement axées sur le problème du bruit. Le chapitre 6 décrit une approche permettant d'estimer un STM de parole bruitée, à partir d'un HMM de bruit et d'un STM de parole propre. Cette méthode permet d'obtenir une approximation des modèles de parole bruitée les plus vraisemblables, sans nécessiter un apprentissage de ces modèles dans le bruit. Ces modèles sont ensuite utilisés pour reconnaître la parole bruitée. Un filtrage, spécifique à chaque état de chaque STM est présenté au chapitre 7. Ce filtrage est non linéaire, et est optimisé selon un critère significatif au niveau de la perception humaine. Les estimateurs sont calculés par une approche numérique, et les filtres sont mis en œuvre sous la forme de tables de transformations dans le domaine cepstral. L'objectif de ce filtrage est de permettre la reconnaissance de la parole dans le bruit, à partir d'un système construit à partir de parole propre. Il ne s'agit donc pas d'une approche pour restaurer le signal de parole propre, mais au contraire d'associer un filtrage à l'étape de reconnaissance du signal. Une méthode d'adaptation des STMs de parole propre aux variations des conditions d'environnement est proposée au chapitre 8. L'adaptation s'effectue par transformation linéaire des vecteurs moyennes des STMs. Les transformations sont spécifiques à des classes de sons, et sont déterminées selon un critère objectif à partir d'un corpus d'adaptation de parole bruitée de taille réduite (environ 20 secondes de parole). L'objectif de l'adaptation est de transformer des modèles entraînés dans des conditions données, dans le but d'obtenir des modèles permettant d'effectuer la reconnaissance dans des conditions différentes. Contrairement aux deux approches précédentes, la nature des variations de environnement n'est pas exploitée, et cette méthode est donc potentiellement applicable pour compenser des variations de locuteur, de canal d'acquisition du signal, ou encore d'ambiance acoustique. Le chapitre 9 conclut cette partie par une évaluation comparée des trois approches précédemment décrites. Les corpus de parole utilisés et les conditions expérimentales sont précisées. Les résultats obtenus sur une tâche de reconnaissance de la parole continue, en mode dépendant du locuteur, pour un vocabulaire d'un millier de mots sont présentés et discutés lorsque la parole est perturbée par différents types et niveaux de bruits additifs.

La partie IV décrit deux approches s'inscrivant également dans la recherche sur l'amélioration de la robustesse des systèmes de RAP. Cependant, ces approches ne peuvent être directement comparées avec les méthodes précédentes, ce qui justifie le développement d'une partie distincte. Le chapitre 10 est une étude sur la possibilité d'exploiter l'analyse linéaire discriminante pour mettre en œuvre un paramétrage du signal de parole robuste au bruit. Nous mettons en évidence que l'analyse linéaire discriminante permet de développer un paramétrage efficace pour la reconnaissance de la parole dans le bruit, en particulier lorsque le rap-

2. *Stochastic Trajectory Models*
3. *Hidden Markov Models*

port signal-à-bruit est faible. Ce paramétrage est évalué sur la même tâche de reconnaissance que précédemment, mais ici l'apprentissage et le test s'effectuent dans la même condition de bruit. Une telle stratégie de test n'est donc pas comparable avec les approches proposées dans la partie III. La robustesse aux variations du niveau de bruit est étudiée expérimentalement. Les variations du rythme d'élocution provoquées en présence d'effet Lombard sont prises en compte dans le chapitre 11, par une méthode d'adaptation des modèles de durée des phonèmes. L'objectif est ici de modifier les modèles de durée de phonèmes de VINICS, construits à partir de parole propre, afin de reconnaître de la parole Lombard. L'adaptation est effectuée sous le cadre général de l'apprentissage Bayésien.

Enfin, en conclusion de ce travail, nous résumons les méthodes proposées et les résultats obtenus, et présentons des perspectives d'études.

Première partie

Étude bibliographique

Introduction

L'insuffisance de robustesse des systèmes de RAP provient des variations statistiquement significatives entre le signal de parole utilisé lors de l'apprentissage des systèmes, et le signal de parole de test. Comme nous l'avons déjà noté, ces variations d'environnement sont principalement liées à la présence d'un bruit de fond, aux distorsions provoquées par le canal d'acquisition du signal, ou encore à l'effet Lombard, qui peuvent se manifester lors du test.

Par conséquent, l'ensemble des approches pour la reconnaissance de la parole dans des environnements difficiles se focalise sur la réduction des différences entre les conditions d'apprentissage et de test. Plusieurs articles synthétisant ces recherches ont été publiés ces dernières années [Juang, 1991; Furui, 1992b; Haton, 1993; Gong, 1995]. Cette partie complète ces études, et dresse un bilan d'environ 340 publications dans le domaine de la reconnaissance de la parole dans des environnements difficiles. La réduction des différences entre conditions de test et d'apprentissage peut être effectuée selon trois familles d'approches, qui, bien que se différenciant par leurs principes de base, conduisent au développement d'idées assez semblables.

On peut considérer d'une part que la configuration du système de reconnaissance est figée, et que par conséquent, la réduction des différentes distorsions s'effectue en traitant le signal de test. Ces traitements peuvent avoir pour objectif de filtrer le bruit, de compenser les variations de microphone entre test et apprentissage, ou encore de réduire les distorsions provoquées par l'effet Lombard. Ces méthodes sont présentées au chapitre 1, où l'accent est mis sur la suppression d'un bruit de fond additif.

D'autre part, il est possible de développer une approche duale, qui consiste à autoriser la présence d'un bruit ou d'une distorsion lors du test, en modifiant la configuration du système de reconnaissance. Le système évolue donc selon ses conditions d'utilisation. Le chapitre 2 décrit cette famille de travaux.

Enfin, on peut reporter l'amélioration de la robustesse sur la recherche d'un paramétrage du signal ainsi que de mesures de distances associées robustes. Ici encore, la robustesse peut se focaliser sur le bruit, les variations du canal d'enregistrement ou l'effet Lombard. Le système de reconnaissance est alors utilisé quelque soit l'environnement, sans modification de sa configuration. Ces approches sont présentées chapitre 3.

Le chapitre 5 conclut cette partie, en rappelant les éléments essentiels qui permettent d'améliorer la robustesse des systèmes de reconnaissance automatique de la parole aux variations des conditions d'environnement.

Chapitre 1

Transformation de la parole

Un système de RAP est utilisé de façon optimale lorsque ses conditions de test et d'apprentissage sont semblables. Étant donné un système entraîné à partir d'un corpus de parole propre, l'objectif du filtrage de bruit est de prétraiter le signal bruité de test, afin de pouvoir l'utiliser comme entrée du système.

La difficulté du filtrage de bruit sur un signal de parole perturbé par un bruit additif large bande réside dans deux points principaux. Tout d'abord, le bruit recouvre le signal à la fois dans le domaine temporel et fréquentiel, et ses caractéristiques sont *a priori* inconnues dans chacun des deux domaines. D'autre part, le signal de parole est non stationnaire ; les distorsions provoquées par le bruit varient au cours du temps et selon les fréquences. Un niveau de bruit constant large bande perturbe donc plus les zones de faible énergie du signal (sons non voisés, transitions) que les zones de fortes énergies (sons voisés). Ainsi, les sons les plus perturbés sont ceux ayant le moins de redondance à exploiter pour définir un filtrage.

Les objectifs du filtrage de bruit sont multiples : d'une part l'amélioration des aspect perceptifs du signal (qualité [1] et intelligibilité [2]), d'autre part l'augmentation de la robustesse au bruit des systèmes de RAP. Qualité et intelligibilité étant évaluées par des tests d'écoute et ne s'exprimant donc pas sous forme mathématique, il est difficile d'évaluer et de concevoir des méthodes permettant d'améliorer ces deux propriétés. De plus, l'amélioration de la qualité se traduit souvent par une dégradation de l'intelligibilité, et les méthodes destinées à l'amélioration de la qualité et l'intelligibilité n'améliorent pas systématiquement la robustesse des systèmes de RAP.

1 Soustraction spectrale

La soustraction spectrale fait partie des méthodes de filtrage du signal de type *Overlap and Add* (OLA). L'OLA consiste à effectuer tout d'abord une analyse spectrale du signal temporel dans des fenêtres successives se recouvrant. Le spectre à court terme de chaque fenêtre est

[1] La qualité est une mesure subjective, caractérisant l'aspect agréable de l'écoute d'un signal.

[2] L'intelligibilité est une mesure objective de la quantité d'informations que peut extraire un auditeur lors de l'écoute d'un signal, indépendamment de sa qualité. Un signal peut donc être de mauvaise qualité, tout en ayant une bonne intelligibilité, ou l'inverse.

ensuite calculé puis transformé selon une certaine stratégie, consistant par exemple à atténuer les effets du bruit. Le signal temporel est alors reconstruit dans chaque fenêtre, en utilisant la phase d'origine. Les fenêtres ainsi traitées sont enfin ajoutées les unes aux autres en respectant leurs décalages temporels initiaux afin de resynthétiser le signal.

La soustraction spectrale consiste à retrancher une estimation de la densité spectrale de puissance du bruit de la densité spectrale de puissance du signal bruité [Weiss *et al.*, 1974; Lim, 1978; Boll, 1979]. En considérant que bruit et parole sont des processus aléatoires non corrélés et stationnaires à court terme, la densité spectrale de puissance peut être estimée par le carré du module de la transformée de Fourier à court terme. La densité spectrale de puissance du bruit est calculée pendant les périodes d'absence de parole, et est utilisée pour filtrer toute la zone de parole qui suit. Le signal temporel est ensuite reconstruit dans chaque fenêtre d'analyse en utilisant la phase du signal bruité initial, l'oreille étant peu sensible aux distorsions de phases [Wang et Lim, 1982].

La différence entre l'estimation de la densité spectrale de puissance du signal bruité et celle du bruit peut devenir négative. Dans ce cas, elle est généralement remplacée par un seuil arbitraire positif de faible valeur. Ce seuillage provoque l'apparition de pointes spectrales qui vont et viennent de façon aléatoire sur le spectre, d'une trame à l'autre. Sur le signal temporel, ce phénomène se traduit par la présence d'une somme de sons purs de fréquences fondamentales aléatoires, d'où l'appellation de bruit « musical ». En dépit d'une énergie très faible par rapport à l'énergie initiale du bruit, ce bruit musical dégrade fortement l'intelligibilité du signal. Les améliorations de la soustraction spectrale visent essentiellement à limiter cet effet.

[Boll, 1979] effectue un lissage de l'amplitude spectrale du signal estimée entre plusieurs trames, afin d'atténuer les pointes spectrales du bruit musical. Dans le même ordre d'idée, [Whipple, 1994] effectue un filtrage passe-bas de l'image temps-fréquence du signal de parole, afin de supprimer les pointes spectrales.

La soustraction spectrale de Boll est utilisée par [Cairns et Hansen, 1992] dans un système temps-réel à base de HMMs discrets, associé à une compensation de l'effet Lombard (cf. § 7) pour une reconnaissance de 35 mots isolés. Une amélioration des performances est obtenue lors du test sur de la parole bruitée par un bruit blanc à 30 dB, mais aucune amélioration n'est obtenue pour un bruit de ventilateur d'ordinateur personnel. Cairns et Hansen expliquent ces mauvais résultats par la déficience de leur détecteur d'activité vocale.

[Berouti *et al.*, 1979] introduisent la soustraction spectrale généralisée, qui consiste à contrôler d'une part la quantité de bruit à soustraire (en multipliant l'estimation du bruit par une constante, appelée facteur de surestimation), à introduire ensuite différents degrés de non-linéarités dans la soustraction, et à utiliser enfin un seuillage correspondant à une fraction de l'estimation du bruit. Il faut cependant noter que le facteur de surestimation optimal (au sens de la distorsion spectrale d'Itakura-Saito [Gray *et al.*, 1980]) varie selon le son et le niveau de bruit. De plus, l'amélioration maximale du SNR ne correspond pas à l'obtention de la distorsion spectrale minimale [Kushner *et al.*, 1989].

[Van Compernolle, 1987; Van Compernolle, 1989a] constate que l'utilisation de la soustraction spectrale de Berouti *et al.* ne conduit pas à l'obtention de bons résultats en reconnaissance de parole lorsqu'on l'applique sur le signal de test et d'apprentissage. Il associe alors

à la soustraction spectrale une méthode de masquage de bruit (cf. § 5), et cette combinaison permet d'améliorer les performances.

[Lockwood et Boudy, 1991] utilisent un coefficient de surestimation du bruit fonction de la fréquence. L'idée est d'utiliser un coefficient de surestimation important dans les zones de fréquences où le SNR est faible, et inversement d'appliquer une faible surestimation du bruit lorsque le SNR est élevé. Le coefficient de surestimation est une fonction non linéaire du rapport signal-à-bruit instantané. La supériorité de la soustraction spectrale non linéaire sur le filtrage de Kalman est mise en évidence dans [Lockwood *et al.*, 1991]. La soustraction spectrale non linéaire combinée à un paramétrage MFCC[3], à l'utilisation de la mesure de projection cepstrale et à l'utilisation de HMMs continus avec lissage des matrices de covariance, conduit à l'obtention d'excellents résultats en reconnaissance de mots isolés dans un environnement de voiture [Lockwood et Boudy, 1991; Lockwood et Boudy, 1992].

[Hirsch et Ruhl, 1989] effectuent une soustraction spectrale sur les sorties d'un banc de filtres répartis sur l'échelle Mel, et obtiennent des résultats satisfaisants en reconnaissance de chiffres avec un système de reconnaissance basé sur l'alignement temporel dynamique (DTW[4]), mais pour un SNR modéré.

Une évaluation des différentes variantes de la soustraction spectrale [Weiss *et al.*, 1974; Lim, 1978; Berouti *et al.*, 1979; Boll, 1979] et du filtrage de Wiener implanté dans le domaine temporel et fréquentiel est effectuée dans [Ahmed, 1989]. L'évaluation porte sur l'amélioration de différentes mesures de distorsions fondées sur une analyse spectrale par LPC[5]. La supériorité du filtrage MMSE[6] fréquentiel par rapport aux méthodes concurrentes est mise en évidence pour l'ensemble des rapports signal-à-bruit testés.

Les performances de la soustraction spectrale sont comparées à celles d'un filtrage adaptatif utilisant l'algorithme de [Widrow *et al.*, 1975] (cf. § 2), et à celles d'un filtrage par antenne, dans le cadre de la reconnaissance de mots isolés avec un système basé sur la DTW [Fellbaum et Becker, 1991]. La soustraction spectrale ne donne de bons résultats qu'aux SNR élevés (> 15 dB), et les meilleurs résultats sont en définitive obtenus avec le filtrage par antenne. Il faut cependant remarquer que le test est effectué dans les conditions idéales d'un filtrage adaptatif où entrée de référence et entrée primaire sont enregistrées séparément afin d'éviter le phénomène de *cross-talk*.

[Van Compernolle, 1992] montre que la soustraction spectrale peut être interprétée comme un filtrage de Wiener à phase nulle. Le critère de calcul de la soustraction spectrale correspond à la minimisation de l'erreur quadratique moyenne entre le spectre à court terme de parole propre et son estimation. La simplicité du calcul explique le succès de cette méthode de filtrage, bien que le critère d'estimation ne corresponde pas à des critères perceptifs. En effet, il semble préférable de minimiser les distorsions dans le domaine logarithmique du spectre. Malheureusement, selon la nature des fonctions de densité de probabilité (*pdf*[7]) de parole et du bruit, la minimisation des distorsions dans le domaine log-spectral ne conduit pas toujours à l'obtention d'un estimateur sous forme close ; il est alors nécessaire

3. *Mel Frequency Cepstral Coefficient*
4. *Dynamic Time Warping*
5. *Linear Predictive Coding*
6. *Minimum Mean Square Error*
7. *Probability Density Function*

de procéder à des approximations ou à des résolutions numériques [Ephraim et Malah, 1985; Van Compernolle, 1989b; Erell et Weintraub, 1990]. Sur des tâches de reconnaissance de parole et d'amélioration de la qualité du signal, la minimisation des distorsions dans le domaine du logarithme du spectre s'avère plus efficace que dans le domaine linéaire spectral.

L'efficacité de la soustraction spectrale est fortement conditionnée par la qualité de l'estimation de la densité spectrale de puissance du bruit, obtenue à partir du signal de parole bruitée. En particulier, il est indispensable de disposer d'une bonne détection parole/non-parole [McAulay et Malpass, 1980; Van Compernolle, 1989a; Fellbaum et Becker, 1991]. Dans [Puel et André-Obrecht, 1994], deux méthodes de segmentation parole/non-parole sont présentées. L'une d'elle se base sur une analyse temporelle du signal et exploite le fait que l'abscisse curviligne d'un signal de parole bruitée augmente plus rapidement dans les zones de parole voisée que dans les zones de bruit. L'autre effectue une analyse fréquentielle du signal, et considère que le spectre du bruit est une courbe monotone dont la dérivée est faible tandis que le spectre de la parole bruité présente une forte dynamique. L'analyse temporelle semble donner les meilleurs résultats aux SNR faibles (0 dB), tandis que la meilleure segmentation parole/non-parole est obtenue avec l'analyse fréquentielle pour des SNRs de 15 à 30 dB.

2 Annulation adaptative de bruit

Contrairement à la soustraction spectrale qui permet de filtrer le bruit d'un signal enregistré avec un seul microphone, l'annulation adaptative de bruit (ANC[8]) nécessite l'utilisation de deux microphones pour supprimer le bruit [Widrow *et al.*, 1975]. Un des microphones, appelé entrée primaire, effectue l'acquisition du signal de parole bruitée ($s + d_1$). Le second microphone, ou entrée de référence, enregistre un bruit d_2, corrélé au bruit se superposant à la parole (cf. fig. 1.1). L'objectif du filtrage consiste à calculer le filtre adaptatif H^*, permettant d'estimer le bruit perturbant l'entrée de référence, afin de le soustraire du signal bruité. Le calcul du filtre s'effectue généralement dans le domaine temporel, en utilisant l'algorithme LMS[9] [Widrow *et al.*, 1975]. Le filtrage est d'autant plus efficace que la source de parole ne perturbe pas l'entrée de référence et que le bruit présent à l'entrée primaire est le plus corrélé possible avec celui de l'entrée de référence. Différentes variantes de l'algorithme LMS existent cependant lorsque la parole est présente sur les deux entrées (phénomène de *cross-talk*), et sont présentés dans [Le Bouquin, 1991]. Le lecteur trouvera dans [Baudois *et al.*, 1989] un large panorama des différents algorithmes d'estimation du filtre adaptatif. Dans [Feder *et al.*, 1987; Feder *et al.*, 1988] un cadre général est proposé pour l'estimation de filtres adaptatifs en utilisant l'algorithme EM[10] [Dempster *et al.*, 1977].

Le grand intérêt de l'ANC est qu'elle ne nécessite pas de connaître *a priori* les statistiques du bruit ou de la parole. De plus, et contrairement à la soustraction spectrale, cette méthode est applicable à des bruits non stationnaires. Malheureusement, les performances de l'ANC sont fortement affectées par les choix d'implantation, en particulier les positions respectives des

8. *Adaptive Noise Cancelation*
9. *Least Mean Square*
10. *Expectation-Maximisation*

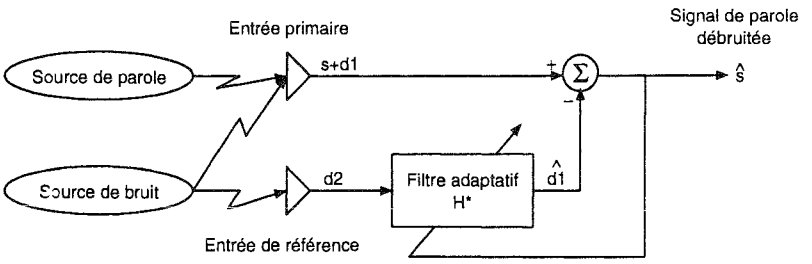

FIG. 1.1 - *Schéma de principe de l'annulation adaptative de bruit.*

microphones. [Dabis et Wrench, 1991] obtiennent une amélioration des taux de reconnaissance en utilisant un filtrage ANC, mais dans des conditions idéales de test où les positions des sources de parole et bruit par rapport aux microphones sont contrôlées. [Dal Degan et Prati, 1988] constatent que l'ANC échoue dans un environnement de voiture, à cause de la faible cohérence du bruit entre l'entrée primaire et de référence. [Lecomte *et al.*, 1989] rapportent des observations analogues et concluent que l'ANC n'est pas applicable dans un tel environnement.

Dans les environnements fortement bruités où l'estimation *a priori* des caractéristiques de la parole et du bruit sont peu précises, l'ANC fournit cependant une suppression effective du bruit [Boll et Pulsipher, 1980].

Les filtrages de bruit utilisant un nombre de microphones supérieur à deux et fondés sur les techniques de formation de voies sont encore marginaux dans la communauté reconnaissance de parole, principalement à cause de l'absence de corpus de parole enregistrés dans de telles conditions. [Van Compernolle *et al.*, 1990] obtiennent de bons résultats de reconnaissance de parole en utilisant une matrice de microphones et la méthode de formation de voies de [Griffiths et Jim, 1982], en présence de bruits stationnaires ou de bruits de parole d'un autre locuteur. [Giuliani *et al.*, 1994] observent qu'une méthode de formation de voies utilisant quatre microphones permet d'obtenir de meilleurs taux de reconnaissance dans le bruit, par rapport à la combinaison d'une soustraction spectrale et d'une normalisation cepstrale (cf. chap. 3, § 4) n'utilisant qu'une seule voie. La combinaison des trois approches conduit à une amélioration supplémentaire des taux de reconnaissance. [Lin *et al.*, 1994] combinent un filtrage à base de matrice de microphones à une transformation d'espace par réseau de neurones (cf. § 3). La matrice de microphones est utilisée pour permettre l'enregistrement du signal dans un environnement bruyant et réverbérant (p.ex. salle de conférence), et le réseau de neurones transforme l'espace cepstral du signal filtré vers l'espace cepstral du microphone de proximité ayant servi à entraîner le système de RAP.

L'introduction de contraintes auditives dans le processus de filtrage permet d'améliorer la qualité des traitements. [Nandkumar et Hansen, 1992; Nandkumar et Hansen, 1994] développent un filtrage à deux voies, basé sur EM, utilisant des contraintes auditives. Leur méthode incorpore les filtres en bandes critiques, la conversion d'intensité en sonie et l'inhibition latérale dans le processus de filtrage (cf. chap. 3, page 45). De plus, les contraintes appliquées sont spécifiques à des classes phonétiques grossières (sons voisés, non voisés, transitions). Évaluée sur un grand corpus, cette méthode permet d'améliorer la qualité sur toutes les zones

du signal, tout en maintenant l'intelligibilité.

[Sullivan et Stern, 1993] s'inspirent du traitement binaural humain et proposent un filtrage multivoies qui utilise en entrée les sorties du banc de filtres de [Seneff, 1988] (cf. chap. 3, § 3), et exploite les corrélations entre les voies. Par rapport à un traitement similaire sur un signal monovoie, l'utilisation de plusieurs voies permet d'améliorer les taux de reconnaissance sur une tâche de reconnaissance de mots isolés.

3 Transformation d'espace

L'objectif de la transformation d'espace est de définir une transformation permettant de recouvrer la parole propre dans le domaine temporel ou dans un espace de paramètres, à partir de la parole bruitée. Contrairement aux approches fondées sur le filtrage de bruit (cf. § 1 et § 2), la transformation est déterminée sans présumer de la nature de la combinaison entre parole et bruit, c.-à-d. sans connaître la nature des différences entre l'espace de parole de référence et l'espace de parole de test. Ces méthodes permettent donc de prendre en compte des différences entre conditions de test et d'apprentissage, que ces différences proviennent d'un changement de locuteur, de bruit d'environnement ou bien de microphone.

La transformation, c.-à-d. la correspondance entre les espaces de référence et de test, est généralement établie à partir de l'observation d'une même séquence de parole dans les deux environnements. Dans le cas d'une adaptation au locuteur, les locuteurs de test et de référence doivent prononcer un même ensemble de phrases. Dans le cas d'une adaptation à un environnement bruité (bruit additif), la séquence de parole dans l'espace de test s'obtient directement en ajoutant le bruit à la parole de référence. L'ensemble des données utilisées pour déterminer la transformation est appelé corpus d'adaptation. Bien évidemment, on recherche en général à utiliser un corpus d'adaptation le plus réduit possible.

Une première famille de transformations se base sur l'utilisation d'une correspondance une à une entre les vecteurs du répertoire de prototypes (*codebook*[11]) de l'espace de référence et ceux de l'espace de test (cf. § 3.1). Dans une seconde famille de méthodes, des transformations analytiques explicites peuvent être définies selon un critère objectif pour projeter un espace sur un autre (cf. § 3.2). Enfin, les transformations d'espaces peuvent être mises en œuvre par des réseaux de neurones (cf. § 3.3).

3.1 Transformation de *codebook*

Les transformations de *codebook* sont issues des travaux sur l'adaptation au locuteur des systèmes de RAP à base de DTW ou de HMMs discrets [Shikano *et al.*, 1986; Nakamura et Shikano, 1989]. La transformation s'effectue généralement de la façon suivante. Les répertoires de prototypes de l'espace de test et de référence sont générés à l'aide de l'algorithme de Lloyd [Linde *et al.*, 1980] à partir des corpus de parole représentatifs des deux espaces acoustiques. Une quantification vectorielle de chaque mot du corpus d'apprentissage est effectuée trame par trame. Un mot est alors caractérisé par la séquence des indices des vec-

[11] Un *codebook* est un ensemble fini de vecteurs, ou prototypes, représentatif d'un espace vectoriel donné.

teurs du *codebook* qui le constitue. La correspondance optimale entre les vecteurs des mêmes mots prononcés dans les environnements de test et de référence est déterminée par DTW. On obtient alors un ensemble de couples de vecteurs en correspondance. Lorsque le corpus de test correspond à une version bruitée du corpus de référence, la correspondance est immédiate. Ensuite, une matrice des correspondances (appelée histogramme) entre chaque vecteur i du corpus de référence et chaque vecteur j du corpus de test est construite. Chaque vecteur du *codebook* de test (respectivement référence) est remplacé par une combinaison linéaire des vecteurs du *codebook* de référence (respectivement test), dont les poids sont fournis par l'histogramme des correspondances, selon que l'on souhaite effectuer une transformation de l'espace de référence vers l'espace de test ou inversement de l'espace de test vers l'espace de référence. [Roe, 1987] applique une telle méthode sur des tâches simples de reconnaissance de chiffres dans le bruit, et divise ainsi le taux d'erreur de reconnaissance par 4.

[Ohkura et Sugiyama, 1991] utilisent une méthode de transformation de *codebook* similaire, associée à une quantification vectorielle floue [Tseng *et al.*, 1987] pour obtenir une transformation continue entre les deux espaces. Les taux de reconnaissance sont améliorés d'environ 20% sur une application de reconnaissance de mots isolés pour un SNR de 20 dB en présence d'un bruit rose. Ohkura et Sugiyama comparent leur transformation de *codebook* avec une méthode de filtrage du signal temporel utilisant des réseaux de neurones. Sur une tâche de reconnaissance des phonèmes /b,d,g/, et en présence d'un bruit de ventilateur pour un SNR de 5 dB, la transformation par réseaux de neurones s'avère légèrement plus efficace.

[O'Shaughnessy, 1988] recherche à améliorer l'intelligibilité d'un signal de parole bruitée en resynthétisant la parole à l'aide d'un ensemble de prototypes obtenu par quantification vectorielle de l'espace LPC de la parole propre. La distance entre parole propre et bruitée utilise les fréquences et largeurs de bande des formants, relativement robustes au bruit. La parole resynthétisée ne comporte donc pas de bruit (car elle est construite à partir d'un ensemble d'observations de parole propre), mais une distorsion spectrale liée à une mauvaise classification des trames bruitées.

Dans [Juang et Rabiner, 1987], une correspondance une à une est obtenue entre trames de parole propre et bruitée à partir d'un corpus de parole et de sa version bruitée. Étant donné un vecteur de parole bruitée, un ensemble de trames bruitées, proches au sens de la distorsion d'Itakura-Saito [Gray *et al.*, 1980], est déterminée. La moyenne des trames propres associées à cet ensemble de trames bruitées fournit une version filtrée du vecteur initial. Une amélioration de la distorsion d'Itakura-Saito correspondant à un gain de 10 dB est obtenue à partir d'un signal perturbé par un bruit Gaussien à un SNR de 14 dB. Une méthode relativement proche avait été utilisée auparavant par [Porter et Boll, 1984].

Dans les travaux de [Gong, 1993], une base est associée à un espace de parole propre. Un vecteur de l'espace s'exprime comme une combinaison linéaire des vecteurs de la base. Les vecteurs de la base sont obtenus à partir de phonèmes issus d'un corpus d'adaptation étiqueté de taille réduite (20 secondes de parole). Une version propre et une version bruitée du corpus d'adaptation permettent d'obtenir une base d'un espace de parole propre et la base correspondante dans l'espace de parole bruitée. Un vecteur de parole bruitée, combinaison linéaire des éléments de la base bruitée, est transformé en remplaçant la base bruitée par la base propre, les coefficients de la combinaison linéaire restant inchangés. Sur une tâche de reconnaissance de 206 mots isolés en présence d'un bruit Gaussien, à un SNR de 10 dB, un

taux de reconnaissance de 90% est obtenu, le taux de reconnaissance en parole propre étant 95%. Une base est cependant spécifique à une condition de bruit donnée. Cette méthode est alors étendue pour s'affranchir de la dépendance au type et au niveau de bruit [Treurniet et Gong, 1994]. Différentes bases sont déterminées *a priori* pour différentes conditions de bruit. Lors de la reconnaissance, l'identification de la condition de bruit est d'abord effectuée. La parole bruitée est alors transformée avec la base associée au bruit reconnu. Sur la même tâche de reconnaissance que précédemment, cette méthode offre des performances meilleures que celles obtenues par combinaison de modèles (cf. chap. 6) et par entraînement à partir de parole bruitée [Siohan *et al.*, 1994].

[Furui, 1989a; Furui, 1989b] effectue une mise en correspondance de répertoires de prototypes, dans un but d'adaptation au locuteur, en utilisant une approche non supervisée basée sur une classification hiérarchique des vecteurs de cepstre. Son approche minimise la distorsion de quantification vectorielle entre l'espace acoustique de référence et l'espace acoustique de test. Cette transformation est utilisée dans [Cung et Normandin, 1992; Cung et Normandin, 1993] sur une application de reconnaissance de parole bruitée, ainsi que dans [Fissore *et al.*, 1992] pour effectuer une adaptation des prototypes d'un HMM discret aux variations de canal d'enregistrement.

3.2 Transformation analytique d'espace

Un autre type de transformation d'espace consiste à rechercher une transformation analytique, calculée de façon optimale selon un certain critère. Là encore, la plupart de ces méthodes proviennent des travaux sur l'adaptation au locuteur. [Gu et Mason, 1989] définissent une transformation linéaire pour projeter les vecteurs de parole de l'espace de test sur l'espace de référence. La transformation est définie à partir d'un ensemble de couples de vecteurs provenant de l'alignement temporel des mêmes phrases prononcées dans l'espace de test et de référence. La transformation, obtenue par régression linéaire multiple, minimise l'erreur quadratique moyenne entre les vecteurs de référence et les vecteurs de test transformés. La transformation étant déterminée, il semble intéressant de corriger l'alignement initial en réalignant le corpus d'adaptation de l'espace de test transformé sur le corpus de référence [Gong *et al.*, 1992], puis de réestimer une nouvelle transformation.

L'approche de Gu et Mason est utilisée dans [Mokbel, 1992] pour transformer le corpus de référence d'un système de RAP à base de DTW, afin de le rendre plus proche des conditions bruitées de test. Sur une application de reconnaissance de mots isolés dans un environnement de voiture, la transformation du corpus de référence fournit des taux de reconnaissance supérieurs à ceux obtenu par filtrage de Kalman [Mokbel et Chollet, 1991a]. Dans [Mokbel *et al.*, 1992b], cette transformation d'espace conduit à l'obtention de taux de reconnaissance supérieurs à ceux obtenus par une soustraction spectrale non linéaire. Il semble également préférable de transformer les données de référence plutôt que celles de test, mais aucun autre travail ne confirme, à notre connaissance, cette assertion. Dans [Mokbel *et al.*, 1992a; Mokbel *et al.*, 1992b], une transformation d'espace calculée de façon semblable est utilisée pour transformer les moyennes et variances de HMMs entraînés à partir de parole propre, afin de reconnaître la parole bruité.

[Seide et Mertins, 1994] utilisent une transformation non linéaire entre un espace de pa-

role propre et un espace de parole bruitée. Les vecteurs des deux espaces sont complétés par des combinaisons polynomiales de leurs composantes. La transformation non linéaire est en fait une transformation linéaire calculée par régression linéaire multiple à partir des vecteurs ainsi étendus. Ce traitement divise par 5 le taux d'erreur de reconnaissance des 10 chiffres, perturbés par un bruit Gaussien à un SNR de 10 dB.

Étant donné que le bruit ne perturbe pas toutes les zones du signal de parole de la même façon, il est intéressant d'appliquer des transformations d'espace spécifiques à différents sous-espaces, plutôt qu'une transformation globale. [Gu et Mason, 1989] décomposent l'espace de test en classes par l'algorithme de Lloyd [Linde *et al.*, 1980]. À chaque classe est associée une transformation linéaire, calculée par régression linéaire multiple. Lors de la reconnaissance, la classe associée au vecteur de test courant est d'abord identifiée, la transformation spécifique à cette classe est ensuite appliquée.

[Neumeyer et Weintraub, 1994] étendent la méthode de Gu et Mason et définissent une transformation probabiliste d'espace, qui consiste tout d'abord à décomposer l'espace de référence en classes avec l'algorithme de Lloyd. Ensuite, une transformation linéaire estimée par régression linéaire multiple est associée à chaque classe de l'espace. Le vecteur transformé correspond alors à la somme de chaque transformation linéaire appliquée au vecteur, pondérée par la probabilité que le vecteur appartiennent à chaque classe. Cette même transformation est utilisée avec succès pour reconnaître de la parole téléphonique à partir d'un système entraîné sur de la parole propre de bonne qualité [Weintraub et Neumeyer, 1994]. Un algorithme équivalent est utilisé par [Gish *et al.*, 1990; Ng *et al.*, 1992], mais Gish *et al.* conditionnent en plus les transformations par le SNR instantané.

3.3 Transformation d'espace par réseaux de neurones

Dans de nombreux travaux, des transformations complexes d'espaces sont mises en œuvre par des réseaux de neurones (RN). Les variations entres ces travaux se situent principalement au niveau de la nature des espaces à transformer. [Tamura et Waibel, 1988; Tamura, 1989] utilisent des RNs pour effectuer une réduction de bruit directement sur le signal temporel, et obtiennent une amélioration de la qualité du signal par rapport à un traitement à base de soustraction spectrale. L'intelligibilité n'est cependant pas améliorée. [Ohkura et Sugiyama, 1991] transforment également le signal temporel et obtiennent des résultats en reconnaissance de mots isolés légèrement meilleurs que ceux obtenus par transformation de *codebook*.

[Sorensen, 1991] choisit de transformer des espaces de paramètres cepstraux, tout comme [Barbier et Chollet, 1991; Barbier, 1992] qui obtiennent de bons résultats en reconnaissance des 10 chiffres dans un environnement de voiture.

L'espace de paramètres issus d'un modèle auditif [Gao *et al.*, 1992b] est transformé dans [Gao et Haton, 1993] en utilisant un RN, et les vecteurs transformés sont utilisés en entrée d'un système de RAP à base de LPNNs[12]. Dans [Gao et Haton, 1994], le RN servant au filtrage du bruit et les LPNNs sont intégrés dans une structure unifiée, entraînée selon un critère d'optimisation conjointe de la réduction de bruit et de la prédiction du signal. La

[12] *Linked Predictive Neural Networks*

réduction de bruit n'est donc plus indépendante de la partie du système assurant la reconnaissance. De plus, le filtrage étant directement associé aux différents modèles LPNNs, la transformation d'espace s'effectue de façon spécifique à chaque modèle.

D'après [Trompf, 1992], l'introduction du contexte de la trame courante en entrée du RN permet d'améliorer les performances en reconnaissance, particulièrement pour les SNR faibles.

Dans [Paliwal, 1990], les performances de différents classifieurs entraînés sur de la parole propre (Perceptrons multi-couches, k-moyennes, classifieur du maximum de vraisemblance) sont comparées sur une tâche de reconnaissance de voyelles dans le bruit, pour différents paramétrages du signal. Les meilleurs résultats sont obtenus en utilisant les paramètres MFCC associés à un Perceptron multi-couches.

Étant donné leur capacité d'approximation de n'importe quelle fonction, les RN permettent d'obtenir des résultats satisfaisants en transformation d'espace. Les RN n'utilisant aucun modèle de bruit ou de parole, ils demeurent cependant peu robustes aux variations du niveau et du type de bruit. [Trompf, 1992] propose d'effectuer un apprentissage des RN en utilisant plusieurs SNRs différents. Pour des bruits stationnaires, cette méthode donne des résultats satisfaisants sur une application de reconnaissance de 30 mots isolés.

Dans [Xie et Van Compernolle, 1994], les données d'entraînement d'un RN sont générées à partir de modèles stochastiques de parole et de bruit. Lors de l'apprentissage, les données de parole bruitée sont appliquées à l'entrée du RN et les données de parole propre en sortie. De plus, les moyennes et variances de la parole propre et du bruit sont utilisées comme entrées supplémentaires du réseau, ce qui permet d'éviter un surentraînement du réseau sur une condition de bruit donnée.

4 Exploitation de la structure harmonique du signal de parole

Certaines méthodes de filtrage exploitent la structure harmonique du signal de parole, et en particulier la périodicité des sons voisés. De telles méthodes sont donc limitées et ont généralement un comportement médiocre sur les zones non voisées du signal. De plus, l'estimation et le suivi de la fréquence fondamentale du signal devient délicate en présence de bruit. En présence d'un bruit large bande, la distorsion affectant les zones non voisées de faible énergie sera forte par rapport à celle affectant les zones voisées, et ces méthodes n'améliorent généralement ni la qualité ni l'intelligibilité du signal. Par contre, ces techniques sont plus utiles pour filtrer un bruit additif à bande étroite, ou bien des bruits de parole [O'Shaughnessy, 1989].

[Sambur, 1978] développe une méthode de filtrage adaptatif basée sur l'ANC, en utilisant une seule voie d'acquisition du signal bruité. À l'entrée de référence du filtre adaptatif, il injecte une version du signal bruité retardée d'une ou deux périodes de la fréquence fondamentale. L'entrée primaire est constituée du signal de parole bruitée. Comme les composantes de parole du signal bruité et du signal retardé sont fortement corrélées, et que les composantes de bruit entre l'entrée primaire et de référence sont décorrélées, le filtre adaptatif fournit en

sortie une estimation du signal de parole propre.

La structure harmonique des sons voisés peut être exploitée pour filtrer le bruit à l'aide d'un filtre en peigne [Shields, 1970]. Ce filtrage consiste à ne conserver que les harmoniques du signal, peu perturbées par le bruit, tout en rejetant les composantes fréquentielles situées entre ces harmoniques. Le filtrage peut être mis en œuvre dans le domaine fréquentiel à l'aide d'une séquence de distributions de Dirac espacées de la fréquence fondamentale, ou bien dans le domaine temporel, où un segment de parole d'une durée égale à la période de la fréquence fondamentale est filtré par moyennage avec les segments adjacents.

[Perlmutter *et al.*, 1977; Lim, 1978] évaluent l'influence du filtrage en peigne sur le SNR de plusieurs signaux de parole perturbés par différents bruits. Si l'amélioration du SNR est effective, le filtrage réduit cependant l'intelligibilité, à cause de la non-stationnarité du signal de parole.

Lorsque l'interférence perturbant le signal est structurée (p.ex. bruit périodique), le filtrage en peigne traditionnel n'atténue que faiblement le bruit. [Malah et Cox, 1982; Cox et Malah, 1981] développent un modèle généralisé de filtrage en peigne consistant à pondérer chaque segment associé à une période de la fréquence fondamentale par une fonction du temps et non plus par une constante comme dans le cadre du filtre en peigne classique. L'atténuation d'un bruit à structure périodique est alors possible.

Les performances du filtre en peigne se dégradent lorsque la fréquence fondamentale varie rapidement. [Frazier *et al.*, 1976] proposent une méthode adaptative permettant de prendre en compte de telles variations. Dans des travaux plus récents, [Graf et Hubing, 1993] améliorent le filtre de Frazier *et al.* en effectuant un alignement temporel dynamique du segment courant sur les segments adjacents, afin de corriger les variations de la fréquence fondamentale de segment en segment.

[Ramalho et Mammone, 1994] utilisent une idée assez proche et appliquent un filtre en peigne sur le signal temporel expansé (c.-à-d. rendu périodique). Après filtrage, le signal expansé est compressé pour retrouver la quasi-périodicité initiale. Sur une tâche d'identification du locuteur, ce filtrage améliore de façon significative la robustesse au bruit du système.

[Erell et Weintraub, 1994] exploitent la quasi-périodicité du signal de parole pour estimer la log-amplitude de la transformée de Fourier du signal, selon un critère MMSE. Leur algorithme est une amélioration des travaux de [Porter et Boll, 1984], et consiste à prendre en compte les corrélations entre les composantes spectrales : d'une part les corrélations large bande déterminées par le conduit vocal, d'autre part les corrélations en bande étroite, introduites par la quasi-périodicité de l'excitation. Les contraintes spectrales large bande améliorent l'estimation sur tout le signal, et l'utilisation de la structure périodique des sons voisés améliore l'estimation des zones voisées. Cependant, comme les zones de basses fréquences des sons non voisés de faible énergie sont les plus perturbées par le bruit, l'amélioration de l'estimation des zones voisées a un effet négligeable sur le taux de reconnaissance.

5 Masquage de bruit

En présence d'un bruit de fond large bande, les régions du spectre de faible énergie sont plus affectées que les zones de forte énergie. Il est donc difficile de définir une mesure de distance entre spectres, les zones les plus perturbées étant celles où la mesure de distance est la moins fiable. [Klatt, 1976] introduit alors le masquage de bruit en sortie d'un banc de filtres, qui consiste à remplacer les zones du spectre dont l'énergie est inférieure à un certain seuil (caractérisant le niveau maximal du bruit de fond), par ce seuil. Avant masquage, les zones de faible énergie porteuses de peu d'informations intervenaient dans les calculs des distances entre spectres ; après masquage, les zones masquées aux mêmes fréquences sur deux spectres différents ont la même valeur et n'interviennent donc plus dans le calcul de la distance entre ces spectres. On supprime ainsi l'influence des régions du spectre portant peu d'informations. Bien évidemment, lorsque le niveau de bruit devient trop important, un trop grand nombre de régions du spectre sont masquées, et la distance entre spectres n'est plus significative [Furui, 1992b].

Le masquage de Klatt est reformulé dans le cadre d'un système de RAP à base de HMM dans les travaux de [Varga *et al.*, 1988; Varga et Ponting, 1989]. Le problème est alors d'obtenir l'estimation de la vraisemblance d'une observation de parole bruitée dans un état donné du HMM, étant donné une connaissance sur le bruit perturbateur. Classer une observation comme parole ou bruit est une décision binaire, basée sur un seuil. Le choix de ce seuil dépend de la nature de l'estimation du bruit, de la nature du bruit et de la relation liant parole et bruit. Ainsi, l'utilisation d'un seuil relatif au niveau moyen du bruit introduit un compromis entre accepter comme parole des informations de bruit, et rejeter des observations de parole en les considérant comme bruit. Le choix optimal du seuil dépend donc de la variance du bruit et de la parole, et de la « distance » entre distributions de bruit et de parole. Différentes utilisations de la connaissance *a priori* du bruit [Holmes et Sedgwic, 1986; Bridle *et al.*, 1984] sont discutées dans [Varga *et al.*, 1988; Varga et Ponting, 1989]. L'approche de Klatt conduit à l'obtention des meilleurs résultats.

Le masquage de bruit est formulé sous un cadre probabiliste par [Nadas *et al.*, 1989; Van Compernolle, 1989a]. Nadas *et al.* modélisent le spectre de la parole bruitée par un vecteur aléatoire correspondant au maximum du spectre de parole propre et du spectre de bruit. La *pdf* du spectre bruité est alors estimée à partir de la *pdf* du bruit déterminée pendant les pauses, et de la *pdf* de la parole propre estimée à partir du corpus d'apprentissage. Cette compensation ne perturbe pas les performances du système de reconnaissance lorsque les conditions de test et d'apprentissage sont semblables, mais améliore de façon significative les performances lorsque le système est entraîné sur de la parole propre, puis testé dans le bruit. Cette méthode est à l'origine des approches de compensation de modèles (cf. chap. 2, § 3).

[Usagawa *et al.*, 1994] exploitent les notions de masquage simultané[13] et proactif[14], et utilisent des connaissances psychoacoustiques pour définir les gabarits des masques. Associé

[13] « Lorsqu'on entend simultanément deux sons purs de fréquences différentes, il arrive que l'un d'eux devienne inaudible » [CALLIOPE, 1989], c'est l'effet de masquage simultané.

[14] « Si deux sons purs sont présentés non pas simultanément, mais séparés par un bref silence, ils peuvent se masquer mutuellement. » [CALLIOPE, 1989] Lorsque le son masquant précède le son masqué, le masquage est dit proactif.

à un filtrage ANC, ce prétraitement s'avère efficace jusqu'à −15 dB pour une tâche de reconnaissance très simple (commande vocale d'un téléviseur en mode dépendant du locuteur).

Dans les approches précédentes, le masquage était effectué dans le domaine de l'énergie (sortie d'un banc de filtres). Les systèmes de reconnaissance de parole fonctionnant mieux avec un paramétrage cepstral, [Mellor et Varga, 1993] implantent un masquage de bruit pour un paramétrage MFCC. Les performances obtenues par masquage du cepstre restent cependant équivalentes à celles obtenues par masquage des sorties d'un banc de filtres. Par rapport à la méthode de décomposition de modèles de [Varga et Moore, 1990] (cf. chap. 2, § 1), le masquage offre des performances inférieures, en particulier pour des SNRs faibles, mais au profit d'une complexité moindre.

6 Estimation à base de modèles

Le signal de parole peut être représenté à court terme par un modèle auto-régressif (AR). À partir du signal de parole bruitée, [Lim et Oppenheim, 1978; Lim et Oppenheim, 1979; Lim et Oppenheim, 1983] cherchent à estimer les paramètres du modèle AR de parole propre ainsi que le signal de parole propre, en utilisant une méthode d'estimation du maximum *a posteriori* (MAP). Une estimation sous optimale est obtenue à l'aide de l'algorithme EM [Dempster *et al.*, 1977], par un procédé itératif. Les paramètres du modèle AR sont tout d'abord déterminés en supposant le signal propre connu. Une estimation du signal propre est ensuite calculée en utilisant le modèle propre estimé et en supposant connue l'estimation de la densité spectrale de puissance du bruit. Cependant, si un vecteur de parole propre ainsi que le modèle AR sont estimés à partir du vecteur correspondant de parole bruitée, le nombre d'inconnues (nombre d'échantillons de parole propre + nombre de coefficients du modèle AR) est supérieur au nombre d'observations bruitées. Il est donc impossible d'obtenir une faible variance à la fois pour l'estimateur des paramètres du modèle AR et pour l'estimateur du signal propre. L'utilisation des trames adjacentes pour filtrer une trame donnée permet d'éviter ce problème [Hansen et Clements, 1987; Hansen et Clements, 1991]. La méthode de Lim et Oppenheim est limitée au bruit blanc Gaussien ; [Hansen et Clements, 1985] l'étendent pour des bruits colorés. Une variante de l'approche de Lim et Oppenheim est développée dans [Paliwal et Basu, 1987], où le modèle AR est d'abord estimé à partir du signal bruité, puis utilisé pour estimer le signal propre.

L'algorithme de Lim et Oppenheim provoque une réduction de la largeur de bande des formants, ainsi que des fluctuations dans la position des formants d'une trame à l'autre, conduisant à l'obtention d'un signal filtré peu naturel. Pour réduire ces effets, [Hansen et Clements, 1987; Hansen et Clements, 1991] appliquent des contraintes spectrales inter et intra-trames, qui imposent la stabilité du modèle AR ainsi que la continuité des caractéristiques du spectre d'une trame à l'autre. Notons que cet algorithme se révèle très coûteux en calculs.

L'algorithme de Hansen et Clements applique le même traitement sur tout le signal. Or certaines sections du signal sont plus perturbées que d'autres par un même bruit. [Arslan et Hansen, 1994] utilisent des HMMs entraînés sur de la parole bruitée, associés à un processus de décision minimisant le coût d'une mauvaise classification, pour partitionner des segments du signal en classes phonétiques grossières. L'algorithme de Hansen et Clements est

ensuite appliqué en contrôlant la terminaison de l'algorithme itératif d'estimation, de façon spécifique à la classe phonétique reconnue. L'évaluation de la méthode montre une amélioration consistante de la distorsion d'Itakura-Saito pour différents SNRs et pour les différentes classes phonétiques.

[Ephraim *et al.*, 1987] estiment de façon itérative le niveau moyen du bruit ainsi que le spectre d'un modèle AR de parole propre. L'estimation minimise la distorsion d'Itakura-Saito entre le spectre de parole bruitée et son estimation, obtenue en combinant le spectre du modèle AR et le spectre estimé du bruit. Les paramètres des modèles AR ainsi déterminés sont ensuite utilisés en entrée d'un système de RAP. Cette méthode peut être appliquée sur les données de test ou d'apprentissage sans nécessiter une connaissance *a priori* du niveau de bruit. Sur une tâche de reconnaissance de mots isolés (alphabet + 10 chiffres), à un SNR de 10 dB et en mode monolocuteur, le filtrage des données de test permet d'augmenter le taux de reconnaissance de 40%.

Pour toutes ces méthodes basées sur l'estimation d'un modèle AR, il faut cependant remarquer que la qualité de l'estimation des pôles se dégrade rapidement lorsque le SNR diminue [O'Shaughnessy, 1988], et limite de ce fait la qualité des traitements en présence d'un niveau de bruit élevé.

Lorsque la *pdf* du signal de parole propre, la *pdf* du bruit, et une mesure de distorsion entre signaux sont connues, il est possible d'estimer le signal de parole propre selon un certain critère. Ce critère est généralement le minimum de l'erreur quadratique moyenne (MMSE [15]), ce qui correspond à utiliser l'erreur quadratique comme mesure de distorsion, ou bien le maximum *a posteriori* (MAP) [Ephraim, 1992b], qui correspond approximativement à l'utilisation d'une distorsion uniforme [Van Trees, 1968]. Ces deux approches, bien qu'impliquant des calculs très lourds, peuvent être interprétées de façon intuitive simple. Supposons que les *pdfs* de parole et de bruit soient modélisées par 2 HMMs de M et $\tilde{M}$ états, et que bruit et parole soient additifs. Une observation issue de n'importe quel état du HMM de bruit peut perturber une observation de parole issue de n'importe quel état du HMM de parole. Il est alors possible de définir $M \times \tilde{M}$ filtres et de calculer les probabilités *a posteriori* d'occurrence de chaque combinaison entre les états de parole et de bruit, étant donné le signal bruité. L'estimation MMSE consiste alors à appliquer tous les filtres sur chaque vecteur bruité, puis à effectuer une somme des sorties des filtres, pondérées par les probabilités *a posteriori* des combinaisons d'états [Ephraim, 1992a; Ephraim, 1992b]. Dans le cas d'une estimation MAP, seul le filtre le plus vraisemblable est appliqué à chaque instant [Ephraim *et al.*, 1989]. Les signaux sont modélisés par des processus AR et le filtrage consiste à décoder le signal bruité le long de la séquence d'états parole-bruit la plus probable, puis à appliquer chaque filtre aux vecteurs bruitées qui lui sont associés. Cette méthode est très proche de celle de [Lim et Oppenheim, 1978], mais ici le modèle AR de parole propre est estimé à partir d'un corpus d'apprentissage de parole propre, et non à partir de l'observation du signal de parole bruitée.

L'estimateur MAP ne peut être directement déterminé sous forme close, et la résolution s'effectue itérativement en utilisant l'algorithme EM [Dempster *et al.*, 1977]. L'estimation MAP est donc beaucoup plus coûteuse en calculs que MMSE. Les performances des estimateurs d'Ephraim sont comparées à celles d'un filtre de Wiener dans [Sheikhzadeh *et al.*,

[15] *Minimum Mean Square Error*

1994] sur une tâche de filtrage de bruit. Il apparaît que l'utilisation de MMSE et MAP donne les meilleurs résultats, autant quant à l'amélioration du SNR que de la qualité du signal ; leur supériorité est d'autant plus nette aux SNRs faibles et pour des bruits non stationnaires.

L'idée d'utiliser un estimateur spécifique à une classe de son a été introduite par [Drucker, 1968], qui classe le signal dans des catégories phonétiques grossières (voyelle, fricative, plosive, nasale, liquide), puis applique des filtres spécifiques à chaque classe. Dans [McAulay et Malpass, 1980], un estimateur MMSE différent est utilisé selon que le signal de parole est présent ou non dans le signal bruité observé, ce qui revient à considérer que le signal peut se trouver dans deux états différents (absence ou présence de parole). Le signal filtré correspond finalement à la somme pondérée des deux estimations, où les pondérations correspondent à la probabilité *a posteriori* des deux états, étant donné le signal bruité. Les estimateurs d'Ephraim sont donc une généralisation de ces approches.

Une lacune commune aux algorithmes d'estimation précédents est que les composantes spectrales sont estimées indépendamment les unes des autres. Or, les différentes fréquences du spectre de parole sont très corrélées entre-elles, et ignorer ces corrélations se traduit par une estimation sous optimale. Dans [Erell et Weintraub, 1993a], un estimateur MMSE des log-énergies en sortie d'un banc de filtres est déterminé. La corrélation entre les fréquences est partiellement utilisée en conditionnant l'estimation dans une bande de fréquence par l'énergie totale de la trame. Sur un test de reconnaissance de parole bruitée avec un SNR de 10 dB, l'utilisation de cette dépendance sur l'énergie totale permet de diviser par deux le taux d'erreur, et les performances sont proches de celles obtenues lorsque les conditions de test et d'apprentissage sont identiques. L'algorithme est étendu dans [Erell et Weintraub, 1993b] où le critère d'estimation utilisé est la minimisation de la distance Euclidienne entre les log-énergies en sortie d'un banc de filtres. De plus, l'estimation d'une trame est conditionnée par les trames adjacentes en utilisant un modèle de Markov, ce qui permet de conditionner l'estimation sur une séquence de trames afin de prendre en compte les corrélations temporelles du signal. Dans un environnement de bruit blanc stationnaire à 10 dB, les taux de reconnaissance sont équivalents à ceux obtenus lorsque test et apprentissage sont effectués dans les mêmes conditions.

[Acero et Stern, 1990] proposent une correction additive du cepstre, appelée CDCN[16], qui permet d'effectuer une normalisation de l'environnement. L'algorithme CDCN compense simultanément les effets d'un bruit additif et d'un filtrage linéaire inconnus. Le cepstre de parole propre peut s'exprimer comme la somme du cepstre de parole bruitée avec la fonction de transfert du filtre dans le domaine cepstral et une fonction non linéaire du cepstre de parole bruitée, du cepstre de la fonction de transfert du filtre et du cepstre du bruit. Le problème consiste à estimer les deux vecteurs de cepstre inconnus, que sont le cepstre du bruit et le cepstre de la fonction de transfert du filtre. L'estimation s'effectue selon un critère du maximum de la vraisemblance, qui minimise la distorsion entre les vecteurs de cepstre filtrés et un ensemble de prototypes de parole propre « normalisée », par un calcul itératif fondé sur EM. L'algorithme CDCN permet d'améliorer de façon significative les taux de reconnaissance lorsque apprentissage et test s'effectuent en utilisant des micros différents. Cependant, CDCN est très coûteux en calculs, et différentes versions légèrement moins efficaces

16. *Codeword-Dependent Cepstral Normalization*

mais nécessitant moins de calculs sont proposées dans [Acero et Stern, 1991; Acero, 1992; Acero et Stern, 1992].

7 Compensation de l'effet Lombard

[Takizawa et Hamada, 1990] proposent de compenser un vecteur de cepstre par un biais rendant compte du déplacement des formants de la parole Lombard, par rapport à la parole normale. Le biais est déterminé après analyse des différences entre parole normale et Lombard. Cette méthode s'avère plus efficace en reconnaissance que l'utilisation de la pondération RPS [Schroeder, 1981] du cepstre (cf. chap. 3, § 1).

[Hansen et Clements, 1989] développent des algorithmes pour compenser de la parole prononcée sous diverses conditions de stress (parole Lombard, parole criée, prononciation lente, rapide, etc.) en utilisant des tables de compensations spécifiques à chaque type de parole. Les tables sont déterminées après analyse des corpus de parole entre conditions neutres et de stress. Les paramètres compensés portent sur la position des formants et leurs largeurs de bande, ainsi que l'énergie du signal. Ces compensations sont combinées avec le filtrage du signal utilisant des contraintes spectrales intra et inter-trames [Hansen et Clements, 1987; Hansen et Clements, 1988] (cf. § 6) sur une tâche de reconnaissance d'un petit vocabulaire de mots isolés de parole Lombard bruitée. Une telle méthode s'avère très coûteuse en calcul et nécessite de connaître les positions des frontières entre phonèmes.

Dans [Chen, 1988], le cepstre du signal de test est compensé par une composante additive. Un mot à reconnaître est décodé par un système à base de HMMs, et le vecteur de stress est défini par la différence entre la moyenne des vecteurs de cepstre du mot et la moyenne des vecteurs moyennes des états du HMM alignés sur ce signal. Ayant constaté que l'amplitude des composantes de cette différence décroît de façon exponentielle en fonction de l'indice des composantes, Chen estime le vecteur de stress par une fonction exponentielle. Le cepstre du signal à reconnaître est alors corrigé par ce vecteur de stress lissé, et la reconnaissance s'effectue en utilisant les HMMs initiaux. Cette compensation ne corrige cependant que les variations de la pente spectrale entre parole propre et Lombard. Chen précise que les variations du rythme d'élocution et le déplacement des formants ne sont pas pris en compte, et sont responsables de la plupart des erreurs non éliminées.

Dans la compensation exponentielle de Chen, toutes les trames du mot à reconnaître sont compensées de la même façon. Cette méthode est étendue par [Hansen et Bria, 1990], qui dissocient la compensation des zones voisées de celle des zones non voisées. Dans [Hansen, 1993], cette méthode est associée à un filtrage de bruit utilisant des contraintes spectrales morphologiques [Hansen, 1991], et conduit à une amélioration de 38% des taux de reconnaissance, sur de la parole Lombard bruitée, avec un petit vocabulaire.

8 Conclusion

Les approches de prétraitement du signal, qui permettent de minimiser les variabilités provoquées par le bruit, peuvent être à l'heure actuelle considérées comme viables pour la

reconnaissance de la parole dans le bruit, ce qui n'était pas le cas il y a quelques années (p.ex. très mauvais comportement de la soustraction spectrale classique [Van Compernolle, 1989a] ou du filtrage de Kalman [Mokbel, 1992]).

Les progrès proviennent d'une part de la mise en œuvre de traitements spécifiques à des classes de sons, de la minimisation des distorsions provoquées par les techniques classiques de filtrage, et de la prise en compte du rapport signal-à-bruit instantané, éléments qui se retrouvent dans de nombreuses approches de filtrage à base de modèles, de transformation d'espace, ou encore de soustraction spectrale non linéaire. De tels traitements sont justifiés car les différentes zones d'un signal de parole ne sont pas modifiées de façon consistante par un bruit stationnaire, et peuvent alors être traitées spécifiquement en fonction de leurs caractéristiques.

La qualité des traitements est également améliorée par l'exploitation de modèles *a priori* du signal de parole et des corrélations spectrales du signal, qui permettent de prendre en considération les redondances et spécificités du signal de parole.

L'introduction de connaissances et contraintes psychoacoustiques dans les processus de filtrage contribue aussi à l'amélioration des performances, particulièrement en présence d'un niveau de bruit important.

Enfin, les approches de filtrage fondées sur l'utilisation de plusieurs microphones semblent très prometteuses, en particulier pour limiter les effets des bruits non stationnaires et traiter les problèmes d'échos, même si de tels traitements sont encore peu répandus en reconnaissance de parole.

Chapitre 2

Transformation des systèmes de reconnaissance

Dans ces familles de méthodes, le système de reconnaissance de parole est modifié afin de tenir compte de la présence d'un bruit lors de la reconnaissance. La modification peut s'effectuer d'une part au niveau du processus de décodage, pour autoriser la présence d'un signal concurrent perturbant la parole (cf. § 1); d'autre part, il est également possible d'introduire une étape de filtrage dans le processus de décodage, un filtre pouvant être associé à chaque modèle, ou à chaque état d'un modèle stochastique (cf. § 2). Des modèles spécifiques aux nouvelles conditions de test peuvent aussi être déterminés à partir des modèles initiaux de parole propre, en utilisant ou non des connaissances sur la nature de la perturbation (cf. § 3 et § 4). L'utilisation de critères discriminants d'apprentissage permet de lutter contre la source de variabilité introduite par le bruit (cf. § 5). Enfin, effectuer un apprentissage dans différentes conditions prédéfinies de bruit reste une solution efficace, bien que difficilement réalisable en pratique (cf. § 6).

1 Composition/décomposition de modèles

L'observation d'un signal de parole bruitée correspond à l'observation simultanée de deux signaux (parole et bruit) se combinant selon une certaine relation. Les signaux de parole et de bruit peuvent tous deux être représentés par des modèles de Markov, et le processus de décodage de la parole ou du bruit s'effectue en recherchant dans les modèles la séquence d'états la plus vraisemblable qui explique le signal observé (algorithme de Viterbi). À chaque instant, le signal de parole bruitée correspond donc à la combinaison d'une observation de parole associée à un état du HMM de parole, avec une observation de bruit associée à un état du HMM de bruit. Il est alors possible de déterminer la vraisemblance d'une observation de parole bruitée, et par conséquent de cheminer simultanément dans l'espace des états du HMM de bruit et du HMM de parole pour rechercher la séquence composite qui explique le signal bruité observé. Avec cette approche, proposée dans [Varga et Moore, 1990], des événements concurrents sont reconnus simultanément par un décodage de Viterbi dans l'espace combiné de parole et de bruit. Il est en général impossible de déterminer sous forme close l'expression de la fonction

de densité de probabilité d'une observation de parole bruitée, connaissant la *pdf* de la parole propre, la *pdf* du bruit et la relation de combinaison entre parole et bruit. Il est alors nécessaire de poser des hypothèses simplificatrices pour conduire les calculs. [Varga et Moore, 1990; Varga et Moore, 1991] représentent les signaux par les logarithmes des énergies en sortie d'un banc de filtres. En supposant que parole et bruit sont additifs, le logarithme de l'énergie du signal de parole bruitée est alors égal, en première approximation, au maximum du logarithme de l'énergie du bruit et de la parole. Une telle approximation a été initialement introduite par [Nadas *et al.*, 1989]. Sur une tâche de reconnaissance des 10 chiffres isolés dans différents bruits stationnaires ou non, la décomposition du signal bruité conduit à l'obtention de taux de reconnaissance supérieurs à ceux fournis par le masquage de Klatt [Varga et Moore, 1991]. [Kadirkamanathan, 1992] représente la combinaison entre parole et bruit par une fonction linéaire à 3 morceaux au lieu d'utiliser l'approximation du maximum, et améliore légèrement les performances de la méthode. Dans tous les cas, la complexité du décodage reste cependant très importante. Un parallèle peut être établi entre la décomposition de modèle et le filtrage de [Ephraim, 1992b], qui consiste à définir un filtre de Wiener pour toutes les combinaisons possibles entre un état d'un HMM de parole et un état d'un HMM de bruit. [Sorensen et Hartmann, 1994] associent la décomposition de modèles de Varga et Moore avec les HMM/RBF[1] de [Singer et Lippmann, 1992]. Les RBF sont utilisés pour modéliser les probabilités d'émission de symboles, et la décomposition hybride ainsi définie s'avère efficace et robuste pour la reconnaissance de chiffres dans un environnement de voiture. [Kobayashi *et al.*, 1994] développent une forme simplifiée de la décomposition de Varga et Moore, en introduisant en plus des paramètres dynamiques, associée avec une étape de soustraction spectrale [Boll, 1979].

Des idées proches de celles de Varga et Moore sont développées dans [Young, 1992a], avec la combinaison parallèle de modèles (PMC[2]), qui consiste à construire un modèle de parole bruitée à partir d'un modèle de parole propre et d'un modèle de bruit. Les modèles combinés sont des HMMs ; si le HMM de parole propre comporte M états et le HMM de bruit N états, on construit alors un HMM de parole bruitée de $M \times N$ états, qui s'exprime entièrement à l'aide des paramètres des HMMs de bruit et de parole propre. Si les signaux de parole et de bruit sont modélisés dans l'espace MFCC par des loi normales, il est impossible de déterminer sous forme close l'expression de la *pdf* de la parole bruitée. Young utilise alors l'hypothèse de log-normalité, qui consiste à supposer que la somme de deux variables aléatoires distribuées de façon log-normale est également une variable aléatoire log-normale. Sous cette hypothèse, il est possible de déterminer l'expression des HMMs de parole bruitée. Ce travail a été largement développé dans [Gales et Young, 1992; Gales et Young, 1993a; Gales et Young, 1993c] pour des tâches de reconnaissance de mots isolés dans le bruit, puis étendu pour la reconnaissance de parole continue dans le bruit [Gales et Young, 1994a; Gales et Young, 1994b] mais pour des rapports signal-à-bruit modérés (18-20 dB). Gales et Young ont enrichi la méthode pour prendre en compte la différence de canal d'enregistrement entre les conditions de test et d'apprentissage [Gales et Young, 1993d; Gales et Young, 1993b]. La PMC est également appliquée par [Martin *et al.*, 1993], sur un espace de paramètres LPC-cepstre, pour la reconnaissance de parole continue dans différents bruits, station-

1. *Radial Basis Function*
2. *Parallel Model Combination*

naires ou non. La possibilité d'appliquer cette méthode pour prendre en compte des bruits non stationnaires est un atout important ; lorsque les statistiques du bruit changent, il est possible de réestimer les modèles de bruit pendant les pauses de parole, et la combinaison entres modèles de parole propre et bruit est peu coûteuse. Dans [Siohan *et al.*, 1994], la PMC est utilisée pour combiner un HMM de bruit à un STM de parole propre, et conduit à l'obtention d'un STM composite de parole bruitée. Dans toutes ces applications, la combinaison de modèles se révèle très efficace en présence d'un bruit modéré, et ne dégrade pas les performances pour la reconnaissance de parole propre.

Lorsque le bruit devient important, la grande variance des signaux réduit la discrimination, et il est intéressant d'appliquer un filtrage de bruit [Nolazco Flores et Young, 1993; Nolazco Flores et Young, 1994]. Or, un filtrage de type soustraction spectrale introduit une distorsion sur le signal. Cette distorsion peut être modélisée dans un cadre stochastique, et il est possible d'exprimer la moyenne de la parole filtrée en fonction de la moyenne de la parole propre et d'un coefficient de compensation. [Nolazco Flores et Young, 1993; Nolazco Flores et Young, 1994] utilisent le cadre de la PMC combiné avec une soustraction spectrale de bruit, où la soustraction spectrale est utilisée pour atténuer le bruit de fond, tandis que la PMC est utilisée pour générer des modèles de parole filtrée à partir de modèles de parole propre et de bruit, et des paramètres de la soustraction spectrale. Sur une tâche de reconnaissance de mots isolés ou connectés dans le bruit, cet algorithme s'avère plus efficace que la PMC initiale, en particulier en présence d'un niveau de bruit important.

Ces approches ont toutes en commun le problème de la détermination de la *pdf* de la parole bruitée, connaissant la *pdf* du signal propre et la *pdf* du bruit. [Nadas *et al.*, 1989; Varga et Moore, 1990; Gales et Young, 1992; Kadirkamanathan, 1992; Rose *et al.*, 1994] considèrent que le signal de parole bruitée correspond au maximum du signal de parole propre et du bruit, dans le domaine log-spectral. Selon [Rose *et al.*, 1994], l'approximation par le maximum dans le domaine logarithmique est une hypothèse raisonnable lorsque les signaux sont additifs. L'approximation linéaire à 3 segments de Kadirkamanathan est cependant plus précise. Dans [Young, 1992a; Gales et Young, 1993a; Martin *et al.*, 1993; Siohan *et al.*, 1994], l'hypothèse de log-normalité des observations bruitées dans le domaine spectral est utilisée, ce qui revient à utiliser une distribution normale pour la parole bruitée dans le domaine cepstral. D'après [Openshaw et Mason, 1994], la *pdf* de la parole bruitée dans le domaine cepstral est bimodale, et ne peut donc pas être modélisée par une loi normale ; les performances de la PMC se dégradent donc rapidement lorsque l'énergie du bruit devient importante [Siohan *et al.*, 1994].

2 Filtrage par état

Le filtrage de Wiener s'avère efficace lorsqu'on l'applique sur un signal stationnaire corrompu par un bruit additif, mais reste limité par la non-stationnarité sous-jacente du signal de parole. Or, les HMMs découpent automatiquement la parole en segments quasi-stationnaires, correspondant aux états des modèles [Beattie et Young, 1991]. Il est alors possible d'associer à chaque état d'un HMM un filtre de Wiener, et d'appliquer un filtrage lors de la reconnaissance de la parole bruitée. [Beattie et Young, 1991] utilisent cette méthode pour effectuer

une reconnaissance de mots isolés dans le bruit. Les HMMs sont entraînés avec de la parole propre, et l'algorithme de décodage (Viterbi) est modifié pour effectuer un filtrage de Wiener par état. Les coefficients du filtre sont associés à chaque état des HMMs et le filtrage s'effectue sur les vecteurs de paramètres, un filtre étant spécifique à chaque composante du vecteur (les vecteurs de paramètres sont constitués par les sorties d'un banc de filtres). Les coefficients des filtres sont calculés une fois pour toute avant la phase de reconnaissance, mais il est possible de les remettre à jour périodiquement, pour prendre en compte les variations des caractéristiques du bruit. Avec une telle méthode, le filtrage est effectué lors de la reconnaissance, et de façon spécifique à chaque état. Dans [Beattie et Young, 1992b; Beattie et Young, 1992a], ce travail est étendu pour effectuer un filtrage dans le domaine cepstral, toujours dans le cadre d'un filtrage de Wiener, et se traduit alors par une correction additive du cepstre bruité. Un vecteur de compensation cepstrale est associé à chaque état d'un HMM, et le vecteur bruité est translaté par le vecteur de compensation lors de la reconnaissance. L'avantage de cette méthode est son faible coût en calculs lors de la reconnaissance. Évaluée sur une application de reconnaissance de mots isolés dans un environnement de voiture, cette méthode s'avère plus efficace que la soustraction spectrale non linéaire de [Lockwood et Boudy, 1992], la décomposition de [Gales et Young, 1992] et l'utilisation de la distance de projection cepstrale.

[Vaseghi et Milner, 1992; Vaseghi et Milner, 1993b] mettent en œuvre un filtrage par état, mais d'une façon différente de celle de Beattie et Young, pour une reconnaissance de l'alphabet dans le bruit. Le signal bruité est décodé avec chacun des HMMs de mots, entraînés sur de la parole propre. Les N modèles candidats ayant les scores les plus élevés sont sélectionnés, et pour chacun de ces modèles, la séquence d'états la plus vraisemblable est utilisée pour calculer un filtre de Wiener. Le signal bruité est alors filtré en utilisant le filtre associé à chacun des N modèles candidats, et un nouveau score est recalculé ; le modèle conduisant au score le plus élevé est finalement sélectionné. Les performances sont supérieures à celles de la soustraction spectrale de [Porter et Boll, 1984], mais il faut noter que la séquence d'états obtenue en alignant la parole bruitée sur les modèles propres devient incorrecte lorsque le SNR est faible. L'approche de [Vaseghi et Milner, 1992] peut être considérée comme une extension des travaux de [Berstein et Shallom, 1991], qui effectuent un filtrage de Wiener dans le domaine cepstral, lors du décodage avec un système de reconnaissance basé sur la DTW. Dans [Vaseghi et Milner, 1993a], il apparaît qu'il est préférable d'utiliser un filtrage par états dans le domaine cepstral similaire à celui de [Beattie et Young, 1992b], plutôt qu'appliquer le filtrage calculé à partir de la séquence d'états la plus probable. Dans [Vaseghi *et al.*, 1994], le filtrage par états est étendu à un paramétrage temps-cepstre [Pai et Wang, 1992], et s'avère plus efficace que la soustraction spectrale de [Boll, 1979] sur une application de reconnaissance de chiffres isolés dans le bruit. Le paramétrage temps-cepstre, qui modélise l'enveloppe temps-fréquence de la parole semble très robuste pour la reconnaissance de la parole dans le bruit, comme le confirment également les travaux de [Kitamura *et al.*, 1992].

3 Adaptation des modèles acoustiques

Un grand nombre de méthodes ont été développées pour permettre de reconnaître de la parole bruitée à partir d'un HMM initialement entraîné sur de la parole propre. Bien souvent, ces approches sont dérivées des travaux sur l'adaptation au locuteur des systèmes de reconnaissance, et consistent à transformer une fois pour toute les modèles de parole propre, la ou les transformations étant déterminées à partir d'un corpus d'adaptation. Une fois les modèles adaptés, ces méthodes ont l'avantage de ne nécessiter aucun calcul supplémentaire.

Nous ne nous attarderons pas sur l'adaptation des prototypes des systèmes à base de HMMs discrets, utilisant des procédés du type adaptation au locuteur [Nakamura et Shikano, 1989; Ohkura et Sugiyama, 1991; Shikano *et al.*, 1986] qui s'apparentent plus à une transformation de la parole, et qui ont déjà été présentés dans la partie sur la transformation de *codebook* (cf. chap. 1, § 3).

Les méthodes de transformation d'espace peuvent également être utilisées pour transformer les paramètres des modèles. Ainsi, [Mokbel *et al.*, 1992a; Mokbel, 1992; Mokbel *et al.*, 1992b] appliquent une transformation linéaire sur les paramètres des *pdf* d'un HMM (moyennes et variances), déterminée à partir d'un ensemble de couples de vecteurs de parole propre et de parole bruitée. Cette transformation s'avère plus efficace que la soustraction spectrale non linéaire pour une application de reconnaissance de mots isolés dans une voiture en déplacement.

Les méthodes d'adaptation de modèles peuvent être considérées comme des problèmes de réentraînement de modèles à partir d'un faible volume de données. [Stern et Lasry, 1987] effectuent l'adaptation au locuteur d'un système de reconnaissance de mots isolés, en utilisant le cadre de l'apprentissage Bayésien. Les vecteurs moyennes des lois normales sont réestimés en utilisant les vecteurs moyennes initiaux multilocuteurs, ainsi qu'un corpus d'adaptation spécifique au nouveau locuteur. [Lee *et al.*, 1990; Lee *et al.*, 1991] utilisent également l'apprentissage du maximum *a posteriori* pour adapter les paramètres d'un HMM continu à un nouveau locuteur. L'intérêt principal de cette approche est que les estimateurs convergent vers les estimateurs du maximum de la vraisemblance lorsque le volume de données d'adaptation est important. Cette approche est également publiée dans [Gauvain et Lee, 1992; Gauvain et Lee, 1994; Lee et Gauvain, 1993] sous un cadre unifié pour l'adaptation au locuteur et à l'environnement, le lissage des paramètres, et l'apprentissage correctif. Sur une tâche d'adaptation au locuteur d'un système multilocuteur, l'exploitation d'un corpus d'adaptation constitué de 5 minutes de parole permet d'obtenir des performances équivalentes à celles d'un système dépendant du locuteur entraîné à partir de 30 minutes de parole [Gauvain et Lee, 1992].

[Ohkura *et al.*, 1992] développent l'algorithme de VFS[3] pour adapter les vecteurs moyennes d'un HMM continu. L'adaptation s'effectue en trois étapes. Tout d'abord, les modèles spécifiques aux nouvelles conditions sont initialisés avec les modèles de référence, puis réentraînés avec les données d'adaptation, dans le but de déterminer un vecteur de translation de chaque moyenne des modèles pour lesquels il existe des données d'adaptation. Ensuite, un vecteur de translation est déterminé par interpolation pour chaque état des modèles auxquels

[3] *Vector Field Smoothing*

n'était associée aucune donnée d'adaptation. Enfin, chaque vecteur de translation est lissé en fonction de ses voisins les plus proches, et ce vecteur lissé est utilisé pour translater la moyenne de l'état qui lui est associé. En utilisant une minute de parole d'adaptation, sur une tâche d'adaptation au locuteur, le taux de reconnaissance obtenu correspond à 96% du taux de reconnaissance en mode dépendant du locuteur. [Takahashi et Sagayama, 1994] utilisent l'algorithme de VFS pour adapter des modèles multilocuteurs entraînés sur de la parole de bonne qualité, afin de reconnaître des mots isolés au téléphone. Avec seulement 10 mots d'adaptation, l'algorithme de VFS permet de réduire le taux d'erreur de 45% par rapport au système non adapté initial.

Les modèles peuvent également être adaptés pour prendre en compte les variations provoquées par l'effet Lombard. Dans [Suzuki *et al.*, 1994], les vecteurs moyennes de HMMs sont adaptés, et trois facteurs spécifiques à l'effet Lombard sont compensés : le déplacement des formants, la réduction de leur largeur de bande et la modification de la pente spectrale. Suzuki *et al.* alignent de la parole Lombard sur un HMM entraîné à partir de parole propre. L'enveloppe spectrale du vecteur moyenne de chaque état du HMM est alignée sur les enveloppes spectrales des trames de parole Lombard associées. Une différence moyenne entre les spectres alignés de parole Lombard et parole propre est calculée, et cette différence est utilisée pour corriger le vecteur moyenne du HMM. L'opération d'alignement de la parole Lombard sur les HMMs, puis de correction des HMMs est répétée plusieurs fois jusqu'à atteindre la stabilité de l'alignement. Sur une tâche de reconnaissance de 100 mots isolés en parole Lombard (sans le bruit), la correction des modèles améliore le taux de reconnaissance d'environ 11%.

4 Adaptation des modèles de durée

En présence d'un bruit d'environnement important, le locuteur modifie son effort vocal pour que ses propos restent intelligibles. Cela se traduit par une modification des caractéristiques spectrales du signal, mais également par une grande variation du rythme d'élocution [Junqua, 1993]. Si un système de RAP modélise de façon explicite les durées des phonèmes, il est alors possible d'adapter les modèles de durée pour prendre en compte ce phénomène. Dans [Siohan *et al.*, 1993], l'adaptation Bayésienne est utilisée pour transformer les modèles de durée des phonèmes. Sans corriger les modèles acoustiques, cette compensation des modèles de durée permet d'améliorer les taux de reconnaissance de mots isolés dans le bruit, et confirme ainsi que l'utilisation d'un modèle de durée permet d'améliorer les performances d'un système de RAP.

En présence d'un bruit modéré, l'effet Lombard ne se manifeste pas [Das *et al.*, 1993] ; le bruit provoque une distorsion spectrale mais le rythme d'élocution varie peu. Les statistiques concernant la modélisation de la durée de séjour dans un état d'un HMM sont donc plus robustes que les statistiques d'émission de vecteur acoustique. [Nicol *et al.*, 1992] exploitent cette robustesse et introduisent une modélisation de la durée en décomposant chaque état d'un HMM en une séquence de sous-états ayant chacun la même probabilité d'émission de vecteur. Sur une tâche de reconnaissance de 23 mots isolés dans le bruit, cette modélisation de la durée améliore le taux de reconnaissance de 13 à 20% pour des SNRs de 20 dB et

10 dB.

5 Apprentissage discriminant

Bien souvent, l'apprentissage de modèles stochastiques est basé sur le critère du maximum de la vraisemblance (MLE[4]) [Bahl *et al.*, 1983], et n'optimise donc pas la discrimination entre les classes à identifier. Ce critère d'estimation garantit cependant l'optimalité de l'apprentissage si les modèles correspondent aux données [Bahl *et al.*, 1987]. En présence de bruit, la discrimination entre les différentes unités de parole devient délicate et l'utilisation d'un critère d'apprentissage discriminant semble donc judicieuse.

[Mizuta et Nakajima, 1992] effectuent un apprentissage discriminant en utilisant de la parole bruitée, pour corriger des HMMs initialement entraînés par MLE sur de la parole propre. Lorsque le corpus bruité utilisé pour l'apprentissage discriminant comporte plusieurs types et niveaux de bruits, les modèles obtenus s'avèrent robustes pour différents environnements bruités de test.

[Frangoulis et Sgardoni, 1991; Frangoulis et Gaganelis, 1992] adaptent les vecteurs moyennes d'un HMM continu entraîné à partir de parole propre, en utilisant un corpus d'adaptation de parole bruitée. L'adaptation consiste à déplacer par essais successifs les vecteurs moyennes afin d'augmenter la discrimination sur le corpus d'adaptation. La méthode s'apparente à un apprentissage correctif, mais reste très grossière, utilise de nombreux seuils et ne semble fonctionner que parce que la quantité de données d'adaptation est grande et la tâche très simple.

Le critère d'apprentissage du minimum d'erreur de classification (ME[5]) [Chou *et al.*, 1992; Euler et Zinke, 1992] est comparé avec l'apprentissage MLE dans [Ohkura *et al.*, 1993] sur une tâche de reconnaissance de mots isolés. Il apparaît que ME est plus robuste aux variations d'environnement que MLE, quelques soient les conditions d'apprentissage (propre ou bruitée). Ainsi, des modèles entraînés à un SNR donné conduisent aux meilleurs taux de reconnaissance lorsque l'apprentissage est effectué par ME, quelque soit le SNR de test. La même conclusion persiste lorsque les modèles sont entraînés à partir d'un corpus d'apprentissage comportant différents niveaux de bruit.

6 Apprentissage multiréférences

Une stratégie possible pour la reconnaissance de la parole dans le bruit consiste à entraîner les systèmes dans le bruit [Dautrich *et al.*, 1983a; Dautrich *et al.*, 1983b]. Cette approche peut être considérée comme une forme de compensation de modèle, qui supprime totalement les différences entre conditions de test et d'apprentissage. [Morii *et al.*, 1990] ajoutent une estimation du bruit aux vecteurs de références d'un système de reconnaissance à base de programmation dynamique. [Das *et al.*, 1993; Das *et al.*, 1994] superposent également un bruit de fond aux références, avant d'appliquer la transformation de *codebook* de [Nadas *et al.*,

4. *Maximum Likelihood Estimation*
5. *Minimum Error*

1989]. [Mokbel et Chollet, 1991b; Mokbel et Chollet, 1991a] suggèrent d'ajouter le bruit au signal temporel de référence, plutôt que de chercher à supprimer le bruit du signal de test. L'ajout du bruit dans les références est simple à mettre en œuvre et permet d'éviter les problèmes liés à l'obtention d'un spectre de puissance négatif lors d'une soustraction spectrale. Il est cependant difficile de prévoir quelles seront les conditions de bruit lors du test, et le système entraîné dans le bruit devient inefficace pour reconnaître la parole propre. De plus, un système entraîné dans le bruit est très sensible aux variations du niveau et du type de bruit [Kitamura *et al.*, 1992]. L'apprentissage multi-style utilisant différents types et niveaux de bruits est possible, mais provoque une diminution sensible des taux de reconnaissance en parole propre [Kitamura *et al.*, 1992].

Afin de prendre en compte les variations provenant du mode d'expression du locuteur, il est possible d'entraîner un système dans différentes conditions. [Lippmann *et al.*, 1987] effectuent un apprentissage multi-style qui consiste à entraîner un système avec de la parole propre, Lombard, criée, etc. Bien qu'efficace, cette méthode nécessite de disposer d'un corpus d'apprentissage enregistré sous différentes conditions de stress, dont la collecte est délicate et coûteuse. [Bou-Ghazale et Hansen, 1994] proposent, après une analyse statistique des différents style de parole (Lombard, lente, voix forte), une méthode permettant de générer artificiellement ces différents styles de parole à partir d'un corpus de parole normale. Le système de RAP est entraîné avec ces données générées artificiellement et il devient inutile de collecter des corpus de parole pour ces différents styles d'élocution.

[Matsuoka et Shikano, 1991] proposent de combiner des modèles entraînés sous différents styles d'élocution, avec la méthode d'interpolation par suppression (*deleted interpolation*) de [Jelinek et Mercer, 1980].

L'apprentissage multi-styles diminue le pouvoir discriminant des modèles. Plutôt qu'effectuer l'apprentissage d'un seul système à partir de plusieurs locuteurs, [Witbrock et Haffner, 1992] entraînent plusieurs systèmes, chacun d'entre-eux étant spécifique à une classe de locuteur. Lors de la reconnaissance, la classe du locuteur courant est identifiée rapidement et le système spécifique à cette classe de locuteurs est utilisé. Une approche analogue est exploitée par [Imamura, 1991] et permet d'obtenir des résultats supérieurs à ceux obtenus avec un système entraîné en mode multilocuteur. [Treurniet et Gong, 1994] développent des idées semblables pour débruiter un espace de paramètres. Plusieurs transformations d'espace sont apprises spécifiquement à une condition de bruit donnée (SNR et type de bruit); lors de la reconnaissance, la condition de bruit est d'abord identifiée, puis la transformation spécifique à ce bruit est employée. Dans le même ordre d'idées, [Hansen *et al.*, 1994] proposent une méthode pour identifier l'état de stress d'un locuteur, ce qui devrait permettre, à terme, d'appliquer des compensations spécifiques au stress.

Lorsque le test et l'apprentissage des systèmes sont effectués dans les mêmes conditions, on considère généralement que cela correspond à la configuration optimale de test. Pourtant, certaines méthodes de compensation de modèles permettent d'obtenir des taux de reconnaissance supérieurs à ceux obtenus lorsque test et apprentissage s'effectuent dans des conditions similaires. Sur une tâche de détection de mots dans un flot de parole bruitée, la transformation d'espace probabiliste de [Gish *et al.*, 1990] permet d'obtenir un gain de performance d'environ 20% par rapport à un apprentissage et un test dans les mêmes conditions. Sur une application similaire, en présence d'un bruit additif de 10 dB, la transformation d'espace de [Ng

et al., 1992] se traduit par des performances supérieures de 15% à celles obtenues lorsque les conditions de test et d'apprentissage sont identiques. [Siohan *et al.*, 1994] rapportent des conclusions semblables sur une tâche de reconnaissance de mots isolés dans le bruit, également confirmées sur une application de reconnaissance de parole continue [Siohan *et al.*, 1995].

7 Conclusion

Comme dans les approches de filtrage de bruit, les méthodes de transformation des systèmes de reconnaissance tirent avantage de la mise en œuvre de compensations associées aux différents sons, c.-à-d. spécifiques aux modèles (combinaisons de modèles, filtrages par états), qui utilisent une connaissance *a priori* du signal de parole propre fournie par les modèles utilisés pour la reconnaissance. La détermination précise des statistiques du bruit est généralement nécessaire, comme dans certaines approches de filtrage du signal, et conditionne fortement la qualité des résultats obtenus. Les méthodes de combinaisons de modèles permettent de prendre en compte des événements concurrents (parole et bruit), en particulier des bruits non stationnaires, tout en ne nécessitant que l'utilisation d'un seul microphone.

Un avantage de certaines approches de transformations de modèles (p. ex. adaptation des paramètres des modèles) réside dans leur caractère généraliste, qui permet d'éviter de se focaliser sur une perturbation particulière comme le bruit additif, pour être potentiellement exploitable pour des applications d'adaptation au locuteur ou à la ligne de transmission. Malheureusement, dans de nombreuses approches d'adaptation de modèles, un corpus d'adaptation de plusieurs minutes de parole est souvent nécessaire afin de déterminer les compensations à appliquer (p.ex. réestimation Bayésienne).

Enfin, il faut noter que certaines méthodes d'adaptation garantissent que leur mise en œuvre ne va pas perturber les performances du système en présence d'un bruit d'environnement faible, ce qui constitue une propriété indispensable pour un système de reconnaissance destiné à fonctionner dans des conditions d'environnement variables.

Chapitre 3

Paramétrages et mesures de similarité robustes

La minimisation des différences entre l'environnement de test et de référence d'un système de RAP peut s'effectuer en recherchant un paramétrage du signal de parole et une mesure de similarité associée robustes aux variations des conditions d'environnement. On s'intéresse donc ici plus aux effets du bruit et à la façon de définir un paramétrage insensible à ces effets, qu'à la façon de supprimer ou d'atténuer ce bruit. La représentation du signal de parole étant supposée indépendante du bruit, un système entraîné sur de la parole propre peut alors être utilisé dans un environnement calme ou bruyant, sans modification de sa configuration.

L'avantage des méthodes mises en œuvre est qu'elles ne nécessitent en général que peu de connaissances sur le bruit perturbateur. En particulier, il est inutile de disposer des statistiques du bruit. Cet avantage peut s'avérer être un inconvénient, dans la mesure où on ne tire aucun parti des caractéristiques spécifiques à un bruit.

1 Représentations acoustiques et distances robustes

L'idée principale pour définir un paramétrage du signal et une mesure associée robustes au bruit consiste à privilégier de façon automatique les zones du spectre les moins perturbées par le bruit, au détriment des zones fortement affectées. La compensation des effets du bruit s'effectue ainsi de façon implicite, par la définition d'une mesure de distance robuste. Un parallèle peut être établi entre ces méthodes et les techniques de masquage de bruit (cf. chap. 1, § 5), qui visent à diminuer l'influence des régions du spectre les plus perturbées par la présence du bruit.

La distorsion d'Itakura-Saito (IS) [Itakura et Saito, 1968; Itakura, 1975] et une de ses variantes, le rapport de vraisemblance (LR [1]), sont les éléments centraux des mesures de similarité entre signaux de parole, et se basent sur la différence entre spectres dans le domaine logarithmique. Des pondérations spectrales (WLR [2]) ont ensuite été introduites pour amélio-

[1] *Likelihood Ratio*

[2] *Weighted Likelihood Ratio*

rer les performances des systèmes de RAP [Shikano et Sugiyama, 1982]. Ces pondérations ont enfin été étendues pour privilégier l'influence des pointes spectrales, plus robustes au bruit que les vallées spectrales.

[Soong et Sondhi, 1987] utilisent des pondérations non uniformes du spectre associées à la distorsion d'IS, et observent que, sur une tâche de reconnaissance de mots isolés dans le bruit, cette mesure donne des résultats équivalents à la distorsion d'IS lorsque le SNR est élevé, mais permet d'obtenir de meilleurs résultats aux SNRs faibles. [Matsumoto et Imai, 1986] rapportent que la WLR améliore les taux de reconnaissance d'une trentaine de mots isolés dans le bruit. Par contre, après une comparaison de différentes mesures de distorsions spectrales, [Nocerino *et al.*, 1985] constatent que la WLR ne permet pas d'obtenir des taux de reconnaissance supérieurs à ceux de la distorsion LR, sur une tâche difficile comme la reconnaissance du "E" *set.*[3]

Il est également possible de privilégier les régions du spectre de forte énergie par des pondérations appliquées sur les coefficients cepstraux. Des pondérations appropriées permettent ainsi d'obtenir de bons résultats de reconnaissance, à la fois en parole propre et bruitée [Matsumoto et Imai, 1986].

Une fonction de pondération généralisée des coefficients cepstraux est introduite par [Itakura et Umezaki, 1987]. Cette pondération conduit à la mesure de distance spectrale basée sur le spectre lissé à retard de groupe (SGD[4]), qui permet de privilégier les pointes spectrales. Sur une tâche de reconnaissance de mots isolés dans le bruit, cette distance conduit à l'obtention de résultats supérieurs à ceux provenant de l'utilisation d'une distance cepstrale [Itakura et Umezaki, 1987].

[Tohkura, 1987] introduit une distance cepstrale où les coefficients du cepstre sont pondérés par l'inverse de leur variance, ce qui permet de diminuer l'influence des coefficients d'ordre faible. Cette distance s'apparente à une distance de Mahalanobis où les termes non diagonaux de la matrice de covariance seraient forcés à zéro. Sur une application de reconnaissance de mots isolés, cette distance cepstrale pondérée s'avère plus performante que la distance Euclidienne.

[Juang *et al.*, 1986; Juang *et al.*, 1987] étudient la variabilité des coefficients cepstraux. Il apparaît que la variabilité des coefficients d'ordre élevé provient d'artefacts de la procédure d'analyse du signal par LPC-cepstre ; leur influence doit donc être pénalisée par rapport à celle des coefficients d'ordre inférieur. La variabilité des coefficients d'ordre faible provient des variations de la ligne de transmission, de l'effort vocal, et de la pente spectrale, éléments préjudiciables pour la reconnaissance multilocuteurs. En appliquant une fenêtre de liftrage sinusoïdale (SWL[5]), Juang *et al.* réduisent l'influence des coefficients d'ordre faible et élevé, par rapport aux coefficients d'ordre moyen. En reconnaissance de parole propre, ce liftrage s'avère plus efficace que la WLR et la pondération cepstrale de Tohkura.

La distance RPS[6] [Schroeder, 1981] consiste à pondérer les coefficients cepstraux par leurs indices. Sur une expérience de reconnaissance de voyelles, [Paliwal, 1982] constate que

---

3. Le "E" *set* est un ensemble de mots difficiles à reconnaître, constitué de chiffre et lettres de l'alphabet anglais : 3, b, c, d, e, g, p, v, t, z.

4. *Smooth Group Delay*

5. *Sine Wave Lifter*

6. *Root Power Sums*

les performances des coefficients cepstraux ainsi pondérés sont meilleures que celles obtenues par distance Euclidienne. [Hanson et Wakita, 1986a] montrent que la distance RPS est une approximation de la distance entre pentes spectrales. Sur une tâche de reconnaissance de mots isolés en parole bruitée, la distance RPS s'avère plus efficace que la distance cepstrale pondérée de Tohkura et que le liftrage cepstral de Juang *et al.* [Hanson et Wakita, 1986b; Hanson et Wakita, 1987]. Si on sature l'indice de pondération au dessus d'un ordre fixé, la distance RPS est très proche de la distance de Tohkura.

La présence d'un bruit dégrade la qualité de l'estimation LPC [Lim et Oppenheim, 1978], et les performances des systèmes de RAP utilisant ce paramétrage chutent rapidement lorsque le SNR diminue [O'Shaughnessy, 1988]. [Mansour et Juang, 1988] utilisent la cohérence entre les segments adjacents du signal pour améliorer la robustesse au bruit. Plutôt que d'effectuer une modélisation tout-pôle du signal temporel, l'autocorrélation du signal est d'abord calculée, suivie d'une compression en racine carrée dans le domaine spectral, avant d'effectuer une analyse LPC. Ces opérations définissent la représentation SMC[7]. Sur une tâche de reconnaissance d'une quarantaine de mots isolés en parole propre, la SMC-cepstre conduit à l'obtention de résultats équivalent à ceux de la LPC-cepstre. Par contre, en reconnaissance de parole bruitée, la SMC-cepstre s'avère plus efficace que la LPC-cepstre [Mansour et Juang, 1988]. [Mena *et al.*, 1990] rapportent la même conclusion. Dans [Gomez-Mena *et al.*, 1991], l'analyse SMC-cepstre est combinée avec une soustraction spectrale, et se révèle plus performante que la LPC-cepstre et que la SMC-cepstre sans filtrage de bruit.

Par contre, [Nakamura *et al.*, 1993] comparent le paramétrage par SMC-cepstre avec la LPC-cepstre, en utilisant différentes mesures de distances (RPS, SWL) sur une tâche de détection de mots dans un environnement de voiture, avec un système de reconnaissance à base de DTW multiréférences. Les résultats des expériences indiquent la supériorité de l'analyse LPC-cepstre associée à une mesure SWL. Les auteurs attribuent les mauvaises performances de la SMC-cepstre à l'utilisation d'un coefficient inapproprié de pré-accentuation du signal. [Erell et Weintraub, 1993b] rapportent également que la pré-accentuation, traitement habituel en reconnaissance de parole propre, dégrade les performances en présence de bruit.

[Hernando et Nadeu, 1991] proposent l'analyse OSALPC[8], qui consiste à effectuer une analyse LPC sur la partie causale de l'autocorrélation du signal. Ce traitement accentue l'influence des zones de forte énergie du spectre. Sur une tâche de reconnaissance de 10 mots isolés perturbés par un bruit blanc, en utilisant un système à base de HMMs, l'analyse OSALPC-cepstre fournit des résultats équivalents à ceux d'une SMC-cepstre de même ordre. Par contre, lorsque l'ordre de la OSALPC augmente, les performances dépassent celles obtenues par la SMC, en particulier pour les SNRs faibles. Dans [Hernando et Nadeu, 1994], l'analyse OSALPC est évaluée et comparée avec d'autres méthodes d'analyse pour la reconnaissance de mots isolés en environnement de voiture. Les auteurs rapportent que l'analyse OSALPC-cepstre combinée à un liftrage cepstral et à l'utilisation de coefficients dynamiques conduit à l'obtention des meilleurs résultats. Sur une tâche d'identification du locuteur dans le bruit, OSALPC-cepstre s'avère plus performant que le paramétrage MFCC et LPC-cepstre [Hernando *et al.*, 1994].

---

7. *Short-time Modified Coherence*
8. *One-Sided Autocorrelation Linear Predictive Coding*

Une unification de l'analyse cepstrale et de l'analyse AR est proposée dans une série de travaux sous le cadre de l'analyse homomorphique en racine [Alexandre *et al.*, 1993; Alexandre, 1993; Alexandre et Lockwood, 1993; Lockwood et Alexandre, 1994]. Ces travaux sont une extension de la déconvolution homomorphique de [Lim, 1979], qui utilise des fonctions puissance $(\cdot)^\gamma$ et racine $(\cdot)^{1/\gamma}$ à la place de l'exponentielle et du logarithme. En faisant varier γ, l'analyse en racine permet de passer d'un modèle tout-pôle ($\gamma = -1$) à un modèle tout-zéro ($\gamma = 1$), en passant par un modèle équiréparti en pôles et zéros ($\gamma = 0$). Cette méthode d'analyse permet d'obtenir un compromis entre assurer une déconvolution correcte modèle/source et approcher au mieux la structure résonnante du signal de parole. Le choix de γ permet donc de concentrer la mesure de distorsion sur les pointes du spectre, les vallées, ou encore de donner autant d'importance aux pointes qu'aux vallées. Une variante est proposée dans [Kobayashi et Imai, 1984], qui garantit une meilleure convergence vers la déconvolution logarithmique quand γ tend vers 0. L'analyse en racine permet de réduire la sensibilité aux zones du spectre de faible énergie, particulièrement affectées par la présence d'un bruit, et se révèle efficace pour la reconnaissance de mots isolés dans une voiture [Alexandre et Lockwood, 1993].

Afin de définir une mesure de distorsion robuste au bruit, il est important de comprendre les effets du bruit sur les paramètres de parole. [Mansour et Juang, 1989] démontrent qu'un bruit blanc additif provoque une diminution de la norme des vecteurs de LPC-cepstre, ainsi qu'une rotation des vecteurs inférieure à $\pi/2$. Pour exploiter le fait que le bruit affecte moins l'angle entre les vecteurs que leurs normes, Mansour et Juang proposent différentes mesures de distorsions entre vecteurs, basées sur la projection de l'un sur l'autre. Des tests de reconnaissance de mots isolés montrent que la mesure de projection cepstrale permet d'obtenir des meilleurs résultats que le liftre sinusoïdal (SWL) cepstral. Des analyses acoustiques montrent que cette distance privilégie les pointes spectrales du spectre [Carlson et Clements, 1991].

La mesure de projection cepstrale est incorporée dans le calcul des vraisemblances de l'algorithme de Viterbi pour les systèmes de reconnaissance à base de HMMs continus. Une évaluation sur une tâche de reconnaissance en présence de bruits blancs [Carlson et Clements, 1991], puis de bruits colorés [Carlson et Clements, 1992] montre la supériorité de la mesure de projection cepstrale par rapport à la distance cepstrale Euclidienne traditionnelle.

La correction de la réduction de la norme proposée par Mansour et Juang est effectuée de façon déterministe. Dans [Juang et Paliwal, 1992], un cadre stochastique est introduit pour effectuer la correction de la norme des vecteurs de cepstre. Sur une tâche de reconnaissance des 10 chiffres, l'amélioration des performances correspond à un gain du SNR d'environ 15-20 dB.

La mesure de projection cepstrale est utilisée par [Lockwood et Boudy, 1992], associée à une soustraction spectrale non linéaire et à un lissage des matrices de covariance. Sur une application de reconnaissance de quelques mots isolés dans une voiture, les résultats obtenus sont satisfaisants. Par contre, sur une tâche similaire, [Hernando et Nadeu, 1994] constatent que la distance de projection cepstrale n'est pas plus efficace que la distance cepstrale Euclidienne. Hernando et Nadeu justifient ce résultat par les caractéristiques du bruit perturbateur (environnement de voiture), très différentes de celles du bruit blanc.

La mesure de projection de Mansour et Juang peut être interprétée comme une transformation linéaire des vecteurs de cepstre bruités [Lee et Wang, 1994]. Dans ce cas, l'effet

du bruit sur les coefficients dynamiques du cepstre (Δcepstre) peut être représenté par une transformation affine, et il est possible de définir une mesure de projection associée. Sur une tâche de reconnaissance de mots isolés dans le bruit, la compensation simultanée des vecteurs de cepstre et de Δcepstre permet alors d'obtenir des performances légèrement supérieures à celles fournies par la compensation du cepstre seul [Lee et Wang, 1994].

Les performances de différentes modélisations tout-pôles combinées à différents liftres cepstraux et à la distance de projection cepstrale sont étudiées pour la reconnaissance de la parole bruitée [Junqua et Wakita, 1989]. Lorsque la parole est produite en environnement calme, la mesure de projection cepstrale associée à une analyse par prédiction linéaire conduit aux meilleurs taux de reconnaissance en mode dépendant du locuteur. Lorsque la parole est produite dans le bruit (effet Lombard), l'analyse PLP-cepstre (cf. § 3) associée à la distance RPS donne les résultats les meilleurs.

[Paliwal et Atal, 1994] comparent les performances de différents paramétrages du signal sur une tâche de reconnaissance de parole en présence d'un bruit additif et de distorsions provoquées par la ligne téléphonique. Ils constatent d'une part que l'analyse par prédiction linéaire permet d'obtenir des coefficients cepstraux plus robustes que l'analyse homomorphique. D'autre part, les coefficients cepstraux calculés à partir du spectre de puissance en sortie d'un banc de filtres sont plus robustes que ceux calculés à partir du spectre de puissance déterminé par FFT[9]. Enfin, l'utilisation d'une échelle de fréquence associée à des propriétés auditives améliore la robustesse au bruit.

Les dérivées d'ordre supérieur de la LPC par rapport au SNR sont utilisées dans [Guan *et al.*, 1993] pour estimer les coefficients LPC « propres », à partir de l'analyse LPC du signal bruité. Cette méthode ne nécessite que la connaissance du rapport bruit-à-signal, et aucune hypothèse n'est faite quant à la distribution du bruit. Pour des SNRs modérés, cette correction des coefficients LPC-cepstre permet d'améliorer les taux de reconnaissance de mots isolés, au prix d'une faible complexité de calculs.

[Lee et Lin, 1993] proposent d'utiliser différents limiteurs de signaux comme prétraitement dans des applications de reconnaissance de parole bruitée. Le limiteur de signal correspond à une transformation non linéaire qui conserve le signe du signal temporel mais ignore son amplitude. Ce traitement est appliqué avant le paramétrage du signal par LPC-cepstre. La motivation est qu'un signal ainsi saturé conserve son intelligibilité, bien que sa qualité soit fortement dégradée. Le signe du signal temporel étant peu affecté par la présence d'un bruit, ce prétraitement se révèle efficace sur une application de reconnaissance de 40 mots isolés en présence d'un bruit blanc ou coloré, avec un système à base de DTW. Par contre, l'application du limiteur dégrade les taux de reconnaissance de la parole propre.

2 Mesures de distorsion et paramétrages discriminants

L'augmentation de la discrimination entre vecteurs de paramètres de parole peut s'effectuer soit en définissant des paramètres discriminants, soit en utilisant des mesures de distorsion discriminantes.

[9] *Fast Fourier Transform*

Les paramètres discriminants sont généralement obtenus par analyse linéaire discriminante (LDA[10]). La LDA est une méthode d'analyse de données permettant d'améliorer la discrimination et d'effectuer une compression d'informations d'un vecteur de paramètres, par application d'une transformation linéaire [Diday *et al.*, 1982]. Le critère habituellement utilisé pour déterminer la transformation est celui qui minimise les variances intra-classes, tout en maximisant les variances inter-classes. Un avantage de la LDA est qu'elle permet de combiner des paramètres hétérogènes (par exemple, paramètres statiques et dynamiques); après transformation, les matrices de covariance intra-classes deviennent des matrices identités, ce qui permet d'utiliser une distance Euclidienne. Il est également possible de ne conserver qu'un sous-ensemble des coefficients du vecteur transformé, afin de réduire la dimension de l'espace de paramètres, sans provoquer pour autant une perte significative d'informations utiles pour la classification [Paliwal, 1992] (cf. chap. 10).

[Doddington, 1989] applique une transformation discriminante spécifique à chaque phonème. Cette transformation minimise la confusion entre les données du corpus d'apprentissage spécifiques au phonème considéré, et les données du corpus d'apprentissage qui peuvent être confondues avec celles de ce phonème. Sur une reconnaissance de chiffres connectés, la transformation discriminante divise le taux d'erreurs par 2. [Haeb-Umbach et Ney, 1992] utilisent la LDA pour projeter un espace de paramètres formé des sorties d'un banc de filtre et de leurs dérivées premières et secondes, sur un espace discriminant de dimension plus réduite. Les classes à discriminer sont associées à des unités acoustiques inférieures au phonème. Sur une application de reconnaissance de parole continue en mode dépendant du locuteur, avec un grand vocabulaire, l'application de la LDA réduit le taux d'erreur de 18% [Haeb-Umbach et Ney, 1992]. En utilisant la LDA, [Roth *et al.*, 1993] observent également une amélioration des performances sur une tâche de reconnaissance de parole continue indépendante du locuteur et pour un grand vocabulaire. Sur une reconnaissance de chiffres connectés [Haeb-Umbach *et al.*, 1993] et d'une centaine de mots enchaînés [Wood *et al.*, 1991], des conclusions analogues sont rapportées.

La LDA est également utilisée en reconnaissance de parole dans le bruit. [Hunt et Lefèbvre, 1988] appliquent une analyse linéaire discriminante pour combiner les sorties d'un modèle auditif. Leur travail est étendu dans [Hunt et Lefèbvre, 1989] et conduit à la définition de l'analyse IMELDA[11]. IMELDA consiste à appliquer une analyse linéaire discriminante sur un ensemble hétérogène de paramètres, extraits des sorties d'un banc de filtres répartis sur l'échelle Mel. Comparé à différents paramétrages (MFCC, distances cepstrales pondérées [Mansour et Juang, 1989], modèle auditif [Hunt et Lefèbvre, 1987]), IMELDA conduit à l'obtention des meilleurs taux de reconnaissance en parole bruitée. Par contre, les MFCC s'avèrent plus efficace que IMELDA en parole propre. Différentes variantes de IMELDA sont proposées dans [Hunt *et al.*, 1991; Lefèbvre *et al.*, 1992].

[Siohan *et al.*, 1994] constatent que la LDA permet d'obtenir un paramétrage efficace pour la reconnaissance de la parole dans le bruit, et que les performances obtenues sont supérieures à celles utilisant différentes approches de compensation de modèles et de transformation d'espace appliquées sur un paramétrage cepstral. Par contre, la nature du bruit conditionne très fortement la robustesse aux variations du SNR des paramètres générés par

[10] *Linear Discriminant Analysis*

[11] *Integrated Mel-scale Linear Discriminant Analysis*

LDA. En présence de bruit blanc, le paramétrage par LDA s'avère très peu robuste (cf. chap. 10) [Siohan, 1995].

[Trompf *et al.*, 1993] génèrent un ensemble de paramètres discriminants par LDA, et effectuent une réduction de la dimension de l'espace initial (LPCC + ΔLPCC + ΔΔLPCC). Une transformation d'espace (espace de parole bruitée vers espace de parole propre) est ensuite effectuée en utilisant un réseau de neurones (cf. chap. 1, § 3), afin de limiter les effets des différences d'environnement. Sur une tâche de reconnaissance de mots isolés dans le bruit, la combinaison de la LDA et de la transformation d'espace permet d'obtenir des résultats meilleurs que ceux obtenus par l'application isolée de la LDA ou de la transformation d'espace.

Dans [Sorensen et Hartmann, 1993], l'analyse IMELDA est combiné avec la soustraction spectrale non linéaire de [Lockwood et Boudy, 1992] pour une application de reconnaissance de mots isolés dans un environnement de voiture. Cette combinaison conduit à l'obtention de meilleurs résultats par rapport à un système utilisant uniquement la soustraction spectrale non linéaire ou IMELDA.

L'augmentation de la discrimination peut être introduite au niveau de la mesure de distance. Étant donné que les zones du signal de forte énergie sont les moins perturbées par le bruit, [Kobatake et Matsunoo, 1994] définissent une distance dans le cadre d'un système de reconnaissance à base de DTW, qui privilégie les chemins où le rapport signal-à-bruit instantané est important. Sur une tâche de reconnaissance de 15 mots isolés, cette distance associée à une pondération RPS fournit les meilleurs résultats. [Anglade *et al.*, 1993] effectuent une reconnaissance de mots difficiles dans le bruit, et focalisent la mesure de distance sur une zone discriminante des mots. Il apparaît par exemple que l'information contenue dans la partie vocalique des mots du "E" *set* intervient peu dans le processus de décision; l'information discriminante n'est située que sur quelques trames de la partie plosive. Le taux de reconnaissance est optimisé pour chaque sous-vocabulaire en ajustant la position de la zone discriminante à utiliser dans le processus de décision. Sur une tâche de reconnaissance de quelques mots isolés en parole propre, Lombard, et Lombard bruitée, cette méthode donne de bons résultats.

3 Paramétrage à base de modèles auditifs

Le système auditif humain est particulièrement résistant aux bruits perturbant le signal de parole. Aussi, dans beaucoup de travaux, des connaissances sur les mécanismes de l'audition sont utilisées pour analyser le signal. Habituellement, on considère que le système auditif se décompose en deux parties. D'une part le système auditif périphérique, qui effectue le codage du signal acoustique en influx nerveux sur les premiers neurones du nerf auditif; d'autre part le système auditif central, qui va des premiers neurones du nerf auditif jusqu'au cortex. Le modèle du système auditif périphérique se décompose en trois parties : tout d'abord une simulation de la sélectivité en fréquence des fibres nerveuses, ensuite une limitation en dynamique et enfin l'adaptation à court terme et la génération des influx nerveux. Le système central comprend en deux ensembles distincts de fibres nerveuses : un système afférent allant de l'oreille au cortex, et un système efférent, allant du cortex à l'oreille, leur interaction étant complexe.

Le lecteur intéressé par les aspects psychoacoustiques pourra se reporter à [Botte *et al.*, 1989; Zwicker et Feldtkeller, 1981].

[Fletcher, 1940] a développé le concept de « filtre auditif », qui permet de modéliser la réponse en fréquence des fibres nerveuses. Le système auditif se comporte comme un banc de filtres se chevauchant, dont les fréquences centrales s'échelonnent continûment, de façon linéaire en basse fréquence, et logarithmique en haute fréquence. Les largeurs de ces filtres sont appelées bandes critiques. L'utilisation d'une échelle psychoacoustique des fréquences comme l'échelle Mel [Zwicker et Feldtkeller, 1981] est désormais courante en RAP. Dans [Davis et Mermelstein, 1980], différents paramétrages du signal de parole (MFCC, LFCC[12], LPCC) sont comparés sur une tâche de détection de mots monosyllabiques dans un flot de parole continue. Les meilleures performances sont obtenues avec le paramétrage MFCC. Ce résultat est confirmé par [Lockwood *et al.*, 1991], où les MFCC conduisent aux meilleurs résultats pour la reconnaissance de mots isolés dans une voiture par rapport aux LPC-cepstres, lorsqu'une méthode de compensation de bruit appropriée est utilisée. [Wang *et al.*, 1993] utilisent une transformation en ondelettes [Meyer, 1992] pour effectuer une analyse à résolution fréquentielle relative constante, suivie d'une compression spectrale non linéaire. Ce traitement, proche de celui réalisé par la cochlée, conduit à une représentation du spectre robuste aux bruits additifs.

[Hermansky *et al.*, 1985; Hermansky, 1990] développent l'analyse par prédiction linéaire perceptive (PLP[13]), qui consiste à effectuer un filtrage en bandes critiques du signal de parole, suivi d'une pré-accentuation à partir de la courbe d'*isosonie*[14] et d'une compression spectrale pour passer de l'intensité à la *sonie*[15]. Le spectre ainsi obtenu est finalement représenté par un modèle tout-pôle, et des coefficients cepstraux peuvent être calculés. [Junqua et Wakita, 1989] montrent que l'analyse par PLP-cepstre associée à la pondération RPS est plus efficace que la LPC-cepstre sur une tâche de reconnaissance de mots isolés prononcés dans le bruit. Sur une tâche d'identification du locuteur, [Xu et Mason, 1989; Xu *et al.*, 1989] constatent également la supériorité de l'analyse PLP sur l'analyse LPC.

Les principaux modèles auditifs utilisés en RAP sont ceux dérivés des modèles de Lyon, Ghitza et Seneff. Le modèle de [Lyon, 1982; Lyon, 1984; Lyon et Dyer, 1986] simule par un contrôle adaptatif de gain le phénomène de limitation en dynamique du système auditif. Le modèle de [Ghitza, 1986; Ghitza, 1987; Ghitza, 1988; Ghitza, 1992] prend un compte le phénomène de *synchronisation des influx*[16]. [Seneff, 1988] incorpore également la synchronisation des influx, qui permet une bonne caractérisation des formants. Ces modèles auditifs se caractérisent par leur grande résolution spectrale, mais les analyses temps-fréquence mises en œuvre conduisent à des charges de calcul importantes.

[Hunt et Lefèbvre, 1986; Hunt et Lefèbvre, 1987; Hunt et Lefèbvre, 1988] développent le modèle de Seneff, et obtiennent une amélioration des taux de reconnaissance de parole bruitée

12. *Linear Frequency Cepstral Coefficient*

13. *Perceptually based Linear Prediction*

14. Les courbes d'*isosonie* relient les niveaux de pression acoustique et fréquence des sons purs qui donnent à l'oreille humaine une égale sensation d'intensité.

15. On appelle *sonie* l'intensité subjective des sons.

16. La *synchronisation des influx* désigne la régularité statistique des intervalles de temps séparant des impulsions isolées sur les fibres nerveuses, corrélées à la fréquence du stimulus acoustique.

en mots isolés, par rapport à un paramétrage MFCC. Le modèle auditif de [Seneff, 1988] est utilisé par [Ohshima et Stern, 1994], et s'avère plus efficace que le paramétrage LPC-cepstre sur une tâche de reconnaissance de parole continue bruitée et filtrée. De plus, après une analyse en composantes principales des sorties du modèle de Seneff, il s'avère qu'il est possible d'utiliser un sous ensemble réduit de ces sorties, sans provoquer une dégradation des taux de reconnaissance. Cependant, la combinaison de l'algorithme CDCN de [Acero, 1992] avec le modèle de Seneff reste moins efficace que l'utilisation de CDCN dans le domaine cepstral, ce qui avait été également rapporté dans [Stern *et al.*, 1992]

Des phénomènes psychoacoustiques plus complexes sont également pris en compte pour améliorer la robustesse des systèmes de RAP. [Cheng et O'Shaughnessy, 1991] tiennent compte du phénomène d'*inhibition latérale*[17] [Shamma, 1985] ; [Cohen, 1985] utilise le modèle d'*adaptation à court terme*[18] de [Schroeder et Hall, 1974]; dans [Aikawa et Saito, 1994], le masquage proactif est incorporé dans un paramétrage cepstral et s'avère efficace à la fois en reconnaissance de parole propre et bruitée. [Gao *et al.*, 1992b] complètent le modèle de Seneff et observent une amélioration des taux de reconnaissance dans le bruit [Gao *et al.*, 1992a].

4 Suppression des variations lentes

La plupart des bruits additifs et des distorsions liées au canal d'enregistrement varient lentement par rapport aux variations du signal de parole. Il est donc possible d'utiliser cette propriété pour définir différents paramétrages insensibles aux variations lentes du signal. Cette opération s'apparente à un filtrage des paramètres, et peut être mise en œuvre dans différents espaces de représentation du signal de parole.

Lorsqu'une phrase est filtrée par l'inverse de l'enveloppe du spectre d'une voyelle, le signal reste intelligible bien que le spectre de la voyelle utilisée soit plat [Morgan *et al.*, 1990]; l'oreille semble donc plus sensible aux variations du spectre du signal de parole, qu'à sa valeur en absolu. Or, la plupart des paramétrages du signal se basent sur les valeurs spectrales absolues, et sont donc très sensibles aux changements de microphone entre les conditions de test et d'apprentissage. Partant de ces constatations, [Hermansky *et al.*, 1991] proposent l'analyse RASTA[19], qui consiste à supprimer les variations lentes du signal. Cette méthode est incorporée à l'analyse PLP [Hermansky *et al.*, 1985; Hermansky, 1990], et consiste à filtrer avec un passe-haut les sorties d'un banc de filtres dans le domaine du logarithme du spectre (afin de supprimer les composantes qui varient lentement), puis à appliquer un opérateur exponentiel pour retourner dans le domaine du spectre de puissance. De nombreux travaux rapportent l'efficacité de RASTA pour compenser les variations de microphone entre test et apprentissage [Hermansky *et al.*, 1991; Hermansky et Morgan, 1992]. Un filtrage semblable peut être effectué dans le domaine des paramètres MFCC [Hirsch *et al.*, 1991], ou encore dans le domaine du PLP-cepstre [Hanson et Applebaum, 1993].

17. Le phénomène d'*inhibition latérale* caractérise le fait que la réponse d'une fibre nerveuse à une excitation peut être affectée par la réponse des fibres nerveuses adjacentes.

18. L'*adaptation à court terme* caractérise les variations des réponses à un son des fibres nerveuses, en fonction des caractéristiques du son précédent, et permet ainsi une prise en compte du contexte.

19. *RelAtive SpecTrAl*

Dans ses premières formulations, RASTA permettait uniquement de compenser des variations de microphone, mais ne luttait pas contre les perturbations provoquées par un bruit additif. Pour atténuer l'influence d'un tel bruit, il est nécessaire d'appliquer le filtrage passe-bas dans le domaine du spectre, et il devient alors difficile d'assurer simultanément une compensation du bruit additif et de la perturbation liée aux variations du canal d'enregistrement. [Hermansky *et al.*, 1993; Morgan et Hermansky, 1992] développent un filtrage passe-bande d'une fonction du spectre, égale à $\log(1 + Jx)$, approximativement logarithmique lorsque l'amplitude x du spectre est élevée, et approximativement linéaire lorsque l'amplitude du spectre est faible. Cette méthode, appelée J-RASTA, permet de compenser principalement les effets du bruit lorsque x est faible, et les effets d'un filtrage linéaire lorsque x est élevé. La valeur optimale de J semble liée au rapport signal-à-bruit instantané [Hermansky *et al.*, 1993], et le coefficient J introduit donc une source de variabilité dans l'analyse. [Koehler *et al.*, 1994] appliquent une transformation empirique linéaire afin de transformer le spectre obtenu avec un J correspondant à de la parole bruitée, pour se rapprocher d'un spectre obtenu avec un J correspondant à de la parole propre. Cette dernière version de J-RASTA-PLP améliore la robustesse aux bruits additifs et aux bruits de convolution. Il faut cependant noter que les analyses de type RASTA augmentent la dépendance au contexte des données, ce qui peut provoquer une dégradation des performances des systèmes de RAP qui utilisent des modèles indépendants du contexte [Koehler *et al.*, 1994].

[Smolders *et al.*, 1994] remarquent que J-RASTA-PLP s'avère plus efficace que la combinaison de RASTA-PLP et la soustraction spectrale de [Berouti *et al.*, 1979] sur une tâche de reconnaissance de 20 mots isolés dans une voiture.

[Hirsch *et al.*, 1991] proposent différents filtrages passe-haut d'enveloppes spectrales dans différentes bandes de fréquences. Un filtre FIR passe-haut permet d'effectuer un filtrage du bruit de fond dans le domaine du module de spectre, tandis qu'un filtre IIR passe-haut, appliqué dans le domaine logarithmique du spectre, améliore la robustesse aux variations de la ligne de transmission, et l'indépendance au locuteur. Une variante des filtres de Hirsch *et al.* ainsi que de Hermansky et Morgan est utilisée par [Dobler *et al.*, 1993], et permet d'augmenter la robustesse à l'environnement sur une tâche de reconnaissance de mots isolés dans une voiture.

La soustraction cepstrale, ou CMN[20] [Atal, 1974], a pour objectif de soustraire à chaque vecteur de cepstre une estimation du cepstre moyen à long terme (calculé sur toute une phrase par exemple). Ce traitement effectue une normalisation de la distribution spectrale de la phrase, ce qui permet de minimiser les variabilités intra-locuteurs [Furui, 1981]. Un tel traitement peut être interprété comme un filtrage inverse qui normalise la réponse en fréquence du canal d'enregistrement, et est donc largement utilisé pour compenser les variations de microphone entre les conditions de test et d'apprentissage des systèmes de RAP.

[Stern *et al.*, 1994] comparent les performances de CMN avec celle de l'algorithme de CDCN [Acero, 1992] sur une tâche de reconnaissance de mots isolés dans une voiture en déplacement. Aux vitesses élevés, les dégradations des performances proviennent des bruits liées au déplacement du véhicule (bruits quasi stationnaires). Aux vitesses faibles, les perturbations proviennent des équipements (radio, essuie-glaces, climatisation, etc.). La méthode

[20] *Cepstral Mean Normalisation*

de CMN améliore les performances dans toutes les conditions, mais l'algorithme CDCN reste cependant plus performant pour les rapport signal-à-bruit les plus faibles. La combinaison de CMN et de CDCN ne conduit pas à une amélioration supplémentaire. Un filtrage adaptatif permet de réduire les erreurs provoquées par la radio lorsque la vitesse du véhicule est faible.

La soustraction cepstrale et le filtrage passe-haut de Hirsch *et al.* sont utilisés par [Mokbel *et al.*, 1993; Mokbel *et al.*, 1994] pour compenser les variations des distorsions apportées par la ligne téléphonique. Le filtrage passe-haut semble légèrement plus efficace que la soustraction cepstrale [Mokbel *et al.*, 1994].

[Anastasakos *et al.*, 1994] comparent les performances de RASTA et de la CMN sur une application de reconnaissance de parole continue grand vocabulaire, lorsque les microphones de test et d'apprentissage diffèrent. Il apparaît que les deux traitements permettent bien de compenser les variations liées à un changement de microphone. Par contre, comparé à une absence de compensation, RASTA dégrade légèrement les performances lorsque les microphones de test et d'apprentissage sont identiques. Sur une application similaire, [Liu *et al.*, 1994] constatent également la dégradation provoquée par RASTA lorsque les microphones sont identiques, et rapportent que la CMN améliore nettement les performances par rapport à RASTA en présence d'une variation de microphones.

Après comparaison de différentes techniques de prétraitement (CMN, RASTA, soustraction spectrale, CDCN), [Chang et Zue, 1994] constatent que la soustraction spectrale suivie de CMN est la méthode la plus efficace lorsque les microphones utilisés pour le test et l'apprentissage sont différents. Par contre, ces méthodes sont peu efficaces lorsque le système est entraîné à partir de parole propre (issue du corpus TIMIT) dont la bande a été limitée, puis testé sur de la parole enregistrée au travers du réseau téléphonique (issue du corpus NTIMIT). La nature des dégradations apportées par le réseau téléphonique ne semble pas pouvoir être modélisée uniquement par un bruit additif et un filtrage linéaire invariant. [Moreno et Stern, 1994] aboutissent à des conclusions analogues : RASTA et CDCN semblent incapable de compenser les distorsions introduites par la ligne téléphonique lorsqu'un système entraîné sur de la parole propre (TIMIT) est testé sur de la parole téléphonique (NTIMIT).

Les paramètres dynamiques semblent jouer un rôle important dans la perception du signal de parole [Furui, 1986a]. Dans [Furui, 1986b; Furui, 1990], les vecteurs de cepstre sont complétés par leurs dérivées premières (Δcepstre), calculées par régression. Sur une tâche de reconnaissance d'une centaine de mots isolés en parole propre, l'introduction des Δcepstres améliore les performances du système. [Hanson et Applebaum, 1990b] utilisent les coefficients dynamiques (Δcepstre) pour la reconnaissance de parole bruitée et de parole Lombard, à partir d'un système entraîné avec de la parole propre. L'introduction des coefficients d'accélération ($\Delta\Delta$cepstre) conduit à une amélioration supplémentaire des performances. Lorsque le bruit est important, il semble également préférable de rejeter les coefficients statiques pour ne conserver que les coefficients dynamiques et d'accélérations. L'ajout des dérivées troisièmes contribue à une amélioration supplémentaire des performances [Hanson et Applebaum, 1990a]. Une discussion sur la largeur des fenêtres d'analyse et sur le nombre de trames à utiliser pour le calcul des dérivées premières et secondes est disponible dans [Applebaum et Hanson, 1991]. Il apparaît que le nombre de trames optimal est différent, selon que l'on effectue une reconnaissance de parole propre ou de parole Lombard bruitée.

5 Conclusion

De nombreuses représentations du signal de parole et mesures de distances ont été proposées ces dernières années. Leurs caractéristiques essentielles sont d'une part de privilégier l'influence des zones du spectre de forte énergie, c.-à-d. celles qui sont les moins perturbées par le bruit, d'autre part, d'exploiter la structure du signal de parole et les effets du bruit sur le signal.

Les mesures de distances et paramétrages spécifiques au bruit sont généralement très efficaces dans le bruit, et les utiliser provoque une amélioration importante des performances par rapport à un paramétrage classique type MFCC, sans nécessiter de modifications de la structure des systèmes de reconnaissance. Les connaissances introduites sur le bruit perturbateur sont souvent minimales et les principales approches ne nécessitent généralement pas de disposer des statistiques du bruit. Par contre, la prise en compte des bruits non stationnaires est difficile à assurer.

En absence de bruit, certains paramétrages spécifiques au bruit s'avèrent moins efficaces que les paramétrages classiques type MFCC. Ceci est généralement le cas pour les paramétrages à base de modèles auditifs, qui sont très appropriés au paramétrage du signal dans le bruit, mais conduisent à une dégradation des performances en absence de bruit, et mettent en œuvre des analyses temps-fréquences coûteuses.

Conclusion

Nous avons présenté dans cette première partie un panorama de la majorité des approches permettant d'améliorer la robustesse des systèmes de RAP dans des environnements bruités. Il est cependant difficile de se prononcer de façon définitive sur la supériorité de telle ou telle méthode. En effet, les systèmes de RAP dans le bruit peuvent être évalués sous différentes conditions qui évoluent continûment de situations réelles, difficiles à contrôler, à des situations contrôlées mais artificielles :

- parole prononcée dans le bruit *vs* parole propre mixée avec un bruit ;
- parole spontanée *vs* parole lue ;
- vocabulaire ouvert *vs* vocabulaire spécifique à l'application.

Pour permettre une comparaison objective des différentes approches de RAP dans le bruit, il est nécessaire que les différents laboratoires adoptent des procédures communes d'évaluation de leurs systèmes. Un élément primordial concerne le développement de corpus de parole spécifiques à la RAP dans des environnements difficiles. Deux approches sont possibles : d'une part, la définition de corpus de parole enregistrés dans le bruit, mais qui se heurte au problème de l'explosion combinatoire des configurations en types et niveaux de bruit ; d'autre part, l'utilisation de corpus de parole propre associés à des bruits et des procédures standards pour générer la parole bruitée. Cela passe en particulier par la définition d'une mesure du rapport signal-à-bruit, devant tenir compte des problèmes comme l'utilisation ou non des zones de silence, l'application de pondérations spécifiques à certaines bandes de fréquences, la définition du SNR en présence de bruits non stationnaires. Si l'ajout de bruit reste une approche valide pour la simulation des environnements faiblement bruités, l'utilisation de corpus enregistrés sous des conditions réelles est indispensable dès que le niveau de bruit devient important (présence d'effet Lombard).

Au regard des différents travaux présentés, quelques conclusions peuvent cependant être dégagées sur les intérêts respectifs de différentes approches. Ainsi, il nous semble important de :

- privilégier dans le processus de décision, les zones de forte énergie du spectre qui sont les moins perturbées par le bruit ;
- appliquer des traitements spécifiques à des classes de sons, le spectre n'étant pas perturbé de façon consistante tout le long du signal ;

- exploiter les corrélations des composantes spectrales intra-trames, ainsi que les corrélations inter-trames au court du temps ;
- développer la prise en compte des bruits non stationnaires, à l'aide de combinaisons de modèles de parole et bruit, et de filtrages multi-voies ;
- utiliser des connaissances *a priori* sur la parole et le bruit dans les processus de filtrage ou de compensation de modèles ;
- utiliser des connaissances sur les mécanismes de l'audition, et des critères d'estimation significatifs par rapport à la perception humaine ;
- développer les algorithmes auto-adaptatifs, pour permettre aux systèmes d'évoluer automatiquement en fonction des conditions d'environnement.

Nous avons donc poursuivi quelques directions de recherche suggérées par ces conclusions. Trois approches sont présentées puis comparées expérimentalement sur une même application dans la partie III. Tout d'abord, dans le chapitre 6, nous combinons des modèles de parole propre et de bruit pour construire des modèles de parole bruitée. Ces modèles sont directement utilisés pour reconnaître la parole bruitée. Ensuite, le chapitre 7 décrit une approche de filtrage du signal bruité, spécifique à chaque phonème, et optimisé selon un critère perceptif. Le filtrage est associé au système de reconnaissance, et est appliquée pendant la reconnaissance de la parole bruitée. Enfin, dans le chapitre 8, les modèles de parole propre sont adaptés à l'environnement de test en utilisant des transformations linéaires, associées à des classes de sons, et définies selon un critère objectif. Cela permet de construire des modèles spécifiques au nouvel environnement de test.

La partie IV regroupe deux approches indépendantes ne pouvant être directement comparées avec les trois méthodes précédentes car évaluées dans des configurations différentes. L'analyse linéaire discriminante est utilisée au chapitre 10 pour tenter de définir un paramétrage du signal de parole robuste au bruit. La reconnaissance de la parole bruitée s'effectue alors sans modification du système de reconnaissance de parole propre initial. Dans le chapitre 11, nous adaptons les modèles de durée des phonèmes pour lutter contre les modifications du rythme d'élocution provoquées par l'effet Lombard.

Ces différentes approches ont été évaluées après intégration dans le système de reconnaissance de parole continue VINICS, décrit en partie II.

Deuxième partie

Le système VINICS

Introduction

En reconnaissance de parole continue, il est essentiel de modéliser une unité de parole dans un contexte phonétique complexe. En effet, l'inertie du système vocal, associée à la tendance naturelle du locuteur à minimiser son effort articulatoire, peut conduire à la production d'un son neutre à partir de séquences phonétiques différentes. Ce phénomène est schématiquement représenté fig. 3.1, où le triangle articulatoire /a/-/u/-/i/[21] est esquissé. Il apparaît que le locuteur désirant produire la séquence /a/-/u/-/i/ va en fait produire une séquence déformée se rapprochant de /a/-/ϕ/-/i/, la cible /u/ n'étant pas atteinte. C'est donc plus l'*évolution* des paramètres du signal de parole dans un espace donné — la trajectoire — que la position absolue des unités de parole dans cet espace qui permet de prendre en compte les caractéristiques de la parole continue.

Le modèle de Markov caché ou HMM[22] [Rabiner, 1989] représente le signal de parole comme étant issu d'un automate à états. Les transitions entre états sont régies par des distributions de probabilité et modélisent les fluctuations du rythme d'élocution. Chaque état est générateur de vecteurs d'observation de parole, selon une fonction de densité de probabilité plus ou moins complexe, et modélise ainsi la variabilité acoustique du signal. Le formalisme de HMM implique que les vecteurs d'observations générés par les états sont décorrélés. De plus, la probabilité de la durée de séjour dans un état donné est exponentiellement décroissante. De telles contraintes ne rendent pas compte de la réalité du signal de parole, et ne permettent pas de modéliser correctement les séquences de vecteurs associées à une unité de parole donnée, ni les durées de ces séquences. Nous illustrons ce propos par l'exemple suivant, tiré de [Gong, 1994] et schématisé fig. 3.2.

Soit un HMM continu gauche-droite à 3 états, avec un mélange de 2 lois normales par état, et modélisant le son associé au phonème s. Supposons que les successions de vecteurs de parole associées à s et issues du corpus d'apprentissage se décomposent en 2 grandes classes : d'une part les successions fluctuant autour de la séquence a-x-b, d'autre part les successions fluctuant autour de c-x-d. Une des lois normales associée au premier état du HMM va donc modéliser les fluctuations autour de a, l'autre les variations autour de c. Les 2 lois de l'état central vont modéliser les distributions des observations autour de x et enfin, une des lois normales du dernier état va modéliser les variations autour de b, l'autre les variations autour de d. Lors de la reconnaissance, supposons qu'une trajectoire spécifique au symbole s' et fluctuant autour de a-x-d soit observée. La probabilité que cette trajectoire soit issue du modèle de s sera très élevée car ce modèle représente bien les fluctuations autour de a, de x et de d.

---

21. /a/ → l<u>a</u>, /u/ → l<u>ou</u>p, /i/ → l<u>i</u>t, /ϕ/ → l<u>eu</u>

22. *Hidden Markov Model*

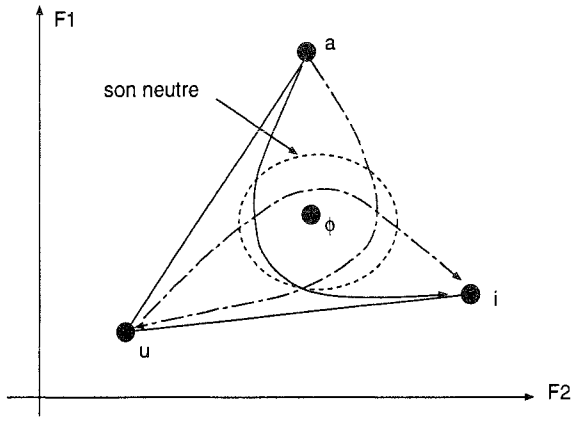

FIG. 3.1 - *Évolution d'une trajectoire de parole continue dans le plan F2–F1. Les cibles situées au centre des trajectoires ne sont pas atteintes.*

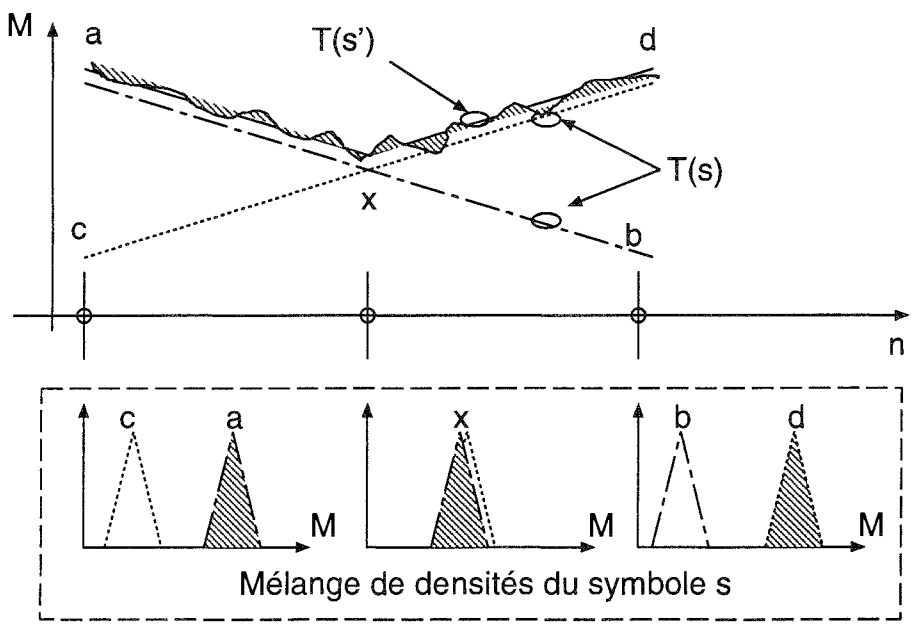

FIG. 3.2 - *Regroupement de trajectoires par un HMM. Une unité de parole s est représentée par les 2 classes de trajectoires a-x-b et c-x-d. Le HMM reconnaît la trajectoire hachurée a-x-d comme étant caractéristique de s, bien que ne faisant pas partie des classes d'apprentissage spécifiques à s [Gong, 1994].*

Pourtant, une telle trajectoire n'était pas présente dans le corpus d'apprentissage du symbole *s*, où seules les trajectoires a-x-b et c-x-d étaient observées. Le HMM a donc modélisé les séquences a-x-b, c-x-d, mais aussi a-x-d et c-x-b, ce qui n'était pas souhaité. Ce phénomène, appelé «recouvrement» de trajectoires, se produit car le formalisme du HMM n'impose pas de corrélations entre les vecteurs d'observations.

Différentes méthodes ont été proposées pour réduire le recouvrement des trajectoires. Leurs objectifs sont d'augmenter la corrélation entre les trames successives du signal, par exemple en complétant un vecteur de paramètres par ses dérivées premières et secondes par rapport au temps [Furui, 1986b]. [Wellekens, 1987] exploite explicitement les corrélations temporelles du signal en conditionnant la distribution de l'observation courante par l'observation précédente, dans un système à base de HMMs continus. Des idées analogues sont développées par [Paliwal, 1993] pour des HMMs discrets, ainsi que par [Takahashi *et al.*, 1993]. L'utilisation de HMMs du second ordre [Gong *et al.*, 1994] permet également d'obtenir une meilleure modélisation de la structure temporelle du signal, tout comme la mise en œuvre de HMMs autorégressifs [Juang et Rabiner, 1985]. Afin d'éviter le recouvrement des trajectoires, il est possible de modéliser explicitement une unité de parole sous la forme d'une séquence de vecteurs d'observations. On peut alors définir un modèle stochastique optimisé au niveau de la séquence de vecteurs, et offrant une bonne représentation des variabilités contextuelles [Ostendorf et Roukos, 1989; Roucos *et al.*, 1988; Digalakis *et al.*, 1992; Lee, 1989; Roukos et Dunham, 1987].

Dans l'espace des trajectoires, plusieurs phénomènes apparaissent.

- Les trajectoires associées à une unité de parole se répartissent en classes en fonction du contexte acoustique, comme l'illustre la figure 3.3. Sur cette figure sont représentées dans le plan C2–C3 (coefficient cepstral 2 et 3), un ensemble de trajectoires réelles d'un locuteur donné, associées au phonème /*m*/. Après application d'un algorithme de classification automatique de l'espace des trajectoires en 2 classes, il apparaît que les trajectoires représentées se répartissent effectivement en 2 zones distinctes dans le plan C2–C3. Définir un modèle qui prenne en compte cette notion de classe de trajectoires permettra de représenter avec plus de précision l'ensemble des trajectoires associées à une unité de parole.

- La variance des observations situées au centre d'une trajectoire est plus faible que celle des observations situées aux extrémités. La figure 3.4 représente les variances de chaque coefficient cepstral d'un ensemble de trajectoires associées au phonème /*õ*/. Les durées des trajectoires ont été normalisées pour obtenir un ensemble de séquences de 5 vecteurs (ou états). Pour chaque coefficient cepstral, la variance est normalisée à 1 dans l'état où elle est minimale. Il apparaît que les variances des coefficients cepstraux sont plus importantes aux extrémités des trajectoires qu'en leurs centres, les variances de la fin des trajectoires étant elles-mêmes plus élevées que celles du début des trajectoires. Il semble alors souhaitable de privilégier lors du processus de reconnaissance, la contribution du centre des trajectoires par rapport à celle des extrémités.

Ces différents phénomènes ne sont pas explicitement pris en compte dans les modèles de trajectoires cités précédemment. De plus, les contraintes sur la durée des trajectoires ne sont pas toujours utilisées.

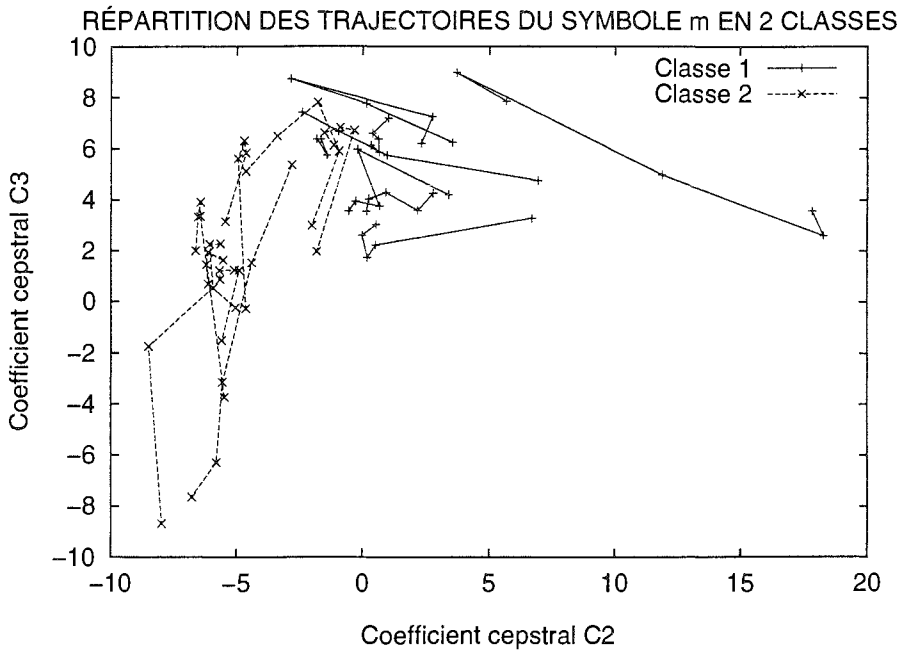

FIG. 3.3 - *Représentation d'un ensemble de trajectoires spécifiques au symbole /m/ dans le plan C2–C3. Les trajectoires forment des classes dans l'espace.*

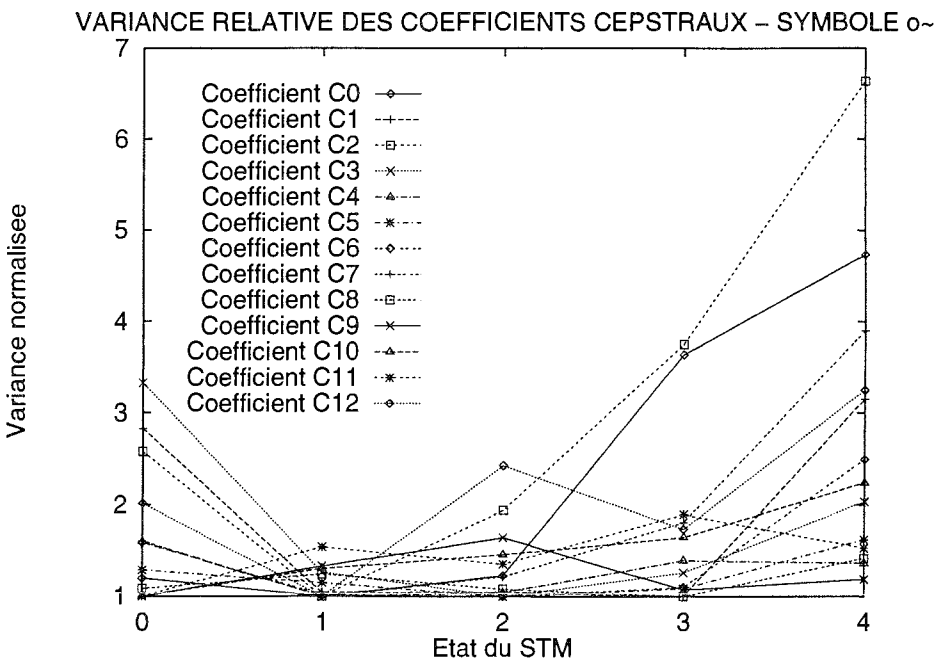

FIG. 3.4 - *Variances des coefficients cepstraux C0 à C12 d'un ensemble de vecteurs d'observations associés au symbole /õ/ en fonction de leurs positions dans la trajectoire de parole. Pour chaque coefficient cepstral, les variances sont normalisées par rapport à leur valeur minimale.*

Les motivations ayant conduit à la définition du modèle stochastique de trajectoires (STM[23]) sont donc les suivantes.

- La notion de trajectoire de parole est plus importante que la position d'une observation dans l'espace (un même point pouvant appartenir à des trajectoires différentes).
- En fonction du contexte acoustique, les trajectoires associées à une même unité de parole forment des classes dans l'espace des trajectoires (cf. fig. 3.3).
- La durée d'une unité de parole permet de définir des modèles plus précis et les contraintes de durée facilitent le décodage d'une séquence d'unités de parole.
- La variance des vecteurs d'observations est plus faible au centre de la trajectoire qu'aux extrémités (cf. fig. 3.4) ; la contribution de la zone centrale doit donc être privilégiée par rapport à celle des zones latérales.

Le formalisme des STMs a été utilisé pour construire différents systèmes de reconnaissance de parole continue [Gong et Haton, 1991; Gong *et al.*, 1991], évalués sur diverses applications [Gong et Haton, 1994; Gong *et al.*, 1994]. Nous nous limitons ici à la description de la version 3 du système (VINICS-III), utilisée pour la majorité de nos expériences. Des descriptions des versions antérieures sont disponibles dans [Gong et Haton, 1991; Gong *et al.*, 1991; Gong et Haton, 1994].

23. *Stochastic Trajectory Model*

Chapitre 4

Modélisation stochastique des trajectoires de parole

1 Trajectoire de parole

Dans un espace de paramètres spécifiques à la parole, le signal de parole est un point qui se déplace lorsque l'articulation évolue. En se déplaçant, ce point décrit une certaine trajectoire dont l'expression analytique est inconnue. Cette trajectoire peut être considérée comme une réalisation d'une fonction aléatoire, ou plus simplement comme une réalisation interpolée d'une séquence de vecteurs aléatoires. Cette séquence sous-jacente de vecteurs aléatoires constitue le modèle de l'ensemble des trajectoires associées à une unité de parole donnée. Bien évidemment, la notion de trajectoire peut caractériser différentes unités de paroles : phonèmes, syllabes, mots, etc., le signal de parole étant constitué d'une concaténation d'unités élémentaires. La reconnaissance de la parole consiste alors à rechercher la succession de modèles expliquant au mieux, selon un certain critère, le signal de parole observé. Dans la suite de la présentation, nous supposons que l'unité de parole associée à la trajectoire est le phonème.

2 Définition du modèle stochastique de trajectoires

2.1 Introduction

Soit $\mathbb{P} \triangleq \{s_1, s_2, \ldots s_H\}$ un ensemble de H symboles représentant des phonèmes. Nous souhaitons définir un modèle caractérisant les trajectoires associées à un symbole s issu de $\mathbb{P}$. Soit un ensemble de N_s trajectoires, spécifiques au symbole s (cf. fig. 4.1). Chaque trajectoire $\mathbf{Y}_i$ est constituée d'une succession de d_i vecteurs, dans un espace de dimension D spécifique à la parole (par exemple, l'espace cepstral). Les d_i vecteurs constituant la trajectoire $\mathbf{Y}_i$ sont notés $\mathbf{y}_0, \ldots \mathbf{y}_{d_i-1}$. Par la suite, d_i sera appelé la durée de la trajectoire $\mathbf{Y}_i$. Ces N_s trajectoires doivent être considérées comme autant de réalisations d'une trajectoire aléatoire sous-jacente qui constitue le modèle.

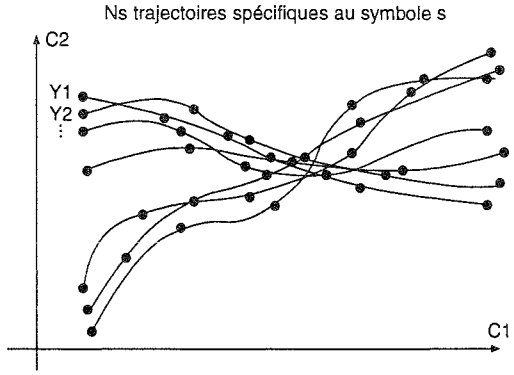

FIG. 4.1 - *N_s trajectoires représentées dans un espace de dimension 2. Chaque trajectoire est constituée de d_i vecteurs, représentés par les points sombres. Les points sont espacés d'une période d'analyse du signal.*

Définir un modèle pour ces trajectoires pose deux problèmes majeurs :

- la prise en compte des variations de durée des trajectoires ;
- la modélisation d'une distribution complexe des trajectoires dans l'espace des trajectoires.

2.2 Rééchantillonnage des trajectoires

Il est plus simple de modéliser sous un cadre stochastique des trajectoires ayant toutes la même durée, plutôt que des trajectoires de durées différentes. Aussi, la première étape de la modélisation consiste à rééchantillonner chaque trajectoire en un nombre fixe de points. Une trajectoire devient alors une séquence de Q vecteurs. Dans le modèle STM, Gong choisit d'effectuer le rééchantillonnage linéairement, de la façon suivante. Soit $\mathbf{Y}$ une trajectoire observée de durée d trames, la trajectoire rééchantillonnée $\mathbf{X}$ s'écrit :

$$\mathbf{Y} = (\mathbf{y}_0, \ldots \mathbf{y}_{d-1}) \longrightarrow \mathbf{X} = (\mathbf{x}_0, \ldots \mathbf{x}_{Q-1}) \qquad \text{avec } \mathbf{x}_i = \mathbf{y}_{i \times \frac{d-1}{Q-1}}, \qquad 0 \leq i < Q \tag{4.1}$$

Si $d < Q$, le rééchantillonnage correspond à un sur-échantillonnage de $\mathbf{Y}$, ce qui signifie que certains vecteurs de $\mathbf{Y}$ peuvent être répétés plusieurs fois dans $\mathbf{X}$ (cf. fig. 4.2). Dans le cas contraire où $d > Q$, la trajectoire $\mathbf{Y}$ va être sous-échantillonnée, ce qui signifie que certains vecteurs constituant $\mathbf{Y}$ n'apparaîtront pas dans $\mathbf{X}$ (cf. fig. 4.2).

En pratique, le rééchantillonnage est effectué en $Q = 5$ points, l'intervalle de temps entre 2 trames consécutives $\mathbf{y}_i$ et $\mathbf{y}_{i+1}$ étant de 10 ms. Il faut remarquer que le sous-échantillonnage n'introduit pas une perte significative d'informations. En effet, le sous-échantillonnage se produit pour les symboles de durée moyenne supérieure à 5 trames, ce qui correspond aux sons stationnaires (voyelles, fricatives) dont les propriétés varient lentement d'une trame à l'autre. Dans ce contexte, rejeter une trame ne se traduit pas par une perte significative d'informations.

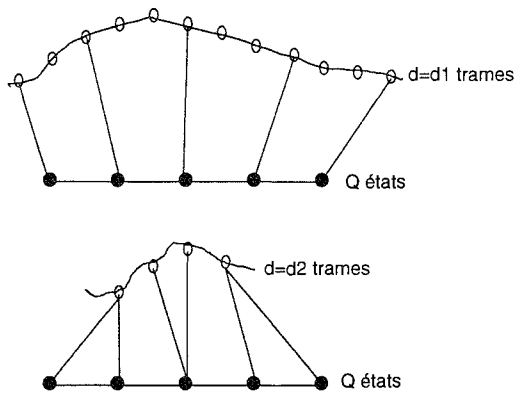

FIG. 4.2 - *Sur-, ou sous-échantillonnage des trajectoires. Les trajectoires de durée d_1 et d_2 sont rééchantillonnées en Q points.*

Différentes stratégies de rééchantillonnage peuvent être envisagées : rééchantillonnage dans l'espace (les points rééchantillonnés sont équidistants le long de la trajectoire), rééchantillonnage linéaire dans le temps avec ou sans interpolation [Ostendorf et Roukos, 1989], rééchantillonnage non linéaire dans le temps [Afify *et al.*, 1994].

Le rééchantillonnage fait disparaître l'information sur la durée des trajectoires. Aussi, dans le formalisme du STM, un modèle de durée, dont nous présenterons l'utilisation par la suite, est introduit pour chaque symbole. La durée d'un symbole s est modélisée par une variable aléatoire $\tilde{d}$, suivant une loi Γ.

2.3 Modélisation des trajectoires rééchantillonnées

Les N_s trajectoires rééchantillonnées $\mathbf{X}_i$ associées au symbole s doivent être considérées comme autant de réalisations d'une trajectoire aléatoire de durée fixe égale à Q points. Cette trajectoire aléatoire constitue le modèle du symbole s, défini par sa fonction de densité de probabilité.

Nous appelons désormais «état» chaque point de la trajectoire aléatoire. À chaque état est associé un vecteur aléatoire, et la trajectoire aléatoire est donc constituée d'une séquence de Q vecteurs aléatoires. Notons $\tilde{\mathbf{x}}_0$ à $\tilde{\mathbf{x}}_{Q-1}$ ces Q vecteurs aléatoires. $p_{\tilde{\mathbf{x}}_0,\ldots\tilde{\mathbf{x}}_{Q-1}}(\mathbf{x}_0,\ldots,\mathbf{x}_{Q-1})$ désigne la *pdf* jointe du Q-uplet $\tilde{\mathbf{X}} = (\tilde{\mathbf{x}}_0,\ldots\tilde{\mathbf{x}}_{Q-1})$.

Nous supposons que les Q vecteurs aléatoires associés à chacun des Q états sont indépendants. La *pdf* de $\tilde{\mathbf{X}}$ peut alors s'écrire :

$$p_{\tilde{\mathbf{X}}}(\mathbf{X}) = p_{\tilde{\mathbf{x}}_0,\ldots\tilde{\mathbf{x}}_{Q-1}}(\mathbf{x}_0,\ldots,\mathbf{x}_{Q-1}) = \prod_{i=0}^{Q-1} p_{\tilde{\mathbf{x}}_i}(\mathbf{x}_i) \tag{4.2}$$

Le vecteur aléatoire associé à l'état i de la trajectoire du symbole s est modélisé par une

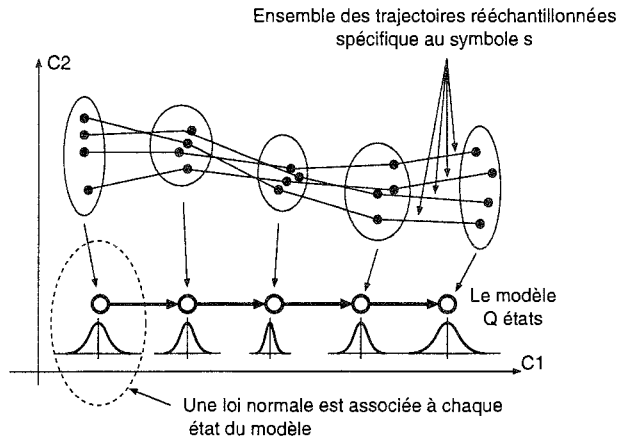

FIG. 4.3 - *Les trajectoires rééchantillonnées sont modélisées par une distribution normale.*

loi normale multidimensionnelle de moyenne $\mathbf{m}_i^s$ et de matrice de covariance $\mathbf{\Sigma}_i^s$:

$$\begin{aligned} p_{\tilde{\mathbf{x}}_i}(\mathbf{x}_i) &\triangleq \mathcal{N}(\mathbf{x}_i; \mathbf{m}_i^s, \mathbf{\Sigma}_i^s) \\ &= \frac{1}{(2\pi)^{D/2}|\mathbf{\Sigma}_i^s|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}_i - \mathbf{m}_i^s)^{\#}\mathbf{\Sigma}_i^{s-1}(\mathbf{x}_i - \mathbf{m}_i^s)\right] \end{aligned} \tag{4.3}$$

où $\mathbf{x}^{\#}$ représente le vecteur transposé de $\mathbf{x}$. Cette modélisation est illustrée figure 4.3.

2.4 Amélioration du modèle

Selon le contexte phonétique, les trajectoires associées à un symbole s occupent des zones distinctes dans l'espace des trajectoires. Ainsi, plusieurs classes de trajectoires peuvent être associées à un même symbole s en fonction du contexte, comme le schématise la figure 4.4.

Pour prendre en compte une distribution complexe des trajectoires dans l'espace des trajectoires, Gong choisit de représenter la séquence aléatoire par un mélange de trajectoires. Cette notion de mélange constitue la principale différence entre le STM et le modèle de trajectoires de Ostendorf et Roukos [Ostendorf et Roukos, 1989]. La densité de probabilité de la séquence aléatoire $\tilde{\mathbf{X}}$ s'écrit :

$$p_{\tilde{\mathbf{X}}}(\mathbf{X}) \triangleq \sum_{t_k \in \mathbb{T}_s} Pr(\tilde{t} = t_k) \cdot p_{\tilde{\mathbf{X}}|\tilde{t}}(\mathbf{X}|t_k) \tag{4.4}$$

où :

- $\tilde{t}$ désigne la variable aléatoire représentant la classe de la trajectoire et prenant ses valeurs dans l'ensemble des classes de trajectoires $\mathbb{T}_s = \{t_k\}$ associées au symbole s ;
- $Pr(\tilde{t} = t_k)$ représente la probabilité *a priori* que la classe de trajectoires soit t_k ;
- $p_{\tilde{\mathbf{X}}|\tilde{t}}(\mathbf{X}|t_k)$ représente la densité de probabilité conditionnelle de la trajectoire $\tilde{\mathbf{X}}$ étant donné la classe de trajectoires t_k.

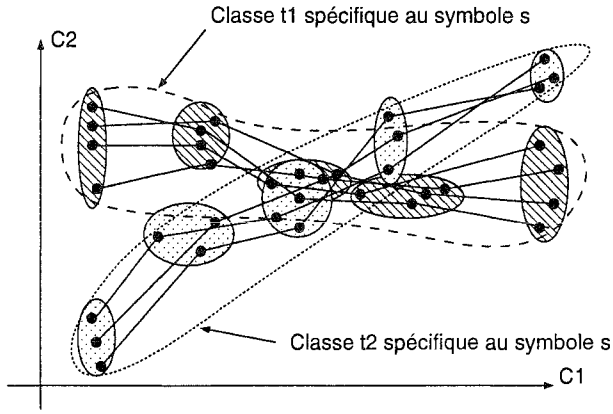

FIG. 4.4 - *Les trajectoires d'un même symbole s se répartissent en classes en fonction du contexte acoustique.*

La classe de trajectoires t_k étant fixée, $p_{\tilde{\mathbf{X}}|\tilde{t}}(\mathbf{X}|t_k)$ est défini comme précédemment, à savoir :

$$\begin{aligned} p_{\tilde{\mathbf{X}}|\tilde{t}}(\mathbf{X}|t_k) &= \prod_{i=0}^{Q-1} p_{\tilde{\mathbf{x}}_i|\tilde{t}}(\mathbf{x}_i|t_k) \\ &= \prod_{i=0}^{Q-1} \mathcal{N}(\mathbf{x}_i; \mathbf{m}_{k,i}^s, \Sigma_{k,i}^s) \end{aligned} \tag{4.5}$$

où $\mathbf{m}_{k,i}^s$ et $\Sigma_{k,i}^s$ désignent respectivement le vecteur moyenne et la matrice de covariance dans l'état i pour la classe t_k de trajectoires du symbole s.

3 Utilisation du modèle en reconnaissance

Jusqu'à présent, nous avons vu comment obtenir le modèle de la trajectoire aléatoire associée à un symbole s. Plusieurs étapes ont été nécessaires pour définir le modèle. Tout d'abord, un rééchantillonnage des trajectoires a été réalisé afin d'obtenir un ensemble de séquences composées de Q vecteurs. Ces séquences sont ensuite considérées comme autant de réalisations d'une trajectoire aléatoire sous-jacente qui constitue le modèle. On considère que cette séquence est constituée de Q vecteurs aléatoires indépendants, chacun d'entre-eux étant représenté par une loi normale multidimensionnelle. Afin de représenter une distribution complexe des séquences dans l'espace des trajectoires, la notion de classe de trajectoires est alors introduite. Le modèle final est constitué d'un mélange de fonctions de densités de probabilités représentant la distribution d'une trajectoire dans une classe donnée sous la forme d'un produit de lois normales multidimensionnelles.

En reconnaissance, le but est de rechercher la succession de symboles dont les modèles expliquent au mieux, selon un certain critère, le signal observé. À l'instant n, il va donc falloir faire des hypothèses sur l'identité du symbole courant s et sur la durée d de la trajectoire

associée à ce symbole. La durée et le symbole sont considérés comme des variables aléatoires, notées $\tilde{d}$ et $\tilde{s}$. Notons $p_{\tilde{\mathbf{X}},\tilde{d},\tilde{s}}(\mathbf{X}_n, d, s)$ la *pdf* conjointe d'une trajectoire rééchantillonnée de Q points, centrée à l'instant n, issue d'une trajectoire de d trames d'un symbole s. D'après la règle de développement en chaîne des probabilités, cette quantité peut s'écrire :

$$p_{\tilde{\mathbf{X}},\tilde{d},\tilde{s}}(\mathbf{X}_n, d, s) = p_{\tilde{\mathbf{X}}|\tilde{d},\tilde{s}}(\mathbf{X}_n|d, s) \cdot p_{\tilde{d}|\tilde{s}}(d|s) \cdot Pr(\tilde{s} = s) \tag{4.6}$$

où :

- $p_{\tilde{\mathbf{X}}|\tilde{d},\tilde{s}}(\mathbf{X}_n|d, s)$ désigne la *pdf* conditionnelle d'une trajectoire rééchantillonnée centrée en n, connaissant la durée d de la trajectoire observée et le symbole associé s ;
- $p_{\tilde{d}|\tilde{s}}(d|s)$ représente la *pdf* de la durée $\tilde{d}$ de la trajectoire, sachant que le symbole est s ;
- $Pr(\tilde{s} = s)$ représente la probabilité *a priori* du symbole s.

Dans les paragraphes précédents, le modèle défini correspondait en fait à $p_{\tilde{\mathbf{X}}|\tilde{d},\tilde{s}}(\mathbf{X}_n|d, s)$. La connaissance de d et s était considérée comme implicite : les N trajectoires étaient associées au symbole s et la connaissance de la durée d de chaque trajectoire avait permis de passer d'un ensemble de N trajectoires observées $\mathbf{Y}_i$ à N trajectoires rééchantillonnées $\mathbf{X}_i$.

Nous avons donc :

$$p_{\tilde{\mathbf{X}}|\tilde{d},\tilde{s}}(\mathbf{X}_n|d, s) \triangleq \sum_{t_k \in \mathbb{T}_s} Pr(\tilde{t} = t_k|\tilde{d} = d, \tilde{s} = s) \cdot p_{\tilde{\mathbf{X}}|\tilde{t},\tilde{d},\tilde{s}}(\mathbf{X}_n|t_k, d, s) \tag{4.7}$$

où :

- $\mathbb{T}_s$ représente l'ensemble des classes possibles des trajectoires aléatoires associées au symbole s ;
- $Pr(\tilde{t} = t_k|\tilde{d} = d, \tilde{s} = s)$ représente la probabilité que la classe de trajectoires soit t_k ; connaissant le symbole s et sa durée d. Par hypothèse, on considère que la probabilité de la classe est indépendante de la durée de la trajectoire. Par conséquent :

$$Pr(\tilde{t} = t_k|\tilde{d} = d, \tilde{s} = s) = Pr(\tilde{t} = t_k|\tilde{s} = s) \tag{4.8}$$

- $p_{\tilde{\mathbf{X}}|\tilde{t},\tilde{d},\tilde{s}}(\mathbf{X}_n|t_k, d, s)$ est la *pdf* de la trajectoire rééchantillonnée centrée en n, connaissant la classe de trajectoires t_k, s et d.

Finalement, on obtient :

$$\begin{aligned} p_{\tilde{\mathbf{X}}|\tilde{t},\tilde{d},\tilde{s}}(\mathbf{X}_n|t_k, d, s) &= \prod_{i=0}^{Q-1} p_{\tilde{\mathbf{x}}_i|\tilde{t},\tilde{d},\tilde{s}}(\mathbf{y}_{n-\frac{Q}{2}+i\frac{d-1}{Q-1}}|t_k, d, s) \\ &= \prod_{i=0}^{Q-1} \mathcal{N}(\mathbf{y}_{n-f(i,d,Q)}; \mathbf{m}^s_{k,i}, \mathbf{\Sigma}^s_{k,i})^{\omega^s_i} \end{aligned} \tag{4.9}$$

où $f(i, d, Q)$ représente le calcul de l'indice permettant de passer d'une séquence $\mathbf{Y}$ de d vecteurs centrés en n à une séquence $\mathbf{X}$ de Q vecteurs. La loi normale est pondérée par un

exposant ω_i^s, utilisé pour augmenter l'influence des états centraux du modèle par rapport aux états latéraux.

Déterminons maintenant l'expression de la fonction de densité de probabilité *a posteriori* du symbole $\tilde{s}$ connaissant la trajectoire rééchantillonnée à l'instant n et la durée d : $p_{\tilde{s}|\tilde{\mathbf{X}},\tilde{d}}(s|\mathbf{X}_n, d)$. Cette probabilité sera utilisée pendant la phase de recherche de la meilleure séquence de symboles constituant un signal donné. On a :

$$\begin{aligned} p_{\tilde{s}|\tilde{\mathbf{X}},\tilde{d}}(s|\mathbf{X}_n, d) &= \frac{p_{\tilde{\mathbf{X}},\tilde{d},\tilde{s}}(\mathbf{X}_n, d, s)}{p_{\tilde{\mathbf{X}},\tilde{d}}(\mathbf{X}_n, d)} \\ &= \frac{p_{\tilde{\mathbf{X}},\tilde{d},\tilde{s}}(\mathbf{X}_n, d, s)}{\sum_{s' \in \mathbb{P}} p_{\tilde{\mathbf{X}},\tilde{d},\tilde{s}}(\mathbf{X}_n, d, s')} \end{aligned} \tag{4.10}$$

$p_{\tilde{s}|\tilde{\mathbf{X}},\tilde{d}}(s|\mathbf{X}_n, d)$ peut se réécrire en utilisant l'équation (4.6), et en introduisant un facteur de pondération λ sur $p_{\tilde{d}|\tilde{s}}(d|s)$ et γ sur $Pr(\tilde{s} = s)$:

$$p_{\tilde{\mathbf{X}},\tilde{d},\tilde{s}}(\mathbf{X}_n, d, s) = p_{\tilde{\mathbf{X}}|\tilde{d},\tilde{s}}(\mathbf{X}_n|d, s) \cdot p_{\tilde{d}|\tilde{s}}(d|s)^{\lambda} \cdot Pr(\tilde{s} = s)^{\gamma} \tag{4.11}$$

Ces pondérations permettent de modifier l'influence de la durée des symboles et la distribution *a priori* des symboles dans le processus de reconnaissance. Leur utilisation permet d'améliorer de façon significative les taux de reconnaissance [Gong, 1994]. En définitive, la densité de probabilité *a posteriori* s'écrit :

$$p_{\tilde{s}|\tilde{\mathbf{X}},\tilde{d}}(s|\mathbf{X}_n, d) = \frac{p_{\tilde{\mathbf{X}}|\tilde{d},\tilde{s}}(\mathbf{X}_n|d, s) \cdot p_{\tilde{d}|\tilde{s}}(d|s)^{\lambda} \cdot Pr(\tilde{s} = s)^{\gamma}}{\sum_{s' \in \mathbb{P}} p_{\tilde{\mathbf{X}}|\tilde{d},\tilde{s}}(\mathbf{X}_n|d, s') \cdot p_{\tilde{d}|\tilde{s}}(d|s')^{\lambda} \cdot Pr(\tilde{s} = s')^{\gamma}} \tag{4.12}$$

Cette densité *a posteriori* sera utilisée dans l'étape de recherche de la meilleure séquence de symboles dans la partie 5.

4 Estimation du modèle

Pour procéder à la reconnaissance, il est nécessaire d'estimer les différents termes de l'équation (4.12), à savoir la distribution *a priori* des symboles $Pr(\tilde{s})$, les paramètres de la *pdf* de la durée connaissant le symbole $p_{\tilde{d}|\tilde{s}}(d|s)$ et enfin les paramètres de la *pdf* de la trajectoire rééchantillonnée connaissant le symbole et la durée $p_{\tilde{\mathbf{X}}|\tilde{d},\tilde{s}}(\mathbf{X}|d, s)$.

Les paramètres de ces différentes *pdfs* et distributions sont estimées selon le critère du maximum de vraisemblance (MLE) à partir d'un corpus d'apprentissage étiqueté au niveau phonétique.

4.1 Estimation de la distribution *a priori* des symboles $Pr(\tilde{s})$

Si N désigne le nombre total d'occurrences de tous les symboles, et N_s le nombre total d'occurrences du symbole s dans le corpus d'apprentissage, alors la distribution *a priori* des symboles est estimée par la distribution empirique :

$$Pr(\tilde{s} = s) = \frac{N_s}{N} \qquad \forall s \in \mathbb{P} \tag{4.13}$$

4.2 Estimation du modèle de durée $p_{\tilde{d}|\tilde{s}}(d|s)$

La fonction de densité de probabilité de la durée $\tilde{d}$ connaissant le symbole s, $p_{\tilde{d}|\tilde{s}}(d|s)$, est représentée par une loi Γ, de paramètres α_s et p_s, définie par :

$$p_{\tilde{d}|\tilde{s}}(d|s) \triangleq \frac{\alpha_s^{p_s} \cdot d^{p_s} \cdot \exp(-\alpha_s \cdot d)}{\Gamma(p_s)} \quad \text{avec } d \geq 0, p_s > 0 \text{ et } \alpha_s > 0 \tag{4.14}$$

et où $\Gamma(p)$ représente la fonction Gamma, définie par :

$$\Gamma(p) = \begin{cases} 1 & \text{si } p = 1 \\ (p-1)\Gamma(p-1) & \text{si } p > 1 \end{cases} \tag{4.15}$$

Nous souhaitons estimer les 2 paramètres de la loi Γ, α_s et p_s, à partir des N_s observations des durées d_i ($1 \leq i \leq N_s$) des symboles s, issus du corpus d'apprentissage. L'estimation s'effectue selon le critère MLE, à savoir :

$$(\hat{\alpha}_s, \hat{p}_s) = \underset{\alpha_s, p_s}{\operatorname{argmax}}\, p_{\tilde{d}_1, \ldots \tilde{d}_{N_s}}(d_1, \ldots, d_{N_s}|\alpha_s, p_s) \tag{4.16}$$

L'espérance mathématique $E(\tilde{d})$ et la variance $\mathrm{var}(\tilde{d})$ de la variable aléatoire $\tilde{d}$ s'écrit :

$$E(\tilde{d}) = \frac{p_s}{\alpha_s} \tag{4.17}$$

$$\mathrm{var}(\tilde{d}) = \frac{p_s}{\alpha_s^2} \tag{4.18}$$

Nous supposons que $E(\tilde{d})$ et $\mathrm{var}(\tilde{d})$ peuvent être estimées respectivement par la moyenne et la variance empirique des échantillons. On a donc :

$$E(\tilde{d}) \approx \frac{1}{N_s} \sum_{i=1}^{N_s} d_i \tag{4.19}$$

$$\mathrm{var}(\tilde{d}) \approx \frac{1}{N_s} \sum_{i=1}^{N_s} \left(d_i - E(\tilde{d})\right)^2 \tag{4.20}$$

soit finalement :

$$\hat{\alpha}_s = \frac{E(\tilde{d})}{\mathrm{var}(\tilde{d})} \tag{4.21}$$

$$\hat{p}_s = \frac{E(\tilde{d})^2}{\mathrm{var}(\tilde{d})} \tag{4.22}$$

4.3 Estimation du modèle de trajectoires $p_{\tilde{\mathbf{X}}|\tilde{d},\tilde{s}}(\mathbf{X}|d,s)$

4.3.1 Estimation par LBG

Nous avons vu que $p_{\tilde{\mathbf{X}}|\tilde{d},\tilde{s}}(\mathbf{X}|d,s)$ se décompose en un mélange de trajectoires, où les variables aléatoires $\tilde{\mathbf{x}}_i$ associées à chaque état sont considérées comme indépendantes :

$$p_{\tilde{\mathbf{X}}|\tilde{d},\tilde{s}}(\mathbf{X}|d,s) \triangleq \sum_{t_k \in \mathbb{T}_s} Pr(\tilde{t} = t_k|\tilde{s} = s) \cdot p_{\tilde{\mathbf{X}}|\tilde{t},\tilde{d},\tilde{s}}(\mathbf{X}|t_k,d,s) \tag{4.23}$$

$$= \sum_{t_k \in \mathbb{T}_s} Pr(\tilde{t} = t_k|\tilde{s} = s) \cdot \prod_{i=0}^{Q-1} \mathcal{N}(\mathbf{x}_i; \mathbf{m}^s_{k,i}, \mathbf{\Sigma}^s_{k,i}) \tag{4.24}$$

Il est également possible de regrouper la séquence des Q vecteurs aléatoires de dimension D en un seul vecteur aléatoire de dimension $D \times Q$. L'équation (4.23) se réécrit alors sous la forme d'un mélange de lois normales multidimensionnelles :

$$p_{\tilde{\mathbf{X}}|\tilde{d},\tilde{s}}(\mathbf{X}|d,s) = \sum_{t_k \in \mathbb{T}_s} Pr(\tilde{t} = t_k|\tilde{s} = s) \cdot \mathcal{N}(\mathbf{X}; \mathbf{m}^s_k, \mathbf{\Sigma}^s_k) \tag{4.25}$$

Si M_s désigne le nombre de classes de trajectoires associées au symbole s ($M_s = \text{card}(\mathbb{T}_s)$), les paramètres à estimer sont :

- M_s densités *a priori* $Pr(\tilde{t} = t_k|\tilde{s} = s)$;
- M_s vecteurs moyennes $\mathbf{m}^s_k$;
- M_s matrices de covariance $\mathbf{\Sigma}^s_k$.

Gong résout le problème d'estimation des paramètres d'un mélange de loi normales en utilisant l'algorithme de classification LBG [Linde *et al.*, 1980], qui fournit une estimation sous optimale. Cet algorithme partitionne l'ensemble des N_s trajectoires associées au symbole s en M_s classes. La mesure de similarité utilisée pour la classification est celle de l'équation (4.9). Remarquons que cette classification est définie au niveau de la séquence de vecteurs et non au niveau d'un vecteur isolé.

Les N_s observations étant réparties dans chacune des M_s classes, l'estimation des paramètres spécifiques à chaque classe s'effectue selon le critère MLE. La moyenne $\mathbf{m}^s_k$ et la matrice de covariance $\mathbf{\Sigma}^s_k$ sont estimées par la moyenne et la covariance des observations associées à la classe t_k. La probabilité *a priori* de la classe, $Pr(\tilde{t} = t_k|\tilde{s} = s)$ est donnée par le nombre d'observations associées à la classe t_k divisé par le nombre total d'observations.

4.3.2 Estimation par EM

La difficulté de l'estimation non supervisée d'un mélange de lois de probabilité selon le critère MLE provient du fait que la catégorie de chaque observation, c.-à-d. la composante du mélange, est *a priori* inconnue. On se retrouve ainsi en présence d'un problème d'estimation à partir de données « incomplètes », qui peut être résolu par l'algorithme EM [Dempster *et al.*, 1977] (cf. annexe A).

Nous avons choisi d'appliquer cet algorithme d'estimation, plutôt que l'algorithme LBG, afin d'estimer le modèle de trajectoires $p_{\tilde{\mathbf{X}}|\tilde{d},\tilde{s}}(\mathbf{X}|d,s)$. Nos expériences ont montré que cet algorithme d'apprentissage permettait d'obtenir des taux de reconnaissance supérieurs à ceux obtenus par apprentissage LBG (cf. chapitre 9). L'algorithme EM, contrairement à LBG, procède à résolution itérative, et exploite à chaque itération l'estimation fournie à l'itération précédente. De fait, l'algorithme est sensible à l'initialisation des paramètres du modèle à estimer, et converge vers un optimum local. Nous avons choisi d'utiliser l'estimation LBG pour amorcer le calcul par EM.

Lorsque le modèle à estimer est un mélange de lois normales multidimensionnelles, les formules de réestimation du modèle sont les suivantes. Soit $\{\mathbf{m}_k^s, \mathbf{\Sigma}_k^s, Pr(\tilde{t}=t_k|\tilde{s}=s)\}$ l'ensemble des paramètres de la loi $p_{\tilde{\mathbf{X}}|\tilde{d},\tilde{s}}(\mathbf{X}|d,s)$ fourni à l'itération j de l'algorithme EM, et $\{\overline{\mathbf{m}}_k^s, \overline{\mathbf{\Sigma}}_k^s, \overline{Pr}(\tilde{t}=t_k|\tilde{s}=s)\}$, l'ensemble des paramètres de la loi $p_{\tilde{\mathbf{X}}|\tilde{d},\tilde{s}}(\mathbf{X}|d,s)$, calculé à l'itération $j+1$. On a :

$$\overline{\mathbf{m}}_k^s = \frac{\sum_{n=1}^{N_s} Pr(\tilde{t}=t_k|\tilde{\mathbf{X}}=\mathbf{X}_n,\tilde{s}=s)\cdot\mathbf{X}_n}{\sum_{n=1}^{N_s} Pr(\tilde{t}=t_k|\tilde{\mathbf{X}}=\mathbf{X}_n,\tilde{s}=s)} \tag{4.26}$$

$$\overline{\mathbf{\Sigma}}_k^s = \frac{\sum_{n=1}^{N_s} Pr(\tilde{t}=t_k|\tilde{\mathbf{X}}=\mathbf{X}_n,\tilde{s}=s)\cdot(\mathbf{X}_n-\overline{\mathbf{m}}_k^s)(\mathbf{X}_n-\overline{\mathbf{m}}_k^s)^{\#}}{\sum_{n=1}^{N_s} Pr(\tilde{t}=t_k|\tilde{\mathbf{X}}=\mathbf{X}_n,\tilde{s}=s)} \tag{4.27}$$

$$\overline{Pr}(\tilde{t}=t_k|\tilde{s}=s) = \frac{1}{N_s}\sum_{n=1}^{N_s} Pr(\tilde{t}=t_k|\tilde{\mathbf{X}}=\mathbf{X}_n,\tilde{s}=s) \tag{4.28}$$

où $Pr(\tilde{t}=t_k|\tilde{\mathbf{X}}=\mathbf{X}_n,\tilde{s}=s)$ est calculé par application de la règle de Bayes :

$$Pr(\tilde{t}=t_k|\tilde{\mathbf{X}}=\mathbf{X}_n,\tilde{s}=s) = \frac{Pr(\tilde{t}=t_k|\tilde{s}=s)\cdot\mathcal{N}(\mathbf{X}_n;\mathbf{m}_k^s,\mathbf{\Sigma}_k^s)}{\sum_{l\in\mathbb{T}_s} Pr(\tilde{t}=t_l|\tilde{s}=s)\cdot\mathcal{N}(\mathbf{X}_n;\mathbf{m}_l^s,\mathbf{\Sigma}_l^s)} \tag{4.29}$$

Le calcul itératif est répété jusqu'à convergence des estimations. Pour des raisons pratiques, nous appliquons une heuristique qui consiste à supprimer une composante du mélange, lorsque le nombre d'observations $\sum_{n=1}^{N_s} Pr(\tilde{t}=t_k|\tilde{\mathbf{X}}=\mathbf{X}_n,\tilde{s}=s)$ associées à cette classe est trop faible. Cela permet d'éviter les imprécisions provoquées par le calcul de la matrice de covariance de cette classe.

5 Récapitulatif

Nous présentons ici l'ensemble des expressions définissant le modèle stochastique de trajectoires. Plusieurs simplifications de notation vont être introduites. Désormais, nous ne préciserons plus le nom des variables aléatoires en indice des *pdfs*. $p_{\tilde{d}|\tilde{s}}(d|s)$ sera noté $p(d|s)$. Cette simplification n'introduit cependant aucune ambiguïté dans nos expressions.

De plus, nous n'effectuerons plus la distinction entre une trajectoire observée $\mathbf{Y}$ et sa version rééchantillonnée $\mathbf{X}$. Ainsi, $p_{\tilde{\mathbf{x}}_i|\tilde{t},\tilde{d},\tilde{s}}(\mathbf{y}_{n-f(i,d,Q)}|t_k,d,s)$ sera noté $p(\mathbf{x}_{n-f(i,d,Q)}|t_k,d,s)$.

$$p(\mathbf{X}_n|d,s) \stackrel{\Delta}{=} \sum_{t_k \in \mathbb{T}_s} Pr(t_k|s) \cdot p(\mathbf{X}_n|t_k,d,s) \tag{4.30}$$

$$\begin{aligned} p(\mathbf{X}_n|t_k,d,s) &\stackrel{\Delta}{=} \prod_{i=0}^{Q-1} p(\mathbf{x}_{n-\frac{Q}{2}+i\frac{d-1}{Q-1}}|t_k,d,s) \\ &= \prod_{i=0}^{Q-1} \mathcal{N}(\mathbf{x}_{n-f(i,d,Q)};\mathbf{m}_{k,i}^s,\mathbf{\Sigma}_{k,i}^s)^{\omega_i^s} \end{aligned} \tag{4.31}$$

$$\mathcal{N}(\mathbf{x};\mathbf{m},\mathbf{\Sigma}) = \frac{1}{(2\pi)^{D/2}|\mathbf{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\mathbf{m})^{\#}\mathbf{\Sigma}^{-1}(\mathbf{x}-\mathbf{m})\right] \tag{4.32}$$

Chapitre 5

Recherche de la meilleure phrase

1 Introduction

Dans le chapitre précédent, nous avons présenté comment déterminer la probabilité d'un symbole s à l'instant n, connaissant la trajectoire rééchantillonnée centrée en n, c.-à-d. en s'étant fixé une durée d pour la trajectoire observée centrée en n (cf. fig. 5.1).

Dans cette partie, nous présentons comment, à partir d'une phrase $\mathbf{X}$ constituée de N vecteurs $\mathbf{X}_0, \ldots \mathbf{X}_{N-1}$, rechercher la meilleure séquence de symboles qui lui correspond.

2 Probabilité d'une séquence de symboles

L'ensemble $\mathbb{F}$ des phrases à reconnaître, spécifique à une application donnée, peut-être représenté par une grammaire. Une phrase w, issue de cette grammaire est constituée d'une succession de $L(w)$ symboles :

$$w \triangleq a_1, \ldots a_h, \ldots a_{L(w)} \qquad \text{avec } \forall h,\ a_h \in \mathbb{P} \tag{5.1}$$

où nous rappelons que $\mathbb{P}$ désigne l'ensemble des symboles phonétiques.

Dans la séquence des N trames de parole, notons n_h la dernière trame associée au symbole a_h. La 1re trame associée à a_h est donc $n_{h-1}+1$ (cf. fig. 5.2). La durée du symbole a_h, c.-à-d. le nombre de trames associées à a_h est $d_h = n_h - n_{h-1}$, et a_h est centrée en :

$$\frac{(n_{h-1}+1)+n_h}{2} = \frac{n_h + n_{h-1} + 1}{2}$$

Lorsque les $L(w)$ frontières n_h ($1 \leq h \leq L(w)$), séparant les $L(w)$ symboles a_h de la phrase w sont connues, il est possible de calculer un score $\theta(w|n_1, \ldots n_{L(w)})$, caractérisant la probabilité de la phrase étiquetée w. Pour alléger les notations, notons :

$$\mu(s, n, d) \triangleq Pr(s|\mathbf{X}_n, d) \tag{5.2}$$

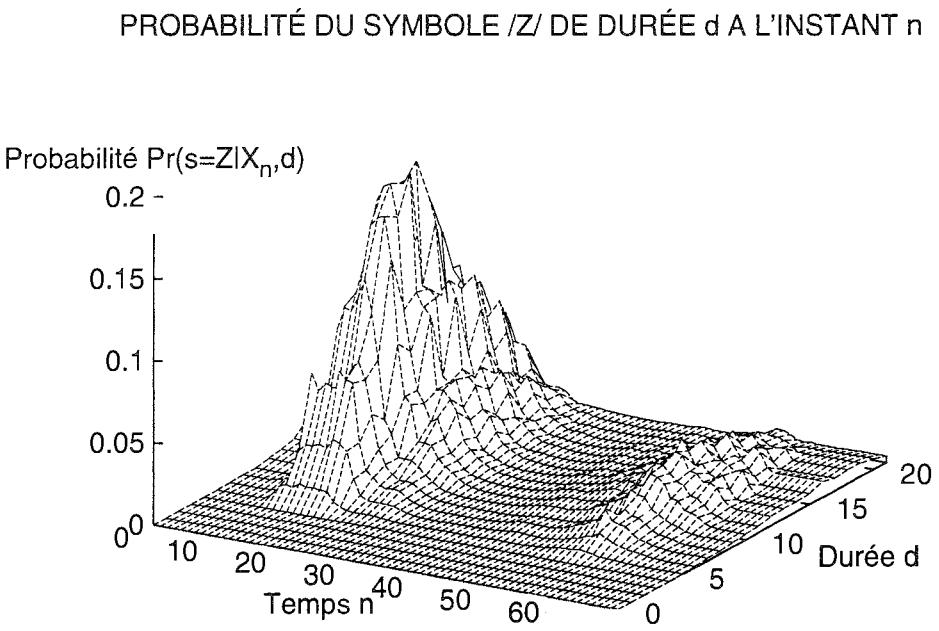

FIG. 5.1 - *Courbes de probabilité* $Pr(s = /Z/|\mathbf{X}_n, d)$ *du symbole /Z/ en fonction du temps* n *et de la durée* d *du symbole. Phrase prononcée :* «*je sors*», */Z s o R/.*

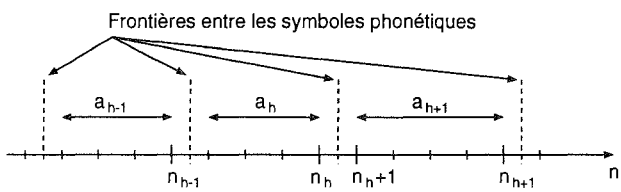

FIG. 5.2 - *Numérotation des trames de symboles consécutifs.*

Le score $\theta(w|n_1, \ldots n_{L(w)})$ est alors défini par :

$$\theta(w|n_1, \ldots n_{L(w)}) = \prod_{h=1}^{L(w)} \mu(a_h, \frac{n_{h-1} + n_h + 1}{2}, n_h - n_{h-1}) \tag{5.3}$$

où l'on pose $n_0 = 0$.

Lors de la reconnaissance, les positions des frontières entre les symboles a_h ne sont pas connues. Le calcul de la probabilité de w, notée $\Theta(w)$, consiste donc à rechercher l'ensemble des frontières $\{n_h\}$ entre symboles qui maximise la probabilité de w, les frontières étant supposées connues.

On note :

$$\begin{aligned} \Theta(w) &= \max_{n_1, \ldots n_{L(w)}} \theta(w|n_1, \ldots n_{L(w)}) \\ &= \max_{n_1, \ldots n_{L(w)}} \prod_{h=1}^{L(w)} \mu(a_h, \frac{n_{h-1} + n_h + 1}{2}, n_h - n_{h-1}) \end{aligned} \tag{5.4}$$

3 Algorithme de calcul de $\Theta(w)$

Notons $\Xi(j)$ la probabilité de la séquence constituée des j premiers symboles de la phrase w, de la trame 0 à la trame $N-1$ du signal, les frontières entre symboles étant fixées. D'après l'équation (5.3), on a :

$$\begin{aligned} \Xi(j) &= \prod_{h=1}^{j} \mu(a_h, \frac{n_{h-1} + n_h + 1}{2}, n_h - n_{h-1}) \\ &= \prod_{h=1}^{j-1} \mu(a_h, \frac{n_{h-1} + n_h + 1}{2}, n_h - n_{h-1}) \times \mu(a_j, \frac{n_{j-1} + n_j + 1}{2}, n_j - n_{j-1}) \\ &= \Xi(j-1) \times \mu(a_j, \frac{n_{j-1} + n_j + 1}{2}, n_j - n_{j-1}) \end{aligned} \tag{5.5}$$

On a donc :

$$\Xi(L(w)) = \theta(w|n_1, \ldots n_{L(w)}) \tag{5.6}$$

Notre objectif est de maximiser $\Xi(L(w))$ en faisant varier les positions des frontières entre les symboles :

$$\Theta(w) = \max_{n_1, \ldots n_{L(w)}} \Xi(L(w)) \tag{5.7}$$

Notons $\Pi(l, j)$ la probabilité de la séquence optimale constituée des j premiers symboles de w, entre la trame 0 et la trame l du signal. Posons $l = n_j$ et $k = n_{j-1}$. Nous recherchons comment exprimer $\Pi(l, j)$ connaissant la probabilité de la séquence optimale des $j-1$ symboles précédents, de la trame 0 à la trame k, $\Pi(k, j-1)$.

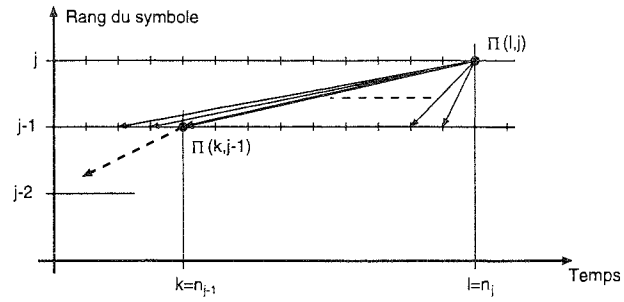

FIG. 5.3 - *Passage du calcul de la probabilité cumulée* $\Pi(k, j-1)$ *à* $\Pi(l, j)$*, entre le symbole de rang* $j-1$ *et le symbole de rang* j.

La probabilité du symbole a_j, entre les instants $k+1$ et l est :

$$\mu(a_j, \frac{k+l+1}{2}, l-k)$$

On a donc (cf. 5.3) :

$$\begin{aligned}\Pi(l,j) &= \max_{n_1,\ldots k,l} \Xi(j) \\ &= \max_{n_1,\ldots k,l} \left\{ \Xi(j-1) \times \mu(a_j, \frac{k+l+1}{2}, l-k) \right\} \\ &= \max_{0 \le k < l} \left\{ \Pi(k, j-1) \times \mu(a_j, \frac{k+l+1}{2}, l-k) \right\}\end{aligned} \tag{5.8}$$

avec $0 \le l < N$ et $1 \le j \le L(w)$.

Finalement, $\Theta(w)$ s'obtient par :

$$\Theta(w) = \Pi(N-1, L(w)) \tag{5.9}$$

calculé à partir de la relation de récurrence (5.8).

En pratique, la recherche de k n'est pas effectuée dans tout l'intervalle $[0, N[$ comme l'indique la relation (5.8). En effet, il est possible de trouver un intervalle de durée $[d_j^m, d_j^M]$ spécifique au symbole a_j, tel qu'il soit très peu probable d'observer un symbole a_j ayant une durée située à l'extérieur de cet intervalle. La recherche de k ne s'effectue alors que dans l'intervalle $[l - d_j^M, l - d_j^m]$, comme l'illustre la figure 5.4.

4 Recherche de la meilleure phrase

Nous avons vu dans le paragraphe précédent, comment déterminer la probabilité d'une phrase donnée w, à partir des probabilités des symboles $\mu(s, n, d)$ fournies par les modèles de trajectoires. La reconnaissance d'une phrase s'effectue finalement de la façon suivante.

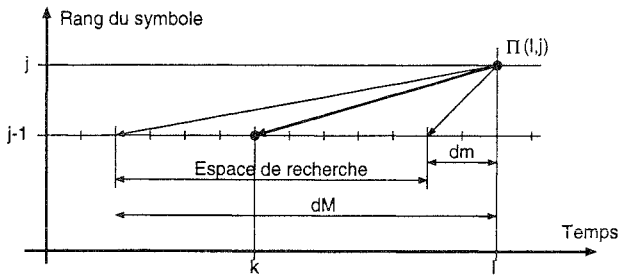

FIG. 5.4 - *Limitation de l'espace de recherche dans le calcul des probabilités cumulées, par des contraintes sur la durée du symbole courant a_j, limitée à l'intervalle $[d_j^m, d_j^M]$.*

À l'aide d'une grammaire, on génère l'ensemble des phrases possibles. Étant donné une séquence de symboles, on calcule la probabilité de cette phrase en utilisant l'équation (5.8). Le système VINICS propose finalement les N phrases dont les probabilités sont les plus élevées.

$$\hat{w} = \underset{w \in \mathbb{F}}{\operatorname{argmax}}\, \Theta(w) \tag{5.10}$$

Lors du calcul des probabilités cumulées d'une séquence de symboles, différentes heuristiques sont introduites pour élaguer l'arbre de recherche. Ces heuristiques ne seront pas présentées ici.

Il faut remarquer que la probabilité d'une phrase ainsi définie pénalise les phrases longues au détriment des phrases courtes. En effet, plus la phrase comporte de symboles, plus la probabilité de la phrase est faible, car plus les produits de probabilités $\mu(a_h, n, d)$ sont nombreux. Par conséquent, il est nécessaire de normaliser le score fourni par l'équation (5.8), pour prendre en compte soit le nombre de symboles dans la phrase, soit la durée totale de la phrase.

Gong propose deux normalisations possibles. La 1re consiste à calculer la moyenne géométrique des probabilités de symboles. La phrase reconnue $\hat{w}$ est donc celle qui maximise sa moyenne géométrique :

$$\hat{w} = \underset{w \in \mathbb{F}}{\operatorname{argmax}}\, \Theta(w)^{\frac{1}{L(w)}} \tag{5.11}$$

La 2e normalisation est similaire à celle introduite par [Ostendorf et Roukos, 1989] et consiste à pondérer chaque $\mu(a_h, n, d)$ par sa durée. L'équation (5.8) devient :

$$\Pi(l, j) = \max_{0 \leq k < l} \left\{ \Pi(k, j-1) \times \mu(a_j, \frac{k+l+1}{2}, l-k)^{l-k} \right\} \tag{5.12}$$

Ceci revient à normaliser la probabilité en fonction du nombre de trames de la phrase. La phrase reconnue est ensuite fournie par l'équation (5.10).

Conclusion

Dans cette partie, nous avons présenté le modèle stochastique de trajectoires et son utilisation en reconnaissance de parole continue. Une trajectoire observée, spécifique à un symbole s, est considérée comme une réalisation d'une trajectoire aléatoire sous-jacente qui constitue le modèle de s. La fonction de densité de probabilité de cette trajectoire aléatoire est représentée par un mélange de trajectoires, ce qui permet de modéliser une distribution complexe des trajectoires dans l'espace des trajectoires, et de faire apparaître des classes de trajectoires. On évite ainsi le phénomène de recouvrement de trajectoires présenté en introduction. Dans chaque classe, une trajectoire aléatoire est définie par une séquence de Q vecteurs aléatoires. À chacun de ces vecteurs est associé une loi normale multidimensionnelle. L'estimation des paramètres de chaque loi normale, ainsi que des probabilités *a priori* de chaque classe est effectuée selon le critère du maximum de la vraisemblance, en utilisant l'algorithme LBG [Linde *et al.*, 1980] ou l'algorithme EM [Dempster *et al.*, 1977]. Bien évidemment, d'autres critères d'estimation pourraient être envisagés. L'optimisation des paramètres du modèle est effectuée au niveau de la séquence d'état (c.-à-d. de la trajectoire), et non au niveau d'un état (comme c'est le cas pour les HMMs).

Les STMs utilisent une modélisation explicite de la durée des symboles, sous la forme d'une loi Γ. Une pondération de l'influence de la durée peut-être mise en oeuvre (paramètre λ), et améliore de façon significative les taux de reconnaissance [Gong, 1994].

Une pondération peut-être introduite pour contrôler l'importance de chaque état d'une trajectoire. Il est ainsi possible de renforcer l'influence des états ayant une faible variance (états centraux) par rapport aux états ayant une variance plus importante (états latéraux). Cette pondération permet d'améliorer les taux de reconnaissance [Gong, 1994].

Le CRIN/INRIA dispose du système de reconnaissance de parole continue VINICS, fondé sur les modèles stochastiques de trajectoires. Toutes nos expériences ont été effectuées en utilisant ce système, conçu pour des applications de reconnaissance de parole continue utilisant un vocabulaire de 100-3000 mots. Le système de reconnaissance est complet, et fournit, à partir d'une phrase prononcée, sa transcription orthographique. Il est possible de construire automatiquement, à partir d'un texte échantillon, les contraintes syntaxiques (la grammaire) utilisées lors de la reconnaissance.

Les dernières évaluations de VINICS, réalisées par Y. Gong, ont porté sur une tâche de reconnaissance de parole continue sur un vocabulaire de 2000 mots, avec une perplexité du langage en mode bigramme d'environ 48. Un ensemble de 33 modèles de phones indépendants du contexte sont construits pour chaque locuteur à partir de 80 phrases (env. 6 min) de parole propre, indépendantes de l'application. Moyenné sur 4 locuteurs de test, le taux de

reconnaissance de mots, obtenu en mode dépendant du locuteur, est d'environ 99% [Gong, 1994].

Troisième partie

Trois approches pour la reconnaissance de la parole bruitée

Introduction

Dans cette partie, nous présentons trois approches pour la reconnaissance de la parole continue en milieu bruité (bruit additif), à partir d'un système de reconnaissance initialement entraîné avec de la parole propre.

La première approche proposée, appelée combinaison de modèles stochastiques (cf. chap. 6), s'inscrit dans le cadre des méthodes de transformation des modèles de parole propre. La combinaison de modèles utilise une connaissance explicite de la perturbation, où l'on considère que le bruit est additif, indépendant du signal de parole, et est modélisé dans le domaine cepstral par un HMM utilisant des *pdfs* Gaussiennes. Cette méthode permet de construire des STMs de parole bruitée sans nécessiter un entraînement des modèles à partir d'un corpus d'apprentissage de parole bruitée. Les modèles ainsi obtenus sont alors directement utilisés pour la reconnaissance de la parole dans le bruit.

La seconde méthode, désignée par filtrage non linéaire par états (cf. chap. 7), consiste à déterminer un ensemble d'estimateurs de parole propre, un estimateur étant spécifique à chaque état de chaque STM de parole propre. Le critère d'estimation minimise une erreur quadratique moyenne dans le domaine du logarithme du spectre, domaine relié à celui de la perception auditive humaine. Le calcul des estimateurs suppose, comme précédemment, que parole et bruit sont additifs. Cette approche se situe entre les méthodes de prétraitement du signal, car un ensemble de filtres est appliqué au signal de parole dans le domaine cepstral, et les méthodes de transformation des systèmes de reconnaissance, car les filtres sont associés spécifiquement à chaque état de chaque modèle, et la structure du système de reconnaissance est donc modifiée. La reconnaissance de la parole dans le bruit s'effectue finalement par filtrage de la parole bruitée, pendant le processus de reconnaissance.

La dernière approche est appelée adaptation au bruit par régression linéaire (cf. chap. 8). Cette méthode consiste à déterminer un ensemble de transformations linéaires, pouvant être utilisées soit pour transformer les modèles de parole propre, dans le but de construire des modèles permettant de reconnaître la parole bruitée, soit de transformer la parole bruitée à reconnaître, afin de pouvoir utiliser les modèles initiaux de parole propre. Selon la mise en œuvre choisie, cette approche s'inscrit donc soit dans les méthodes de transformation des modèles, soit dans les méthodes de transformation de la parole. Dans les deux cas, les transformations utilisées sont déterminées à partir d'un corpus d'adaptation étiqueté de parole bruitée, selon un critère qui consiste à maximiser la vraisemblance de ce corpus d'adaptation. Cette approche n'utilise pas d'information sur l'origine de la perturbation, mais suppose que cette dernière peut être modélisée par des transformations linéaires par partie de l'espace cepstral.

Ces trois méthodes sont évaluées sur une même application de reconnaissance de parole

continue dans le bruit, pour un vocabulaire d'un millier de mots, en mode dépendant du locuteur, en présence de différents types et niveaux de bruits. Une quatrième approche, appelée transformation de base [Gong, 1993] (cf. annexe B) et qui consiste à débruiter le signal par transformation d'espace dans le domaine cepstral, est également évaluée. Les résultats obtenus sont comparés, et des conclusions sont tirées sur les avantages et inconvénients des différentes approches (cf. chap. 9).

Chapitre 6

Combinaison de modèles stochastiques

1 Introduction

Les conditions d'utilisation d'un système de reconnaissance des formes sont optimales lorsqu'elles correspondent à celles présentes lors de l'apprentissage. Le cas idéal de la reconnaissance de parole en environnement bruité s'obtient donc lorsque le système est entraîné dans les mêmes conditions d'environnement.

Effectuer l'apprentissage d'un système de reconnaissance automatique de parole à partir d'un corpus de parole bruitée est rarement envisageable. En effet, il est en pratique difficile d'obtenir un corpus d'apprentissage de parole bruitée présentant les mêmes conditions de bruit que celles qui seront présentes lors du test. De plus, dans le cas d'un apprentissage en parole bruitée, comment prendre en compte lors du test les variations du rapport signal-à-bruit, ou encore la présence d'un bruit inconnu de l'apprentissage, ou enfin l'absence de bruit lors du test.

Supposons maintenant que la nature des variations entre conditions de test et d'apprentissage soit connue. On peut par exemple supposer que les conditions de test correspondent à la présence d'un bruit additif stationnaire de rapport signal-à-bruit connu par rapport aux conditions d'apprentissage. Il est alors possible de perturber les données d'apprentissage et d'utiliser ces données perturbées pour réentraîner les modèles. À nouveau, une telle approche est rarement utilisable car elle nécessite de stocker en permanence la totalité du corpus d'apprentissage. D'autre part, cette approche est très coûteuse en calcul car elle nécessite tout d'abord de perturber l'ensemble du corpus d'apprentissage, et ensuite de réentraîner les modèles de parole à chaque fois que les statistiques du bruit évoluent.

Considérons un problème statistique dans lequel un grand nombre de données est disponible. Le traitement de ces données est parfois simplifié s'il est possible de calculer quelques valeurs numériques, ou statistiques, «résumant» l'information pertinente contenue dans les données. Des analyses valides peuvent alors êtres conduites en utilisant ces valeurs plutôt que l'ensemble des données. De telles valeurs sont appelées statistiques suffisantes [DeGroot, 1970].

Nous nous intéressons ici à une telle notion car nous souhaitons non pas perturber un ensemble de données de parole propre (non disponibles) pour réestimer un modèle, mais plutôt

utiliser un résumé significatif de la parole propre afin de le perturber et de réestimer un modèle de parole bruitée.

Soit un vecteur aléatoire $\mathbf{Y}$, caractérisant une zone stationnaire d'un signal de parole bruitée dans un espace de paramètres. Supposons que $\mathbf{Y}$ soit une fonction de deux vecteurs aléatoires $\mathbf{X}$ et $\mathbf{N}$ représentant respectivement une zone stationnaire d'un signal de parole propre et d'un bruit. On note $\mathbf{Y} = \psi(\mathbf{X}, \mathbf{N})$, et $\psi(\cdot)$ caractérise la nature de la combinaison entre parole propre et bruit (bruit additif, convolutif, etc.).

Supposons que l'on dispose de M observations du vecteur de parole bruitée $\mathbf{Y}$, notées $\mathbf{Y}_1$ à $\mathbf{Y}_M$. Si la *pdf* de $\mathbf{Y}$ est définie par une loi normale multidimensionnelle $p(\mathbf{Y})$ de moyenne $\hat{\mathbf{m}}$ et de matrice de covariances $\hat{\boldsymbol{\Sigma}}$, une estimation de $\hat{\mathbf{m}}$ et $\hat{\boldsymbol{\Sigma}}$ selon le critère du maximum de la vraisemblance est obtenue par les expressions suivantes, où $(\cdot)^{\#}$ représente l'opérateur de transposition :

$$\hat{\mathbf{m}} = \frac{1}{M} \sum_{i=1}^{M} \mathbf{Y}_i \tag{6.1}$$

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{M} \sum_{i=1}^{M} (\mathbf{Y}_i - \hat{\mathbf{m}}^{cep}) (\mathbf{Y}_i - \hat{\mathbf{m}}^{cep})^{\#} \tag{6.2}$$

Si $M \longrightarrow \infty$, il est possible de remplacer la moyenne et la variance empirique par l'espérance et la matrice de covariances du vecteur aléatoire $\mathbf{Y}^{cep}$:

$$\hat{\mathbf{m}} = E(\mathbf{Y}) \tag{6.3}$$

$$\hat{\boldsymbol{\Sigma}} = \mathrm{var}(\mathbf{Y}) \tag{6.4}$$

Développons (6.3) en posant $\mathbf{Y} = \psi(\mathbf{X}, \mathbf{N})$ et en supposant que $\mathbf{X}$ et $\mathbf{N}$ sont indépendant :

$$\begin{aligned} \hat{\mathbf{m}} &= E(\psi(\mathbf{X}, \mathbf{N})) \\ &= \iint \psi(\mathbf{X}, \mathbf{N}) p(\mathbf{X}, \mathbf{N}) d\mathbf{X} d\mathbf{N} \\ &= \iint \psi(\mathbf{X}, \mathbf{N}) p(\mathbf{X}) p(\mathbf{N}) d\mathbf{X} d\mathbf{N} \end{aligned} \tag{6.5}$$

Un développement analogue peut être écrit pour $\hat{\boldsymbol{\Sigma}}$. Il apparaît que pour entraîner le modèle de $\mathbf{Y}$, il n'est donc pas nécessaire de disposer d'un ensemble d'observations de parole bruitée. En effet, $\hat{\mathbf{m}}$ et $\hat{\boldsymbol{\Sigma}}$ peuvent être estimées dès lors que l'on connaît les *pdfs* de $\mathbf{X}$ et de $\mathbf{N}$, ainsi que la fonction $\psi(\cdot)$ liant $\mathbf{Y}$ à $\mathbf{X}$ et $\mathbf{N}$.

En pratique, la *pdf* de la parole propre et du bruit ne sont pas connues. Cependant, il est raisonnable de considérer que le modèle de parole propre contient suffisamment d'informations pertinentes pour caractériser la distribution de $\mathbf{X}$. De même, il est en général aisé d'entraîner un modèle de bruit pour représenter $\mathbf{N}$, la collecte d'observations de bruit ne présentant pas de difficultés pratiques. Si l'on considère que $p(\mathbf{X})$ et $p(\mathbf{N})$ sont représentées par des lois normales multidimensionnelles, respectivement de moyennes $\mathbf{m}$, $\tilde{\mathbf{m}}$ et de covariances $\boldsymbol{\Sigma}$, $\tilde{\boldsymbol{\Sigma}}$, et étant donné la fonction $\psi(\cdot)$, il est souvent impossible de calculer les intégrales de l'expression (6.5) sous forme close. Comme nous allons le voir, c'est en particulier le cas

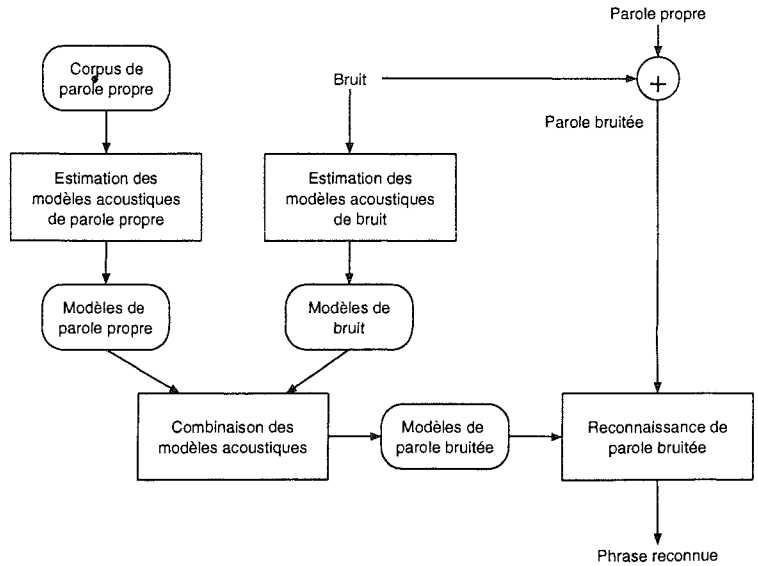

FIG. 6.1 - *Structure d'un système de reconnaissance de parole avec combinaison de modèles.*

lorsque $\mathbf{Y}$, $\mathbf{X}$ et $\mathbf{N}$ sont représentés dans le domaine cepstral et lorsqu'on suppose que parole et bruit sont additifs. Pour exprimer $\hat{\mathbf{m}}$ et $\hat{\mathbf{\Sigma}}$ en fonction de $\mathbf{m}$, $\tilde{\mathbf{m}}$, $\mathbf{\Sigma}$ et $\tilde{\mathbf{\Sigma}}$, il est alors nécessaire d'introduire des hypothèses simplificatrices. Par conséquent, les estimations que nous allons obtenir ne seront pas les estimations MLE du modèle de parole bruitée, mais une approximation des estimateurs MLE.

La structure du système de reconnaissance de parole bruitée que nous cherchons à construire est celle de la fig. 6.1. La première étape consiste à entraîner un système avec de la parole propre. Lors du test, un modèle de bruit est estimé pendant les pauses du locuteur. Ce modèle est combiné avec les modèles de parole propre pour générer un modèle de parole bruitée. La parole bruitée du locuteur est reconnue avec le modèle ainsi construit.

2 Caractérisation de la *pdf* d'un vecteur de parole bruitée

Dans cette partie, nous cherchons à déterminer l'expression de la fonction de densité de probabilité d'un vecteur aléatoire $\mathbf{Y}^{cep}$ représentant le cepstre d'une observation de parole bruitée, connaissant les *pdfs* de la parole propre et du bruit, ainsi que la relation liant parole bruitée, parole propre et bruit.

Soit $x(t)$ le signal temporel de parole propre et $n(t)$ un bruit stationnaire. Sous l'hypothèse d'additivité dans le domaine temporel, le signal $y(t)$, correspondant au signal $x(t)$ perturbé par le bruit $n(t)$ s'écrit, en supposant que l'addition du bruit ne modifie pas la localisation des trames de parole :

$$y(t) = x(t) + n(t) \tag{6.6}$$

Dans le domaine des densités spectrales de puissance, l'équation (6.6) s'écrit, sous l'hypothèse d'indépendance entre le signal de parole propre et le bruit :

$$Y(\omega) = X(\omega) + N(\omega) \tag{6.7}$$

où $Y(\omega)$, $X(\omega)$ et $N(\omega)$ représentent respectivement les densités spectrales de puissance (*PSD*[1]) de $y(t)$, $x(t)$ et $n(t)$.

Soient $\mathbf{Y}^{cep}$, $\mathbf{X}^{cep}$ et $\mathbf{N}^{cep}$ les vecteurs de cepstres de $Y(\omega)$, $X(\omega)$, $N(\omega)$, définis par les expressions suivantes, où IDCT[2] représente la transformée en Cosinus discrète inverse :

$$\mathbf{Y}^{cep} \triangleq \text{IDCT}\{\ln Y(\omega)\}$$
$$\mathbf{X}^{cep} \triangleq \text{IDCT}\{\ln X(\omega)\}$$
$$\mathbf{N}^{cep} \triangleq \text{IDCT}\{\ln N(\omega)\}$$

Nous supposons que $\mathbf{X}^{cep}$ et $\mathbf{N}^{cep}$ peuvent être modélisés avec suffisamment de précision en utilisant des lois normales, caractérisées respectivement par leurs moyennes $\mathbf{m}^{cep}$ et $\tilde{\mathbf{m}}^{cep}$, et leurs matrices de covariances $\mathbf{\Sigma}^{cep}$ et $\tilde{\mathbf{\Sigma}}^{cep}$. Sous cette hypothèse, et étant donné la relation (6.6), nous souhaitons déterminer l'expression de la *pdf* du modèle de parole bruitée.

Soit $\mathbf{X}^{log}$ le vecteur aléatoire correspondant à la transformation de Fourier discrète du vecteurs $\mathbf{X}^{cep}$. Ce vecteur aléatoire caractérise la *PSD* dans le domaine du logarithme du spectre, appelé par la suite domaine log-spectral. Si $\mathbf{W}$ représente la matrice de transformation en cosinus discrète, on a :

$$\mathbf{X}^{log} = \mathbf{W}\mathbf{X}^{cep} \tag{6.8}$$

La transformation en cosinus discrète étant une opération linéaire, la *pdf* de $\mathbf{X}^{log}$ reste une loi normale [Papoulis, 1991], de moyenne $\mathbf{m}^{log}$ et de matrice de covariances $\mathbf{\Sigma}^{log}$, avec :

$$\mathbf{m}^{log} = \mathbf{W}\mathbf{m}^{cep} \tag{6.9}$$
$$\mathbf{\Sigma}^{log} = \mathbf{W}\mathbf{\Sigma}^{cep}\mathbf{W}^{\#} \tag{6.10}$$

Soit $\mathbf{X}^{lin}$ le vecteur aléatoire défini par :

$$\mathbf{X}^{lin} = \exp(\mathbf{X}^{log}) \tag{6.11}$$

Le vecteur aléatoire $\mathbf{X}^{lin}$ caractérise la *PSD* dans le domaine linéaire. La *pdf* de $\mathbf{X}^{lin}$ correspond alors à une loi log-normale[3], de moyenne $\mathbf{m}^{lin}$ et de covariances $\mathbf{\Sigma}^{lin}$, avec [Gales et Young, 1993a] :

$$m_i^{lin} = \exp(m_i^{log} + \Sigma_{i,i}^{log}/2) \tag{6.12}$$
$$\Sigma_{i,j}^{lin} = m_i^{lin} m_j^{lin}[\exp(\Sigma_{i,j}^{log}) - 1] \tag{6.13}$$

1. Power Spectral Densities
2. Inverse Discrete Cosine Transform
3. Une variable aléatoire x suit une loi appelée log-normale si la variable aléatoire $\log(x)$ suit une loi normale

Des expressions équivalentes sont obtenues pour le vecteur de bruit $\mathbf{N}^{lin}$ dans le domaine des densités spectrales de puissance.

A ce stade, nous connaissons donc l'expression des *pdfs* de $\mathbf{X}^{lin}$ et $\mathbf{N}^{lin}$, qui correspondent toutes deux à des lois log-normales, dont les paramètres s'expriment en fonction des paramètres des *pdfs* de $\mathbf{X}^{cep}$ et $\mathbf{N}^{cep}$. Le vecteur aléatoire $\mathbf{Y}^{lin}$ correspond à la somme des vecteurs aléatoires $\mathbf{X}^{lin}$ et $\mathbf{N}^{lin}$ (cf. Eq. (6.7)). La *pdf* de $\mathbf{Y}^{lin}$ est alors définie par le produit de convolution des *pdfs* de $\mathbf{X}^{lin}$ et $\mathbf{N}^{lin}$. Malheureusement, le calcul du produit de convolution entre deux *pdfs* log-normales ne peut s'effectuer sous forme close. Pour simplifier l'expression de la *pdf* de $\mathbf{Y}^{lin}$, nous supposerons que la *pdf* décrivant la somme de deux vecteurs aléatoires distribués selon une loi log-normale est aussi log-normale. Sous cette condition, l'expression de la *pdf* de $\mathbf{Y}^{lin}$ est entièrement déterminée par la connaissance de sa moyenne $\hat{\mathbf{m}}^{lin}$ et de sa matrice de covariance $\hat{\mathbf{\Sigma}}^{lin}$ dans le domaine linéaire des densités spectrales de puissance. On a :

$$E(\mathbf{Y}^{lin}) = E(\mathbf{X}^{lin}) + E(\mathbf{N}^{lin}) \tag{6.14}$$

$$\text{var}(\mathbf{Y}^{lin}) = \text{var}(\mathbf{X}^{lin}) + \text{var}(\mathbf{N}^{lin}) \qquad \text{car indépendance} \tag{6.15}$$

En pratique, nous introduisons un coefficient de correction g, permettant de prendre en compte les variations moyennes d'énergie entre la parole propre d'entraînement et la parole de test prononcée dans le bruit. $\hat{\mathbf{m}}^{lin}$ et $\hat{\mathbf{\Sigma}}^{lin}$ s'expriment donc par :

$$\hat{\mathbf{m}}^{lin} = g\mathbf{m}^{lin} + \tilde{\mathbf{m}}^{lin} \tag{6.16}$$

$$\hat{\mathbf{\Sigma}}^{lin} = g^2\mathbf{\Sigma}^{lin} + \tilde{\mathbf{\Sigma}}^{lin} \tag{6.17}$$

L'expression de la *pdf* de $\mathbf{Y}^{lin}$ est donc entièrement définie dans le domaine linéaire des *PSD*. La *pdf* de $\mathbf{Y}^{cep}$ dans le domaine cepstral s'obtient en appliquant successivement les transformations inverses de celles utilisées précédemment, à savoir un opérateur logarithmique pour revenir au domaine log-spectral suivi d'une transformation en cosinus inverse.

Ainsi, dans le domaine logarithmique des *PSD*, on a, en inversant les équations (6.12) et (6.13) [Gales et Young, 1993a] :

$$\hat{m}_i^{log} = \log(\hat{m}_i^{lin}) - \frac{1}{2}\log\left(\frac{\hat{\Sigma}_{i,i}^{lin}}{\hat{m}_i^{lin}\hat{m}_i^{lin}} + 1\right) \tag{6.18}$$

$$\hat{\Sigma}_{i,j}^{log} = \log\left(\frac{\hat{\Sigma}_{i,j}^{lin}}{\hat{m}_i^{lin}\hat{m}_j^{lin}} + 1\right) \tag{6.19}$$

Par inversion des équations (6.9) et (6.10), on obtient en définitive, dans le domaine cepstral :

$$\hat{\mathbf{m}}^{cep} = \mathbf{W}^{-1}\hat{\mathbf{m}}^{log} \tag{6.20}$$

$$\hat{\mathbf{\Sigma}}^{cep} = \mathbf{W}^{-1}\hat{\mathbf{\Sigma}}^{log}(\mathbf{W}^{-1})^{\#} \tag{6.21}$$

Par la suite et pour alléger les notations, nous appelons ψ_m et ψ_Σ les fonctions permettant de déterminer l'expression des moyennes et variances des modèles de parole bruitée, à partir

des paramètres des modèles de parole propre et du bruit. On pose :

$$\hat{\mathbf{m}}^{cep} = \psi_m(\mathbf{m}^{cep}, \tilde{\mathbf{m}}^{cep}, \mathbf{\Sigma}^{cep}, \tilde{\mathbf{\Sigma}}^{cep}) \tag{6.22}$$

$$\hat{\mathbf{\Sigma}}^{cep} = \psi_\Sigma(\mathbf{m}^{cep}, \tilde{\mathbf{m}}^{cep}, \mathbf{\Sigma}^{cep}, \tilde{\mathbf{\Sigma}}^{cep}) \tag{6.23}$$

$\psi_m(\mathbf{m}^{cep}, \tilde{\mathbf{m}}^{cep}, \mathbf{\Sigma}^{cep}, \tilde{\mathbf{\Sigma}}^{cep})$ est obtenu en appliquant successivement les équations (6.9), (6.12), (6.16), (6.18) et (6.20). De la même façon, $\psi_\Sigma(\mathbf{m}^{cep}, \tilde{\mathbf{m}}^{cep}, \mathbf{\Sigma}^{cep}, \tilde{\mathbf{\Sigma}}^{cep})$ est obtenu en appliquant successivement les équations (6.10), (6.13), (6.17), (6.19) et (6.21).

En définitive, la *pdf* de la parole bruitée $\mathbf{Y}^{cep}$ s'exprime sous la forme d'une loi normale dont les paramètres sont fonctions des paramètres des *pdfs* de la parole propre $\mathbf{X}^{cep}$ et du bruit $\mathbf{N}^{cep}$:

$$p(\mathbf{Y}^{cep}) \propto \mathcal{N}(\mathbf{Y}^{cep}; \psi_m(\mathbf{m}^{cep}, \tilde{\mathbf{m}}^{cep}, \mathbf{\Sigma}^{cep}, \tilde{\mathbf{\Sigma}}^{cep}), \psi_\Sigma(\mathbf{m}^{cep}, \tilde{\mathbf{m}}^{cep}, \mathbf{\Sigma}^{cep}, \tilde{\mathbf{\Sigma}}^{cep})) \tag{6.24}$$

3 Application à la combinaison d'un STM et d'un HMM

3.1 Modèle de bruit à un état

Plaçons nous tout d'abord dans le cas le plus simple où le bruit, supposé stationnaire, est modélisé dans le domaine cepstral par une seule loi normale multidimensionnelle, notée $\mathcal{N}(\mathbf{N}; \tilde{\mathbf{m}}, \tilde{\mathbf{\Sigma}})$. Soit $\mathbf{X}_n$ une trajectoire de parole propre centrée en n. Supposons que cette trajectoire est modélisée par un STM. D'après la définition des STMs donnée partie II, la distribution de $\mathbf{X}_n$, connaissant la classe de trajectoire t_k, le symbole s et la durée d est :

$$p(\mathbf{X}_n|t_k, d, s) = \prod_{i=0}^{Q-1} \mathcal{N}(\mathbf{x}_{n-f(i,d,Q)}; \mathbf{m}^s_{k,i}, \mathbf{\Sigma}^s_{k,i}) \tag{6.25}$$

Le vecteur de parole propre associé à l'état i, caractérisé par une loi normale de moyenne $\mathbf{m}^s_{k,i}$ et de matrice de covariances $\mathbf{\Sigma}^s_{k,i}$, va être perturbé par une observation de bruit. La trajectoire propre $\mathbf{X}_n$ va donc être transformée en une trajectoire bruitée $\mathbf{Y}_n$.

Étant donné les règles de combinaison entre la *pdf* de parole propre et la *pdf* de bruit présentées précédemment, le vecteur de parole bruitée associé à l'état i peut être caractérisé par une loi normale de moyenne $\hat{\mathbf{m}}^s_{k,i}$ et de matrice de covariances $\hat{\mathbf{\Sigma}}^s_{k,i}$, avec $\hat{\mathbf{m}}^s_{k,i} = \psi_m(\mathbf{m}^s_{k,i}, \tilde{\mathbf{m}}, \mathbf{\Sigma}^s_{k,i}, \tilde{\mathbf{\Sigma}})$ et $\hat{\mathbf{\Sigma}}^s_{k,i} = \psi_\Sigma(\mathbf{m}^s_{k,i}, \tilde{\mathbf{m}}, \mathbf{\Sigma}^s_{k,i}, \tilde{\mathbf{\Sigma}})$. On obtient donc :

$$p(\mathbf{Y}_n|t_k, d, s) = \prod_{i=0}^{Q-1} \mathcal{N}(\mathbf{y}_{n-f(i,d,Q)}; \hat{\mathbf{m}}^s_{k,i}, \hat{\mathbf{\Sigma}}^s_{k,i}) \tag{6.26}$$

La reconnaissance de la parole bruitée est donc possible en remplaçant pour chaque symbole s, chaque classe de trajectoire t_k et chaque état i, la moyenne $\mathbf{m}^s_{k,i}$ et la matrice de covariances $\mathbf{\Sigma}^s_{k,i}$ par $\hat{\mathbf{m}}^s_{k,i}$ et $\hat{\mathbf{\Sigma}}^s_{k,i}$. On obtient ainsi directement une approximation d'un STM de parole bruitée (cf. fig. 6.2).

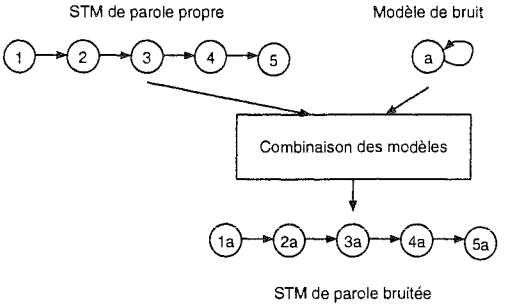

FIG. 6.2 - *Combinaison d'un STM de parole propre avec une loi normale de bruit. Le modèle de parole bruitée reste un STM.*

3.2 HMM de bruit

En pratique, le bruit est rarement stationnaire. Il n'est donc pas possible de le représenter par une seule loi normale et l'équation (6.26) n'est pas directement applicable. Un modèle de bruit couramment utilisé et permettant de prendre en compte les non-stationnarités consiste à représenter la distribution du bruit par un HMM ergodique à N états. Ce HMM est entièrement défini par :

- l'ensemble $I = \{1, \ldots N\}$ représentant les N états du modèles; l'état occupé à l'instant t est noté q_t, avec $q_t \in I$,
- l'ensemble $A = \{a_{ij}\}_{i,j=1}^{N}$ représentant les probabilités de transition entre états ; on note :

$$a_{ij} = Pr(q_{t+1} = j | q_t = i) \qquad 1 \leq i, j \leq N$$

- l'ensemble $B = \{b_i(\mathbf{x})\}_{i=1}^{N}$ des *pdfs* associées à chaque état ; on note :

$$b_i(\mathbf{x}) = \mathcal{N}(\mathbf{x}; \tilde{\mathbf{m}}_i, \tilde{\mathbf{\Sigma}}_i) \qquad 1 \leq i \leq N$$

- l'ensemble $\Pi = \{\pi_i\}_{i=1}^{N}$ des probabilités initiales d'occupation des états ; on note :

$$\pi_i = Pr(q_0 = i) \qquad 1 \leq i \leq N$$

Ce modèle peut être estimé selon le critère MLE en utilisant l'algorithme de Baum et Welch [Rabiner, 1989].

Avec un tel modèle de bruit, il n'est plus possible de savoir *a priori* de quel état du HMM provient l'observation de bruit se combinant avec une observation de parole associée à l'état i du STM. Contrairement au cas précédent où la combinaison entre le modèle de bruit et le STM était unique et conduisait à l'obtention directe d'un STM de parole bruitée, nous obtenons désormais une structure hybride de STM, où chaque état du STM initial peut être combiné avec n'importe quel état du HMM de bruit (cf. fig. 6.3). On modélise ainsi toujours

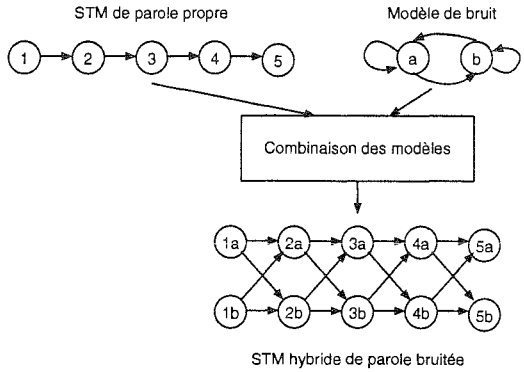

FIG. 6.3 - *Combinaison d'un STM de parole propre avec un HMM de bruit. Le modèle de parole bruitée devient un STM hybride.*

une séquence rééchantillonnée de Q vecteurs, mais à chaque « état », le vecteur bruité est modélisé par 2 distributions qui s'excluent mutuellement.

Étant donné une trajectoire bruitée, nous devons maintenant rechercher la « meilleure » séquence expliquant cette trajectoire (c.-à-d. la plus probable). En d'autres termes, comment calculer $p(\mathbf{Y}_n|t_k, d, s)$?

Notons $\delta_0(j)$, la probabilité d'observer le vecteur de parole bruitée associé à l'état 0 du STM lorsque le HMM de bruit est dans l'état j. On a :

$$\delta_0(j) = \pi_j \cdot \mathcal{N}(\mathbf{y}_{n-f(0,d,Q)}; \psi_m(\mathbf{m}^s_{k,0}, \tilde{\mathbf{m}}_j, \mathbf{\Sigma}^s_{k,0}, \tilde{\mathbf{\Sigma}}_j), \psi_\Sigma(\mathbf{m}^s_{k,0}, \tilde{\mathbf{m}}_j, \mathbf{\Sigma}^s_{k,0}, \tilde{\mathbf{\Sigma}}_j)) \tag{6.27}$$

Connaissant les $\delta_i(l)$ pour $1 \leq l \leq N$, nous recherchons le meilleur chemin permettant d'évoluer de l'état i vers l'état $i+1$ du STM en atteignant l'état j du HMM, soit encore comment calculer $\delta_{i+1}(j)$. Lorsque le HMM évolue d'un état vers un autre, un vecteur d'observation correspondant à une trame de bruit est généré. Lorsque le STM évolue d'un état vers l'état suivant, cela correspond à une évolution moyenne de $\frac{d-1}{Q-1}$ trames dans le signal de parole propre. Les évolutions le long d'une séquence d'observations ne s'effectuent donc pas à la même cadence selon que l'on progresse d'un état dans le HMM ou bien d'un état dans le STM. Notons $Pr(q_t = j|q_{t'} = l)$ la probabilité de passer de l'état l à l'état j du HMM en $t - t'$ transitions. Pour passer de $\delta_i(l)$ à $\delta_{i+1}(j)$, il faut donc effectuer $\frac{d-1}{Q-1}$ transitions dans le HMM. $\delta_{i+1}(j)$ s'obtient donc par la récurrence suivante, qui signifie qu'on recherche l'état l du HMM qui maximise la probabilité de passer de l'état l à l'état j du HMM en $\frac{d-1}{Q-1}$ transitions, et d'observer le vecteur $\mathbf{y}_{n-f(i+1,d,Q)}$ dans l'état $i+1$ du STM lorsqu'on vient de l'état i du STM :

$$\begin{aligned}\delta_{i+1}(j) = \max_{l=1,\ldots N} \Big[& \delta_i(l) \cdot Pr(q_{(i+1)\frac{d-1}{Q-1}} = j | q_{i\frac{d-1}{Q-1}} = l) \\ & \cdot \mathcal{N}(\mathbf{y}_{n-f(i+1,d,Q)}; \psi_m(\mathbf{m}^s_{k,(i+1)}, \tilde{\mathbf{m}}_j, \mathbf{\Sigma}^s_{k,(i+1)}, \tilde{\mathbf{\Sigma}}_j), \psi_\Sigma(\mathbf{m}^s_{k,(i+1)}, \tilde{\mathbf{m}}_j, \mathbf{\Sigma}^s_{k,(i+1)}, \tilde{\mathbf{\Sigma}}_j)) \Big]\end{aligned} \tag{6.28}$$

Nombre d'opérations élémentaires	STM standard ($N = 1$)	STM hybride ($N > 1$)
produits	Q	$\frac{N}{N-1}(2N^Q - N - 1)$
$\mathcal{N}(\cdot)$	Q	NQ

TAB. 6.1 - *Comparaison des nombres d'opérations élémentaires (produits et évaluations de lois normales multidimensionnelles) pour le calcul de $p(\mathbf{Y}_n|t_k, d, s)$ entre un STM standard ($N = 1$) et un STM hybride ($N > 1$).*

On obtient finalement :

$$p(\mathbf{Y}_n|t_k, d, s) = \max_{l=1,\ldots N} \delta_{(Q-1)}(l) \tag{6.29}$$

Lorsque le nombre d'états du HMM de bruit est supérieur à 1, les équations (6.28) et (6.29) sont utilisées à la place de l'équation (6.26) pour déterminer les valeurs de $p(\mathbf{Y}_n|t_k, d, s)$. Le nombre total d'opérations élémentaires (multiplication et calcul d'une loi normale de dimension D) est indiqué tableau 6.1, pour les configurations où $N = 1$ (STM bruité standard) et $N > 1$ (STM bruité hybride).

Il faut enfin noter que les équations (6.28) et (6.29) permettent d'évoluer le long d'un chemin optimal dans le modèle hybride, selon une approche type Viterbi. Une alternative possible consiste à effectuer la somme des probabilités le long de tous les chemins, ce qui correspond à appliquer un décodage de type Baum-Welch [Rabiner, 1989].

4 Conclusion

Dans cette partie, nous avons présenté comment obtenir une approximation d'un STM de parole bruitée à partir d'un STM de parole propre, d'un HMM de bruit et d'une règle de combinaison reliant parole bruitée à parole propre et bruit.

Plusieurs hypothèses ont été nécessaires pour déterminer l'expression des moyennes et matrices de covariances du STM de parole bruitée.

- Les signaux de parole et de bruit sont additifs dans le domaine des densités spectrales de puissance, ce qui constitue une hypothèse couvrant un grand nombre de situations pratiques.
- La présence du bruit ne modifie pas la localisation des trames de parole (ce qui signifie qu'une observation de parole propre affectée à l'état i d'un STM lors l'apprentissage, reste affectée à l'état i du STM après perturbation par le bruit).
- Les signaux de parole et de bruit sont indépendants.
- La somme de deux vecteurs aléatoires de distribution log-normale est un vecteur aléatoire de distribution log-normale. Cette condition revient à considérer que lorsque parole propre et bruit sont modélisés par des lois normales dans le domaine cepstral, les observations de parole bruitée sont également distribuées selon une loi normale.

Par contre, aucune hypothèse n'est faite sur la stationnarité du bruit. Un bruit non stationnaire peut être modélisé par un HMM ergodique à plusieurs états, et la combinaison du STM et du HMM donne naissance à un STM hybride où il est possible de calculer par programmation dynamique la valeur de la *pdf* d'une trajectoire dans une classe donnée.

Cette approche présente l'intérêt d'être non supervisée dès lors que l'on est capable de détecter les zones d'absence de parole dans le signal, afin de réestimer les modèles de bruit. Il est ainsi possible d'obtenir des modèles capable de s'adapter automatiquement aux conditions de bruits.

Les expériences et résultats concernant la combinaison de modèles stochastiques sont présentés et discutés au chapitre 9.

Chapitre 7

Filtrage non linéaire par états

1 Introduction

Dans notre étude bibliographique, nous avons insisté sur la nécessité de mettre en œuvre des traitements spécifiques aux différents sons, afin de prendre en compte les fluctuations des distorsions provoquées par un niveau de bruit constant, le long du signal. En effet, les régions du signal de parole de faible énergie sont plus perturbées par le bruit que les zones où l'énergie est importante.

D'autre part, nous avons également noté qu'il est intéressant, lorsqu'on effectue un filtrage de bruit, d'appliquer un critère d'estimation ayant une signification liée aux mécanismes de perception auditive.

Dans ce chapitre, nous utilisons les modèles stochastiques de trajectoires, estimés pour la reconnaissance, pour segmenter le signal bruité en zones considérées comme quasi-stationnaires, tout comme [Beattie et Young, 1991] utilisent des HMMs pour décomposer de la même façon le signal bruité. Une zone quasi-stationnaire est ainsi associée à chaque état d'un modèle stochastique de trajectoire. On peut considérer que le signal de parole propre paramétré est modélisé avec précision dans chacun des états des STMs, puisque ces modèles de trajectoires permettent d'effectuer la reconnaissance de la parole propre de façon efficace [Gong et Haton, 1994]. De même, il est aisé d'entraîner un modèle de bruit dans cet espace de paramètres. Nous mettons alors en œuvre un filtrage, spécifique à chaque état, et déterminé selon un critère de minimisation de l'erreur quadratique moyenne dans le domaine du logarithme de spectre. Ce critère est utilisé car il semble en accord avec les théories sur l'audition, qui précisent que l'oreille est plus sensible au logarithme de spectre, plutôt qu'au spectre dans le domaine linéaire. En définitive, on obtient un ensemble de filtres associés à chaque état des STMs (entraînés sur de la parole propre) sous la forme de tables de compensations, qui sont utilisées pour filtrer les observations bruitées pendant la reconnaissance. L'architecture générale du système de reconnaissance de parole bruitée est représentée fig. 7.1.

Nous présentons tout d'abord l'estimation spectrale par minimisation de l'erreur quadratique moyenne dans le domaine linéaire du spectre (cf. 2.1). L'estimateur est ensuite étendu pour appliquer le même critère, mais dans le domaine du logarithme du spectre (cf. 2.2). Le

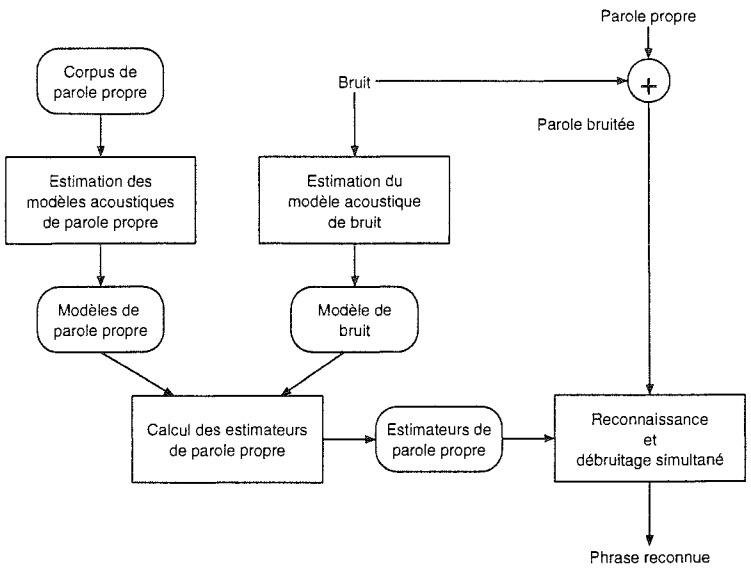

FIG. 7.1 - *Structure d'un système de reconnaissance de parole avec filtrage non linéaire des observations bruitées lors de la reconnaissance.*

calcul des estimateurs ne pouvant être effectué sous forme close, une résolution numérique est décrite au paragraphe 3.1, suivie de son application aux modèles stochastiques de trajectoires entraînés dans le domaine cepstral (cf. 3.2).

2 Estimation spectrale non linéaire

2.1 Soustraction spectrale linéaire

Comme dans le chapitre 6, désignons par $x(t)$ le signal temporel de parole propre, $n(t)$ le signal de bruit et $y(t)$ le signal de parole bruitée. Nous supposons encore que bruit et parole sont additifs dans le domaine temporel, et sont modélisés par des processus stochastiques. Si parole et bruit sont indépendants et stationnaires, l'additivité est conservée dans le domaine des densités spectrales de puissance, où l'on a :

$$Y(\omega) = X(\omega) + N(\omega) \tag{7.1}$$

La densité spectrale de puissance peut être estimée par le spectre de puissance à court terme, obtenu par l'intermédiaire d'une transformée de Fourier discrète. Notons $P_x(\omega)$, $P_n(\omega)$ et $P_y(\omega)$, les spectres de puissance à court terme de $x(t)$, $n(t)$ et $y(t)$, l'équation (7.1) se réécrit :

$$P_y(\omega) \approx P_x(\omega) + P_n(\omega) \tag{7.2}$$

Une estimation de $P_x(\omega)$, notée $\widehat{P}_x(\omega)$, peut être déterminée selon le critère d'estimation

de la minimisation de l'erreur quadratique moyenne (MMSE) :

$$E\{|\widehat{P}_x(\omega) - P_x(\omega)|^2\} \qquad \text{minimal} \tag{7.3}$$

Si les signaux sont modélisés par des processus Gaussiens, un tel critère correspond au filtrage par soustraction spectrale de puissance, qui se caractérise par sa grande simplicité, et on obtient alors :

$$\widehat{P}_x(\omega) = P_y(\omega) - \tilde{P}_n(\omega) \tag{7.4}$$

où $\tilde{P}_n(\omega)$ représente une estimation du spectre de puissance à court terme du bruit, obtenu généralement par la moyenne de $P_n(\omega)$ calculée sur des trames successives de bruit.

Comme nous l'avons déjà noté chapitre 1, § 1, cette soustraction est modifiée pour éviter d'obtenir une estimation négative du spectre de puissance, et des techniques *ad hoc* sont mises en œuvre pour minimiser le bruit musical introduit par ce filtrage.

2.2 Estimation spectrale non linéaire

Bien qu'ayant une énergie faible par rapport au bruit d'origine, le bruit musical introduit par la soustraction spectrale est très pénalisant. Une façon de réduire ce bruit consiste à estimer le spectre du signal filtré dans le domaine logarithmique [Erell et Weintraub, 1993a; Van Compernolle, 1989b]. Ceci est en concordance avec les théories sur l'audition humaine, qui précisent qu'il est préférable de minimiser les distorsions dans le domaine du logarithme du spectre, plutôt que dans le domaine spectral linéaire.

Soit $L_y(\omega)$, $L_x(\omega)$ et $L_n(\omega)$ les représentations de $P_y(\omega)$, $P_x(\omega)$ et $P_n(\omega)$ dans le domaine logarithmique. L'équation (7.2) s'écrit alors :

$$L_y(\omega) \approx \ln(\exp L_x(\omega) + \exp L_n(\omega)) \tag{7.5}$$

Dans le domaine logarithmique du spectre, le critère de minimisation de l'erreur quadratique moyenne de l'équation (7.3) devient :

$$E\{|\widehat{L}_x(\omega) - L_x(\omega)|^2\} \qquad \text{minimal} \tag{7.6}$$

ce qui donne, en laissant tomber l'indice ω :

$$\begin{aligned} \widehat{L}_x &= E\{L_x|L_y\} \\ &= \int L_x \cdot f_{L_x|L_y}(L_x|L_y)dL_x \\ &= \frac{\int L_x \cdot f_{L_xL_y}(L_x, L_y)dL_x}{\int f_{L_xL_y}(L_x, L_y)dL_x} \end{aligned} \tag{7.7}$$

où $f_{L_x|L_y}(L_x|L_y)$ désigne la *pdf* de L_x connaissant L_y et $f_{L_xL_y}(L_x, L_y)$ désigne la *pdf* conjointe de L_x et L_y.

La résolution de l'équation (7.7) nécessite de connaître l'expression de $f_{L_x L_y}(\cdot)$. En pratique, nous ne disposons pas d'une telle information. De plus, étant donné la nature non linéaire de la combinaison entre le logarithme du spectre de la parole propre L_x et du spectre du bruit L_n (équation (7.6)), il n'est pas possible d'exprimer $\widehat{L}_x$ sous forme close, connaissant $f_{L_x}(\cdot)$ et $f_{L_n}(\cdot)$. Par conséquent, nous allons résoudre l'équation (7.7) par une approche numérique semblable à celle utilisée par [Xie et Van Compernolle, 1993].

3 Estimation MMSE numérique

3.1 Principe

Soit un modèle de parole décrit dans le domaine logarithmique spectral par sa *pdf* $f_{L_x}(\cdot)$, de paramètre θ_x, et un modèle de bruit décrit dans le même domaine par $f_{L_n}(\cdot)$, de paramètre θ_n. L'estimation $\widehat{L}_x$ est alors une fonction de L_y et des paramètres θ_x et θ_n, inconnue sous forme close.

Cependant, connaissant $f_{L_x}(\cdot)$ et $f_{L_n}(\cdot)$, nous sommes capables de générer aléatoirement un ensemble très grand d'observations L_x et L_n respectant les *pdfs* $f_{L_x}(\cdot)$ et $f_{L_n}(\cdot)$. Nous pouvons alors déterminer un ensemble d'observations de parole bruitée L_y en combinant par la relation (7.5) les L_x et L_n ainsi générés. Il est alors possible, d'après la loi faible des grands nombres, d'estimer $\widehat{L}_x$ en calculant empiriquement $E\{L_x|L_y\}$. La précision de l'estimation est contrôlée par le nombre d'observations générées. Nous pouvons donc obtenir une table de correspondance entre les L_y et les $\widehat{L}_x$, pour différentes valeurs des paramètres θ_x et θ_n. Cela signifie que l'on peut associer une table de correspondance à différents modèles de parole, c.-à-d. éventuellement une table par modèle de phonème, ou encore, une table par état d'un modèle stochastique de trajectoire. On peut ainsi mettre en œuvre différents filtres spécifiques à des classes de sons.

L'estimation de $\widehat{L}_x$ suit donc le principe suivant [Xie et Van Compernolle, 1993] :

1. Générer N observations L_{x_i} de parole propre, respectant le modèle $f_{L_x}(L_x)$
2. Générer N observations L_{n_i} de bruit, respectant le modèle $f_{L_n}(L_n)$
3. Calculer N observations L_{y_i} en utilisant la relation (7.5)
4. Déterminer l'estimateur MMSE $\widehat{L}_x$ dans le domaine discret :

$$\widehat{L}_x(L_{y_k}) = E\{L_x|L_y \in [L_{y_k} - \Delta L_y/2, L_{y_k} + \Delta L_y/2]\}$$

avec :

$$\begin{aligned} \Delta L_y &= (L_{y_{max}} - L_{y_{min}})/K \\ L_{y_k} &= L_{y_{min}} + k \cdot \Delta L_y - \Delta L_y/2 \end{aligned} \tag{7.8}$$

Soit la fonction $g(L_y, L_{y_k})$ définie par :

$$g(L_y, L_{y_k}) = \begin{cases} 1 & \text{si } L_y \in [L_{y_k} - \Delta L_y/2, L_{y_k} + \Delta L_y/2], \\ 0 & \text{sinon.} \end{cases} \tag{7.9}$$

L'estimateur $\widehat{L}_x$ s'écrit finalement :

$$\widehat{L}_x(L_{y_k}) = \frac{\sum_{i=1}^{N} L_{x_i} \cdot g(L_{y_i}, L_{y_k})}{\sum_{j=1}^{N} g(L_{y_j}, L_{y_k})} \tag{7.10}$$

3.2 Application à des modèles stochastiques définis sur les cepstres

Nous souhaitons associer le processus de filtrage au processus de reconnaissance. Pour cela, nous utilisons comme modèles de parole propre les modèles stochastiques estimés pour la reconnaissance. Un filtre est associé à chaque état d'un STM, et le filtrage est effectué pendant le décodage. Cela signifie qu'une observation associée à un état va être transformée par le filtre avant d'être utilisée pour déterminer sa probabilité, étant donné le modèle de parole propre.

Pour chaque symbole s, nous souhaitons générer N trajectoires aléatoires $\mathbf{X}_i$ respectant le modèle défini par les équations (4.30), (4.31) et (4.32). Cela revient à générer pour chaque symbole s, et pour chaque composante t_k ($t_k \in \mathbb{T}_s$) du mélange de lois normales, $N \times Pr(t_k|s)$ vecteurs de cepstres de dimension D, de *pdf* égale à $\mathcal{N}(\mathbf{x}_i; \mathbf{m}^s_{k,i}, \mathbf{\Sigma}^s_{k,i})$ pour i variant de 0 à $Q-1$. Au total, on génère donc $N \times Q$ vecteurs de cepstre de dimension D, soit N vecteurs pour chaque état $i = 0, \ldots Q-1$ du modèle associé à chaque symbole s.

Nous sommes également capable d'estimer un modèle de bruit, en considérant que les vecteurs de bruit sont représentés dans le domaine cepstral par une simple *pdf* Gaussienne de moyenne $\mathbf{m}_b$ et de covariance $\mathbf{\Sigma}_b$. Un tel modèle peut être estimé en utilisant les zones d'absence de parole sur le signal. Il est alors possible de générer aléatoirement $N \times Q$ vecteurs de bruit respectant ce modèle.

Les calculs des estimateurs doivent être effectués dans le domaine du logarithme de la densité spectrale de puissance. Pour cela, les $N \times Q$ vecteurs de cepstre de parole propre spécifiques à chaque état d'un STM et les $N \times Q$ vecteurs de bruit sont convertis dans le domaine logarithmique par une transformation en cosinus, de la même façon que dans le chapitre 6, équation (6.8). Les $N \times Q$ vecteurs de parole bruitée spécifique à chaque état de chaque modèle sont alors calculés en appliquant la relation (7.5). Les estimateurs optimaux sont ensuite déterminés par calcul empirique des moyennes en utilisant l'approche présentée § 3.1, puis convertis dans le domaine cepstral par transformée en cosinus inverse. En définitive, à chaque état i de chaque modèle STM s est associée une table de compensation T_{si}, qui fournit une estimation d'un vecteur cepstral débruité à partir d'un vecteur de cepstre bruité, dont le critère de calcul est la minimisation des distorsions dans le domaine du logarithme du spectre.

Ces tables de compensation sont utilisées lors de la reconnaissance de la parole bruitée, pour transformer chaque vecteur bruité $\mathbf{y}$ en un vecteur débruité $\hat{\mathbf{x}} = T_{si}(\mathbf{y})$. La probabilité d'une trajectoire bruitée $\mathbf{Y}$, étant donné le symbole s et la composante du mélange t_k s'écrit :

$$p(\mathbf{Y}|t_k, s) = \prod_{i=0}^{Q-1} \mathcal{N}(T_{si}(\mathbf{y}_i); \mathbf{m}^s_{k,i}, \mathbf{\Sigma}^s_{k,i}) \tag{7.11}$$

4 Conclusion

Dans ce chapitre, nous avons présenté une approche permettant de filtrer la parole bruitée à reconnaître dans le domaine cepstral, pendant la reconnaissance. Cette transformation permet d'utiliser les modèles entraînés à partir de parole propre, afin de reconnaître de la parole perturbée par un bruit additif stationnaire.

La compensation mise en œuvre minimise les distorsions dans le domaine du logarithme du spectre de puissance à court terme. Un tel critère est préférable, d'un point de vue perceptif, à la minimisation des distorsions dans le domaine linéaire du spectre.

Associer un filtre à chaque état de chaque STM permet d'appliquer une transformation spécifique à chaque classe de son. Ainsi, une zone de parole n'est pas transformée de la même façon selon qu'elle est supposée provenir de tel ou tel état, de tel ou tel symbole. Cela permet de prendre en considération les fluctuations du rapport signal-à-bruit instantané. Les fonctions de densités de probabilités utilisées pour représenter le signal de parole propre sont précises, car elles proviennent d'un apprentissage réalisé sur une base de données de taille suffisante pour assurer une reconnaissance correcte en absence de bruit. À part l'additivité de la parole et du bruit dans le domaine de la densité spectrale de puissance, aucune hypothèse supplémentaire n'est faite sur le bruit et la parole, en particulier sur les natures des distributions.

Tout comme l'approche présentée au chapitre 6, la méthode est non supervisée à partir du moment où l'on dispose d'un modèle de bruit, estimé pendant les zones d'absence de parole. Cependant, cette approche est très lourde. Lorsque le type et le rapport signal-à-bruit moyen évoluent, il est nécessaire de recalculer les estimateurs, ce qui se traduit par une charge de calcul considérable. Ainsi, il est hors de question de mettre en œuvre une telle méthode pour un fonctionnement en ligne du système de reconnaissance.

Cette approche est évaluée et comparée avec nos autres méthodes au chapitre 9.

Chapitre 8

Adaptation au bruit par régression linéaire

1 Introduction

Nous proposons ici deux approches duales permettant de réduire la dégradation des performances d'un système de reconnaissance automatique de la parole lorsque les conditions de test diffèrent des conditions d'apprentissage. Contrairement aux chapitres 6 et 7 où une connaissance explicite sur la nature des perturbations entre les conditions de test et d'apprentissage était exploitée (parole et bruit additif e. indépendants), l'origine des différences entre environnement de test et d'apprentissage n'est pas ici prise explicitement en considération. En particulier, on ne suppose pas que la perturbation est provoquée par un bruit additif. La connaissance sur l'environnement de test provient uniquement d'un corpus d'adaptation de parole bruitée, de taille réduite. Nous supposons cependant que les différences entre conditions de test et d'apprentissage peuvent être modélisées par des transformations linéaires par morceaux des espaces acoustiques dans le domaine cepstral. Une telle restriction à des transformations linéaires par parties, voire linéaires, se révèle en effet suffisamment efficace pour des problèmes d'adaptation au bruit [Chollet *et al.*, 1990; Mokbel, 1992], au locuteur [Matsukoto et Inoue, 1992; Class *et al.*, 1990; Bellagarda *et al.*, 1992; Tubach *et al.*, 1991], au canal d'enregistrement [Mokbel *et al.*, 1994; Takahashi et Sagayama, 1994] et à l'effet Lombard [Chen, 1988]. Bien évidemment, l'utilisation de transformations non-linéaires permettrait de modéliser avec plus de précision les variations entre espaces de test et d'apprentissage, mais elle ne conduit pas en général à l'obtention de l'expression des transformations sous forme close.

Dans chacune des deux approches, les transformations mises en œuvre sont des transformations linéaires des vecteurs de paramètres, semblables à celles utilisées par [Leggetter et Woodland, 1994b; Leggetter et Woodland, 1994a] pour une tâche d'adaptation au locuteur. Dans notre étude bibliographique, nous avons insisté sur la nécessité d'appliquer des transformations spécifiques à différentes zones de l'espace acoustique, par opposition à une transformation globale de tout l'espace. Ici, nous associons une transformation linéaire à chaque modèle acoustique, ce qui signifie que l'espace acoustique sera transformé linéairement par

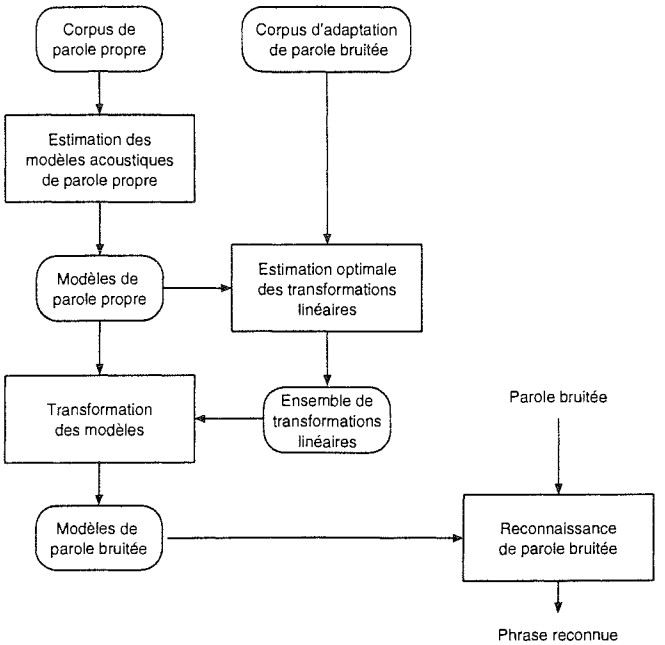

FIG. 8.1 - *Structure d'un système de reconnaissance de parole bruitée avec transformation par régression linéaire des paramètres des STM de parole propre.*

morceaux.

La première approche proposée consiste à transformer les paramètres des modèles de parole propre, afin de construire des modèles spécifiques aux conditions de test sans nécessiter un réentraînement total du système. Les modèles transformés sont ensuite utilisés pour reconnaître la parole de test. En reconnaissance, l'adaptation ne conduit pas à des calculs supplémentaires par rapport au système initial. Le schéma général d'un système de reconnaissance de parole utilisant ce principe est représenté fig. 8.1.

La seconde approche consiste à transformer la parole de test, afin de la rendre la plus « proche » possible de la parole utilisée pour l'apprentissage des modèles. On utilise ensuite les modèles de référence pour reconnaître la parole de test transformée. Étant donné que les transformations sont spécifiques à chaque symbole, il n'est pas possible de pré-traiter directement la parole de test. Les transformations sont en fait appliquées lors du calcul de la probabilité d'une trajectoire (transformée), étant donné un modèle de symbole. Contrairement à l'approche précédente, la transformation de la parole de test conduit à une surcharge importante en calculs. Le principe d'un système de reconnaissance basé sur cette transformation est représenté fig. 8.2.

2 Principe de l'adaptation

La fonction de densité de probabilité d'une trajectoire $\mathbf{X}$ connaissant la durée d et le symbole s est définie par l'équation (4.30). Afin de simplifier les notations, nous ne faisons d'une part plus apparaître cette dépendance sur d et s, et d'autre part, nous introduisons Φ^s, qui re-

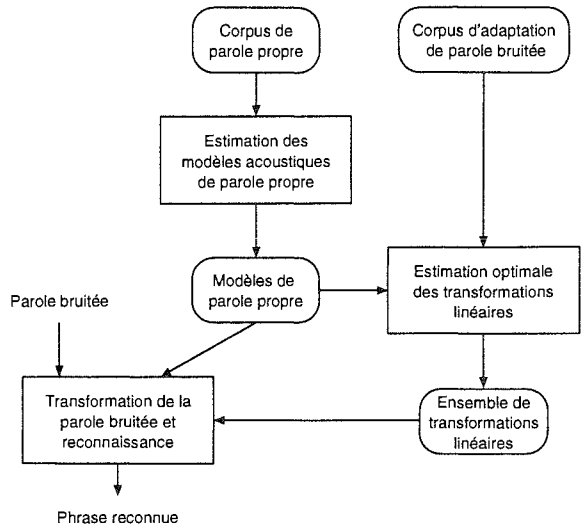

FIG. 8.2 - *Structure d'un système de reconnaissance de parole bruitée avec transformation des observations de parole bruitée lors de la reconnaissance.*

présente l'ensemble des paramètres du STM du symbole s. L'équation (4.30) s'écrit alors :

$$p(\mathbf{X}_n|\Phi^s) \triangleq \sum_{t_k \in \mathbb{T}_s} Pr(t_k|\Phi^s) \cdot p(\mathbf{X}_n|t_k, \Phi^s) \tag{8.1}$$

Selon que l'adaptation porte sur le modèle ou sur la parole de test, l'équation (8.1) sera différemment modifiée. Le paragraphe 2.1 présente comment le calcul de la *pdf* d'une trajectoire de parole bruitée va être transformée par l'adaptation du modèle. Dans le paragraphe 2.2, ce même calcul est décrit lorsque la transformation porte sur la trajectoire bruitée. Le paragraphe 2.3 présente le critère utilisé pour estimer les transformations.

2.1 Transformation des paramètres d'un STM

Nous nous limitons à l'adaptation des vecteurs moyennes des STMs. Les matrices de covariances et les coefficients de mixtures ne seront pas adaptés et conserveront par conséquent leurs valeurs initiales. Adapter uniquement les vecteurs moyennes des modèles est une approche courante pour l'adaptation au bruit des systèmes à base de HMMs continus [Frangoulis et Gaganelis, 1992; Mizuta et Nakajima, 1992; Morii *et al.*, 1990; Nolazco Flores et Young, 1994], dont une forme équivalente est l'adaptation des prototypes des HMMs discrets. L'adaptation des vecteurs moyennes est généralement plus simple à mettre en œuvre que l'adaptation des matrices de covariances, et conduit, comme nous le verrons, à des résultats satisfaisants.

Le vecteur moyenne de la loi normale multidimensionnelle représentant la *pdf* d'une trajectoire $\mathbf{X}$ connaissant la classe de trajectoires t_k est adapté en utilisant une transformation linéaire. Si $\mathbf{m}_k^s$ désigne le vecteur moyenne pour une classe de trajectoire t_k, avec

$\mathbf{m}_k^{s\#} = [m_{k,1}^s, \ldots m_{k,L}^s]$, la moyenne étendue $\hat{\mathbf{m}}_k^s$, avec $\hat{\mathbf{m}}_k^{s\#} = [\hat{m}_{k,1}^s, \ldots \hat{m}_{k,L+1}^s]$, est définie par l'expression suivante où ω représente un terme d'offset constant :

$$\hat{\mathbf{m}}_k^s \triangleq \begin{bmatrix} m_{k,1}^s \\ \vdots \\ m_{k,L}^s \\ \omega \end{bmatrix} \tag{8.2}$$

La moyenne adaptée $\tilde{\mathbf{m}}_k^s$ est obtenue par transformation linéaire en utilisant la matrice de transformation $\mathbf{W}$ de dimension $L \times (L+1)$:

$$\tilde{\mathbf{m}}_k^s = \mathbf{W}\hat{\mathbf{m}}_k^s \tag{8.3}$$

La *pdf* d'une trajectoire de test $\mathbf{X}$ étant donnée t_k s'écrit en utilisant un modèle adapté :

$$p(\mathbf{X}|t_k, \Phi^s) = \frac{1}{(2\pi)^{L/2}|\boldsymbol{\Sigma}_k^s|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{X} - \tilde{\mathbf{m}}_k^s)^{\#}\boldsymbol{\Sigma}_k^{s\,-1}(\mathbf{X} - \tilde{\mathbf{m}}_k^s)\right] \tag{8.4}$$

Lorsque les moyennes des différents STMs ont été adaptées, les modèles adaptés sont utilisés pour la reconnaissance de la parole de test, et la méthode n'introduit pas de calculs supplémentaires.

2.2 Transformation des vecteurs de test

D ns cette approche, duale de la précédente, les vecteurs de test sont transformés à l'aide d'une transformation linéaire, et les modèles ne sont pas modifiés. Si $\mathbf{X}$ désigne une trajectoire de test, la trajectoire étendue $\hat{\mathbf{X}}$, avec $\hat{\mathbf{X}}^{\#} = [\hat{x}_1, \ldots \hat{x}_{L+1}]$, est définie par l'expression suivante où ω représente un terme d'offset constant :

$$\hat{\mathbf{X}} \triangleq \begin{bmatrix} x_1 \\ \vdots \\ x_L \\ \omega \end{bmatrix} \tag{8.5}$$

La trajectoire adaptée $\tilde{\mathbf{X}}$ est obtenue par transformation linéaire en utilisant la matrice de transformation $\mathbf{W}$ de dimension $L \times (L+1)$:

$$\tilde{\mathbf{X}} = \mathbf{W}\hat{\mathbf{X}} \tag{8.6}$$

La *pdf* d'une trajectoire de test $\mathbf{X}$ étant donnée t_k s'écrit après adaptation :

$$p(\mathbf{X}|t_k, \Phi^s) = \frac{1}{(2\pi)^{L/2}|\boldsymbol{\Sigma}_k^s|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{W}\hat{\mathbf{X}} - \mathbf{m}_k^s)^{\#}\boldsymbol{\Sigma}_k^{s\,-1}(\mathbf{W}\hat{\mathbf{X}} - \mathbf{m}_k^s)\right] \tag{8.7}$$

La matrice de transformation est associée à chaque modèle, et la transformation est donc spécifique au symbole. À chaque instant n et pour chaque durée d, la trajectoire $\mathbf{X}_n$ va être transformée par la matrice associée à chaque modèle s. La transformation introduit donc une surcharge de calcul importante par rapport à l'approche précédente.

2.3 Critère d'estimation des transformations

Soit $\mathcal{X}$ le corpus d'adaptation, constitué d'un ensemble de N trajectoires rééchantillonnées : $\mathcal{X} = \{\mathbf{X}_1, \ldots \mathbf{X}_N\}$. Nous supposons que ces trajectoires sont étiquetées : à chaque trajectoire $\mathbf{X}_n$ est associé le symbole s correspondant. Si on dispose au départ de H modèles acoustiques, la probabilité d'observer les N trajectoires d'adaptation s'écrit :

$$p(\mathbf{X}_1, \ldots \mathbf{X}_n, \ldots \mathbf{X}_N) = \prod_{s=1}^{H} \prod_{n=1}^{N} p(\mathbf{X}_n|\Phi^s)^{\delta_s(n)} \tag{8.8}$$

avec

$$\delta_s(n) = \begin{cases} 1 & \text{si } \mathbf{X}_n \text{ est étiqueté } s, \\ 0 & \text{sinon.} \end{cases} \tag{8.9}$$

L'objectif du calcul est de déterminer l'expression de la matrice de transformation $\mathbf{W}^s$ associée au symbole s, qui maximise la vraisemblance du corpus d'adaptation, soit :

$$\mathbf{W}^s = \underset{\mathbf{W}^s}{\operatorname{argmax}}\, p(\mathbf{X}_1, \ldots \mathbf{X}_N) \tag{8.10}$$

Nous souhaitons que le corpus d'adaptation destiné à estimer les transformations soit de taille réduite. Or, lorsque le volume de données d'adaptation est faible, il devient impossible d'associer une transformation à chaque modèle de référence. En effet, pour certains modèles il peut n'y avoir aucune données d'adaptation, alors que pour d'autres le nombre de données d'adaptation peut être insuffisant pour permettre une estimation robuste des transformations.

Par conséquent, afin d'éviter une estimation non robuste des transformations lorsque le volume de données d'adaptation est faible, nous introduisons la possibilité d'associer une même transformation linéaire à différents modèles. La quantité de données d'adaptation disponible détermine le degré de partage d'une transformation entre différents modèles. Plus le volume de données d'adaptation sera faible, plus il sera nécessaire d'associer une même transformation à différents modèles.

3 Adaptation des paramètres des STMs

La détermination des matrices de transformations selon l'équation (8.10) n'est pas immédiate. Afin de faire apparaître les différentes étapes de la résolution, et la signification des transformations, nous étudions tout d'abord, paragraphe 3.1, le cas où les STMs ne comportent qu'une seule mixture. Dans le paragraphe 3.2, nous étendons le calcul au cas où le nombre de mixtures est supérieur à 1; il n'est alors plus possible d'exprimer les matrices $\mathbf{W}^s$ sous forme close, et la résolution s'effectue par un procédé itératif à l'aide de l'algorithme EM.

3.1 STMs sans mixtures

Dans un premier temps, et afin de simplifier la présentation, nous considérons qu'à chaque symbole s est associé une matrice de transformation $\mathbf{W}^s$. Dans le paragraphe 3.1.2, nous

généralisons les expressions obtenues pour permettre le partage d'une même transformation entre un ensemble de modèles.

3.1.1 Une matrice de régression par symbole

Exprimons l'équation (8.8) dans le domaine logarithmique. La log-vraisemblance des données d'adaptation s'écrit :

$$\begin{aligned}\log p(\mathcal{X}) &= \sum_{s=1}^{H}\sum_{n=1}^{N}\delta_s(n)\log p(\mathbf{X}_n|\Phi^s) \\ &= \sum_{s=1}^{H}\sum_{n=1}^{N}\delta_s(n)\Big[-\frac{L}{2}\log(2\pi) - \frac{1}{2}\log|\mathbf{\Sigma}^s| \\ &\qquad - \frac{1}{2}(\mathbf{X}_n - \mathbf{W}^s\hat{\mathbf{m}}^s)^{\#}\mathbf{\Sigma}^{s-1}(\mathbf{X}_n - \mathbf{W}^s\hat{\mathbf{m}}^s)\Big]\end{aligned} \tag{8.11}$$

Pour un symbole s donné, on cherche le $\mathbf{W}^s$ qui maximise $\log p(\mathcal{X})$. $\mathbf{W}^s$ est tel que :

$$\frac{\partial}{\partial \mathbf{W}^s}\log p(\mathcal{X}) = \mathbf{0} \tag{8.12}$$

soit :

$$\sum_{n=1}^{N}\delta_s(n)\left[\mathbf{\Sigma}^{s-1}(\mathbf{X}_n - \mathbf{W}^s\hat{\mathbf{m}}^s)\hat{\mathbf{m}}^{s\#}\right] = \mathbf{0} \tag{8.13}$$

On obtient finalement :

$$\sum_{n=1}^{N}\delta_s(n)\mathbf{\Sigma}^{s-1}\mathbf{X}_n\hat{\mathbf{m}}^{s\#} = \sum_{n=1}^{N}\delta_s(n)\mathbf{\Sigma}^{s-1}\mathbf{W}^s\hat{\mathbf{m}}^s\hat{\mathbf{m}}^{s\#} \tag{8.14}$$

Cette expression se simplifie, et la moyenne adaptée pour le symbole s est donc :

$$\tilde{\mathbf{m}}^s = \mathbf{W}^s\hat{\mathbf{m}}^s = \frac{\sum_{n=1}^{N}\delta_s(n)\mathbf{X}_n}{\sum_{n=1}^{N}\delta_s(n)} \tag{8.15}$$

On obtient ainsi directement l'expression de la moyenne adaptée, ce qui correspond à la forme habituelle de l'estimation de la moyenne au sens du maximum de la vraisemblance.

3.1.2 Partage des matrices de transformation entre symboles

L'expression (8.14) suppose que la matrice de transformation $\mathbf{W}^s$ est spécifique au symbole s. Nous généralisons cette expression pour permettre le partage d'une même matrice de transformation entre différents symboles. La sommation doit alors être effectuée sur tous

les symboles partagés. Supposons que $\mathbf{W}^s$ soit partagé entre les R symboles $\{s_1, \ldots s_R\}$. L'équation (8.14) devient :

$$\sum_{n=1}^{N}\sum_{r=1}^{R}\delta_{s_r}(n)\Sigma^{s_r\,-1}\mathbf{X}_n\hat{\mathbf{m}}^{s_r\#} = \sum_{n=1}^{N}\sum_{r=1}^{R}\delta_{s_r}(n)\Sigma^{s_r\,-1}\mathbf{W}^s\hat{\mathbf{m}}^{s_r}\hat{\mathbf{m}}^{s_r\#} \tag{8.16}$$

Notons $\mathbf{Z}$ la matrice de dimension $L \times (L+1)$, partie gauche de l'équation (8.16). Soit Υ^{s_r} la matrice $\Sigma^{s_r\,-1}$ d'éléments $v_{ip}^{s_r}$, $x_{n,p}$ la p-ième composante du vecteur $\mathbf{X}_n$, et $\hat{m}_j^{s_r}$ la j-ième composante du vecteur $\hat{\mathbf{m}}^{s_r}$. Les éléments z_{ij} de $\mathbf{Z}$ s'écrivent :

$$z_{ij} = \sum_{n=1}^{N}\sum_{r=1}^{R}\delta_{s_r}(n)\cdot\left(\sum_{p=1}^{L} v_{ip}^{s_r}\cdot x_{n,p}\right)\cdot\hat{m}_j^{s_r} \qquad \text{avec} \quad \begin{matrix}1\le i\le L\\ 1\le j\le(L+1)\end{matrix} \tag{8.17}$$

Les matrices de précision Υ^{s_r} étant diagonales, l'équation (8.17) se simplifie :

$$z_{ij} = \sum_{n=1}^{N}\sum_{r=1}^{R}\delta_{s_r}(n)\cdot v_{ii}^{s_r}\cdot x_{n,i}\cdot\hat{m}_j^{s_r} \tag{8.18}$$

Notons $\mathbf{Y}$ la matrice de dimension $L \times (L+1)$, partie droite de l'équation (8.16). Les éléments y_{ij} de $\mathbf{Y}$ s'écrivent :

$$y_{ij} = \sum_{n=1}^{N}\sum_{r=1}^{R}\delta_{s_r}(n)\cdot\left(\sum_{q=1}^{L+1}\sum_{p=1}^{L} v_{ip}^{s_r}\cdot w_{pq}^{s}\cdot\hat{m}_q^{s_r}\right)\cdot\hat{m}_j^{s_r} \qquad \text{avec} \quad \begin{matrix}1\le i\le L\\ 1\le j\le L+1\end{matrix} \tag{8.19}$$

Les matrices de précision Υ^{s_r} étant diagonales, l'équation (8.19) devient :

$$y_{ij} = \sum_{q=1}^{L+1} w_{iq}^{s}\cdot\left(\sum_{n=1}^{N}\sum_{r=1}^{R}\delta_{s_r}(n)\cdot v_{ii}^{s_r}\cdot\hat{m}_j^{s_r}\cdot\hat{m}_q^{s_r}\right) \tag{8.20}$$

Soit $\mathbf{G}^i$ la matrice $(L+1)\times(L+1)$, dont les termes g_{jq}^i sont définis par :

$$g_{jq}^{i} = \sum_{n=1}^{N}\sum_{r=1}^{R}\delta_{s_r}(n)\cdot v_{ii}^{s_r}\cdot\hat{m}_j^{s_r}\cdot\hat{m}_q^{s_r} \qquad \text{avec} \quad \begin{matrix}1\le i\le L\\ 1\le j,q\le(L+1)\end{matrix} \tag{8.21}$$

L'équation (8.16) peut se réécrire, en utilisant (8.18), (8.20) et (8.21) sous la forme de L systèmes de $(L+1)$ équations linéaires à $(L+1)$ inconnues :

$$z_{ij} = \sum_{q=1}^{L+1} w_{iq}^{s} g_{jq}^{i} \qquad \text{avec} \quad \begin{matrix}1\le i\le L\\ 1\le j\le(L+1)\end{matrix} \tag{8.22}$$

La ligne i de la matrice $\mathbf{W}^s$ est finalement obtenue en résolvant le système suivant :

$$\mathbf{z}_i^{\#} = \mathbf{G}^i\mathbf{w}_i^{s\#} \tag{8.23}$$

où $\mathbf{w}_i^s$ désigne la i-ième ligne de la matrice $\mathbf{W}^s$ et $\mathbf{z}_i$ représente la i-ième ligne de la matrice $\mathbf{Z}$. La résolution du système (8.23) s'effectue sans difficulté par une méthode numérique type décomposition LU [Press *et al.*, 1988].

3.2 STMs avec mixtures

3.2.1 Calcul de la fonction auxiliaire $Q(\Phi, \overline{\Phi})$

Nous nous plaçons maintenant dans le cas où les modèles de trajectoires comportent plusieurs mixtures. La vraisemblance du corpus d'adaptation s'écrit :

$$p(\mathbf{X}_1, \ldots \mathbf{X}_n, \ldots \mathbf{X}_N) = \prod_{s=1}^{H} \prod_{n=1}^{N} \left(\sum_{t_k \in \mathbb{T}_s} Pr(t_k|\Phi^s) \cdot p(\mathbf{X}_n|t_k, \Phi^s) \right)^{\delta_s(n)} \tag{8.24}$$

Rappelons que $p(\mathbf{X}_n|t_k, \Phi^s)$ est calculé selon l'équation (8.4). Nous allons tout d'abord chercher à maximiser la vraisemblance pour une seule trajectoire $\mathbf{X}_n$; nous généraliserons par la suite le calcul pour prendre en considération l'ensemble des N observations.

La maximisation de la vraisemblance de $\mathbf{X}_n$ n'est pas possible directement car nous ne connaissons pas *a priori* la mixture associée à la trajectoire $\mathbf{X}_n$. Il est cependant possible de résoudre le problème en utilisant l'algorithme EM. D'après EM, la maximisation de la log-vraisemblance d'une donnée « incomplète » peut être obtenue en maximisant de façon itérative une fonction auxiliaire notée $Q_n(\Phi^s, \overline{\Phi}^s)$, représentant l'espérance mathématique de la log-vraisemblance de la donnée «complète», connaissant Φ^s et la donnée incomplète. Dans le cas présent, la donnée incomplète est $\mathbf{X}_n$, et nous formons la donnée complète $(\mathbf{X}_n, t_{y_n})$ en introduisant t_{y_n}, donnée non observée, désignant la classe de trajectoire de $p(\mathbf{X}_n|t_{y_n}, \Phi^s)$ et de $Pr(t_{y_n}|\Phi^s)$. Φ^s représente les paramètres courants du STM du symbole s et $\overline{\Phi}^s$ la nouvelle estimation que l'on cherche à déterminer.

Par définition, la fonction auxiliaire $Q_n(\Phi^s, \overline{\Phi}^s)$ s'écrit :

$$\begin{aligned} Q_n(\Phi^s, \overline{\Phi}^s) &\triangleq E\left\{\log p(\mathbf{X}_n, t_{y_n}|\overline{\Phi}^s)|\mathbf{X}_n, \Phi^s\right\} \\ &= \int \log p(\mathbf{X}_n, t_{y_n}|\overline{\Phi}^s) \cdot Pr(t_{y_n}|\mathbf{X}_n, \Phi^s) dt_{y_n} \\ &= \sum_{t_{y_n} \in \mathbb{T}_s} \log p(\mathbf{X}_n, t_{y_n}|\overline{\Phi}^s) \cdot Pr(t_{y_n}|\mathbf{X}_n, \Phi^s) \quad \text{car discret} \end{aligned} \tag{8.25}$$

où $Pr(t_{y_n}|\mathbf{X}_n, \Phi^s)$ s'exprime par la relation de Bayes :

$$\begin{aligned} Pr(t_{y_n}|\mathbf{X}_n, \Phi^s) &= \frac{p(\mathbf{X}_n, t_{y_n}|\Phi^s)}{p(\mathbf{X}_n|\Phi^s)} \\ &= \frac{Pr(t_{y_n}|\Phi^s) p(\mathbf{X}_n|t_{y_n}, \Phi^s)}{\sum_{t_{y_n} \in \mathbb{T}_s} Pr(t_{y_n}|\Phi^s) p(\mathbf{X}_n|t_{y_n}, \Phi^s)} \end{aligned} \tag{8.26}$$

Supposons maintenant que l'on dispose de N observations $\mathbf{X}_n$ étiquetées. La fonction

auxiliaire totale est la somme des fonctions auxiliaires de chaque observation $\mathbf{X}_n$:

$$\begin{aligned} Q(\Phi, \overline{\Phi}) &= \sum_{s=1}^{H} \sum_{n=1}^{N} \delta_s(n) Q_n(\Phi^s, \overline{\Phi}^s) \\ &= \sum_{s=1}^{H} \sum_{n=1}^{N} \sum_{t_{y_n} \in \mathbb{T}_s} \delta_s(n) Pr(t_{y_n}|\mathbf{X}_n, \Phi^s) \log p(\mathbf{X}_n, t_{y_n}|\overline{\Phi}^s) \\ &= \sum_{s=1}^{H} \sum_{n=1}^{N} \sum_{t_{y_n} \in \mathbb{T}_s} \delta_s(n) Pr(t_{y_n}|\mathbf{X}_n, \Phi^s) \log \left[\overline{Pr}(t_{y_n}|\Phi^s) p(\mathbf{X}_n|t_{y_n}, \overline{\Phi}^s)\right] \end{aligned} \tag{8.27}$$

La somme est effectuée sur tous les y_n, par conséquent y_n ne dépend pas de n, et on le remplace par k.

$$\begin{aligned} Q(\Phi, \overline{\Phi}) = &\sum_{s=1}^{H} \sum_{n=1}^{N} \sum_{t_k \in \mathbb{T}_s} \delta_s(n) Pr(t_k|\mathbf{X}_n, \Phi^s) \log \overline{Pr}(t_k|\Phi^s) \\ &+ \sum_{s=1}^{H} \sum_{n=1}^{N} \sum_{t_k \in \mathbb{T}_s} \delta_s(n) Pr(t_k|\mathbf{X}_n, \Phi^s) \log p(\mathbf{X}_n|t_k, \overline{\Phi}^s) \end{aligned} \tag{8.28}$$

3.2.2 Dérivation des expressions de réestimation de $\overline{\mathbf{W}}^s$

Puisque nous ne réestimons que la matrice de transformation $\overline{\mathbf{W}}^s$ seul le second terme $Q_{\overline{\mathbf{W}}^s}$ de la partie droite de l'équation (8.28) doit être maximisé :

$$Q_{\overline{\mathbf{W}}^s} = \sum_{s=1}^{H} \sum_{n=1}^{N} \sum_{t_k \in \mathbb{T}_s} \delta_s(n) Pr(t_k|\mathbf{X}_n, \Phi^s) \log p(\mathbf{X}_n|t_k, \overline{\Phi}^s) \tag{8.29}$$

Ce calcul est similaire à celui effectué pour le cas comportant une seule mixture. Les L systèmes de $(L+1)$ équations à résoudre à chaque itération de EM sont obtenus à partir de :

$$\begin{aligned} \sum_{t_k \in \mathbb{T}_{s_r}} \sum_{n=1}^{N} \sum_{r=1}^{R} \delta_{s_r}(n) Pr(t_k|\mathbf{X}_n, \Phi^{s_r}) {\Sigma_k^{s_r}}^{-1} \mathbf{X}_n \hat{\mathbf{m}}_k^{s_r \#} = \\ \sum_{t_k \in \mathbb{T}_{s_r}} \sum_{n=1}^{N} \sum_{r=1}^{R} \delta_{s_r}(n) Pr(t_k|\mathbf{X}_n, \Phi^{s_r}) {\Sigma_k^{s_r}}^{-1} \overline{\mathbf{W}}^s \hat{\mathbf{m}}_k^{s_r} \hat{\mathbf{m}}_k^{s_r \#} \end{aligned} \tag{8.30}$$

La valeur de $Pr(t_k|\mathbf{X}_n, \Phi^{s_r})$ est obtenue par application de la règle de Bayes, comme dans l'équation (8.26).

Comme précédemment, désignons par $\mathbf{Z}$ la partie gauche de l'équation (8.30). Les éléments z_{ij} de la matrice $\mathbf{Z}$ sont :

$$z_{ij} = \sum_{t_k \in \mathbb{T}_{s_r}} \sum_{n=1}^{N} \sum_{r=1}^{R} \delta_{s_r}(n) \cdot Pr(t_k|\mathbf{X}_n, \Phi^{s_r}) \cdot \left(\sum_{p=1}^{L} v_{k,ip}^{s_r} \cdot x_{n,p} \right) \cdot \hat{m}_{k,j}^{s_r}$$

$$\text{avec} \quad \begin{matrix} 1 \leq i, p \leq L \\ 1 \leq j \leq (L+1) \end{matrix} \tag{8.31}$$

Comme $\Upsilon_k^{s_r}$ est diagonale, l'équation (8.31) devient :

$$z_{ij} = \sum_{t_k \in \mathbb{T}_{s_r}} \sum_{n=1}^{N} \sum_{r=1}^{R} \delta_{s_r}(n) \cdot Pr(t_k|\mathbf{X}_n, \Phi^{s_r}) \cdot \upsilon_{k,ii}^{s_r} \cdot x_{n,i} \cdot \hat{m}_{k,j}^{s_r} \tag{8.32}$$

Soit $\mathbf{Y}$ la partie droite de l'équation (8.30). Les éléments y_{ij} de $\mathbf{Y}$ s'écrivent :

$$y_{ij} = \sum_{t_k \in \mathbb{T}_{s_r}} \sum_{n=1}^{N} \sum_{r=1}^{R} \delta_{s_r}(n) \cdot Pr(t_k|\mathbf{X}_n, \Phi^{s_r}) \cdot \left(\sum_{q=1}^{L+1} \sum_{p=1}^{L} \upsilon_{k,ip}^{s_r} \cdot w_{pq}^{s} \cdot \hat{m}_{k,q}^{s_r} \right) \cdot \hat{m}_{k,j}^{s_r}$$

$$\text{avec} \quad \begin{matrix} 1 \leq i \leq L \\ 1 \leq j \leq L+1 \end{matrix} \tag{8.33}$$

Les matrices de précision Υ^{s_r} étant diagonales, l'équation (8.33) se simplifie :

$$y_{ij} = \sum_{q=1}^{L+1} w_{iq}^{s} \cdot \left(\sum_{t_k \in \mathbb{T}_{s_r}} \sum_{n=1}^{N} \sum_{r=1}^{R} \delta_{s_r}(n) \cdot Pr(t_k|\mathbf{X}_n, \Phi^{s_r}) \cdot \upsilon_{k,ii}^{s_r} \cdot \hat{m}_{k,j}^{s_r} \cdot \hat{m}_{k,q}^{s_r} \right) \tag{8.34}$$

Soit $\mathbf{G}^i$ la matrice $(L+1) \times (L+1)$, dont les termes g_{jq}^i sont définis par :

$$g_{jq}^{i} = \sum_{t_k \in \mathbb{T}_{s_r}} \sum_{n=1}^{N} \sum_{r=1}^{R} \delta_{s_r}(n) \cdot Pr(t_k|\mathbf{X}_n, \Phi^{s_r}) \cdot \upsilon_{k,ii}^{s_r} \cdot \hat{m}_{k,j}^{s_r} \cdot \hat{m}_{k,q}^{s_r}$$

$$\text{avec} \quad \begin{matrix} 1 \leq i \leq L \\ 1 \leq j, q \leq (L+1) \end{matrix} \tag{8.35}$$

En utilisant (8.32), (8.34), et (8.35), l'équation (8.30) peut se réécrire sous une forme identique à (8.22). Une estimation de $\overline{\mathbf{W}}^s$ est alors obtenue à chaque itération de EM par résolution des L systèmes décrits par l'expression (8.23).

4 Adaptation des vecteurs de test

Cette approche est la duale de la précédente. La fonction objective à maximiser est toujours l'équation (8.8), mais avec le $p(\mathbf{X}_n|\Phi^s)$ correspondant à l'équation (8.7).

Comme précédemment, nous traitons d'abord le cas où les STMs ne comportent qu'une seule mixture. Nous développons par la suite le calcul pour prendre en compte un nombre de mixtures supérieur à 1.

4.1 STMs sans mixtures

Nous traitons directement le cas où plusieurs symboles partagent une même transformation. Les L systèmes de $(L+1)$ équations sont représentés par :

$$\sum_{n=1}^{N} \sum_{r=1}^{R} \delta_{s_r}(n) \boldsymbol{\Sigma}^{s_r\,-1} \mathbf{m}^{s_r} \hat{\mathbf{X}}_n^{\#} = \sum_{n=1}^{N} \sum_{r=1}^{R} \delta_{s_r}(n) \boldsymbol{\Sigma}^{s_r\,-1} \mathbf{W}^{s} \hat{\mathbf{X}}_n \hat{\mathbf{X}}_n^{\#} \tag{8.36}$$

Notons Z la matrice de dimension $L \times (L+1)$, partie gauche de l'équation (8.36). Les éléments z_{ij} de Z s'écrivent :

$$z_{ij} = \sum_{n=1}^{N} \sum_{r=1}^{R} \delta_{s_r}(n) \cdot \left(\sum_{p=1}^{L} v_{ip}^{s_r} \cdot m_p^{s_r} \right) \cdot \hat{x}_{n,j} \qquad \text{avec} \quad \begin{matrix} 1 \leq i \leq L \\ 1 \leq j \leq (L+1) \end{matrix} \tag{8.37}$$

Les matrices de précision Υ^{s_r} étant diagonales, l'équation (8.37) se simplifie :

$$z_{ij} = \sum_{n=1}^{N} \sum_{r=1}^{R} \delta_{s_r}(n) \cdot v_{ii}^{s_r} \cdot m_i^{s_r} \cdot \hat{x}_{n,j} \tag{8.38}$$

Notons Y la matrice de dimension $L \times (L+1)$, partie droite de l'équation (8.36). Les éléments y_{ij} de Y s'écrivent :

$$y_{ij} = \sum_{n=1}^{N} \sum_{r=1}^{R} \delta_{s_r}(n) \cdot \left(\sum_{q=1}^{L+1} \sum_{p=1}^{L} v_{ip}^{s_r} \cdot w_{pq}^{s} \cdot \hat{x}_{n,q} \right) \cdot \hat{x}_{n,j} \qquad \text{avec} \quad \begin{matrix} 1 \leq i \leq L \\ 1 \leq j \leq L+1 \end{matrix} \tag{8.39}$$

Les matrices de précision Υ^{s_r} étant diagonales, l'équation (8.39) devient :

$$y_{ij} = \sum_{q=1}^{L+1} w_{iq}^{s} \cdot \left(\sum_{n=1}^{N} \sum_{r=1}^{R} \delta_{s_r}(n) \cdot v_{ii}^{s_r} \cdot \hat{x}_{n,j} \cdot \hat{x}_{n,q} \right) \tag{8.40}$$

Soit G^i la matrice $(L+1) \times (L+1)$, dont les termes g_{jq}^i sont définis par :

$$g_{jq}^{i} = \sum_{n=1}^{N} \sum_{r=1}^{R} \delta_{s_r}(n) \cdot v_{ii}^{s_r} \cdot \hat{x}_{n,j} \cdot x_{n,q} \qquad \text{avec} \quad \begin{matrix} 1 \leq i \leq L \\ 1 \leq j, q \leq (L+1) \end{matrix} \tag{8.41}$$

Encore une fois, l'équation (8.36) peut se réécrire en utilisant (8.38), (8.40) et (8.39), sous la forme de L systèmes de $(L+1)$ équations linéaires à $(L+1)$ inconnues (cf. (8.22)). La matrice $\mathbf{W}^s$ est calculée ligne par ligne en résolvant le système de $(L+1)$ équations (8.23).

4.2 STMs avec mixtures

L'extension du calcul à des STMs comportant plusieurs mixtures est immédiate. L'équation (8.36) devient :

$$\sum_{t_k \in \mathbb{T}_{s_r}} \sum_{n=1}^{N} \sum_{r=1}^{R} \delta_{s_r}(n) Pr(t_k | \mathbf{X}_n, \Phi^{s_r}) \Sigma_k^{s_r -1} \mathbf{m}_k^{s_r} \hat{\mathbf{X}}_n^{\#} =$$

$$\sum_{t_k \in \mathbb{T}_{s_r}} \sum_{n=1}^{N} \sum_{r=1}^{R} \delta_{s_r}(n) Pr(t_k | \mathbf{X}_n, \Phi^{s_r}) \Sigma_k^{s_r -1} \overline{\mathbf{W}}^{s} \hat{\mathbf{X}}_n \hat{\mathbf{X}}_n^{\#} \tag{8.42}$$

Les éléments z_{ij} de la matrice Z sont :

$$z_{ij} = \sum_{t_k \in \mathbb{T}_{s_r}} \sum_{n=1}^{N} \sum_{r=1}^{R} \delta_{s_r}(n) \cdot Pr(t_k | \mathbf{X}_n, \Phi^{s_r}) \cdot v_{k,ii}^{s_r} \cdot m_{k,i}^{s_r} \cdot \hat{x}_{n,j} \tag{8.43}$$

La matrice Y s'écrit :

$$y_{ij} = \sum_{q=1}^{L+1} w_{iq}^{s} \cdot \left(\sum_{t_k \in \mathbb{T}_{s_r}} \sum_{n=1}^{N} \sum_{r=1}^{R} \delta_{s_r}(n) \cdot Pr(t_k|\mathbf{X}_n, \Phi^{s_r}) \cdot v_{k,ii}^{s_r} \cdot \hat{x}_{n,j} \cdot \hat{x}_{n,q} \right) \tag{8.44}$$

Les éléments de la matrice $\mathbf{G}^i$ sont :

$$g_{jq}^{i} = \sum_{t_k \in \mathbb{T}_{s_r}} \sum_{n=1}^{N} \sum_{r=1}^{R} \delta_{s_r}(n) \cdot Pr(t_k|\mathbf{X}_n, \Phi^{s_r}) \cdot v_{k,ii}^{s_r} \cdot \hat{x}_{n,j} \cdot \hat{x}_{n,q} \tag{8.45}$$

Comme précédemment, à chaque itération de EM, une estimation de la matrice $\mathbf{W}^s$ est obtenue ligne par ligne en résolvant le système de $(L+1)$ équations (8.23).

5 Réduction du nombre de paramètres des transformations

Le nombre de paramètres définissant les transformations W doit être inférieur au nombre total de paramètres des STMs. Dans le cas contraire, il est préférable de réentraîner directement les STMs plutôt qu'estimer les transformations. Lorsque les matrices de transformations sont pleines, le nombre total de paramètres devient rapidement prohibitif, et il devient nécessaire d'imposer des contraintes sur la nature des matrices de transformations. En effet :

Nombre de paramètres des STMs :
Soit H le nombre total de STMs, et M_{moy} le nombre moyen de mixtures par STM. Comme les matrices de covariances $\mathbf{\Sigma}_k^s$ sont diagonales, le nombre total de paramètres (N_{STM}) associés aux STMs est : $N_{STM} = H \times M_{moy} \times (2L+1)$.

Nombre de paramètres pour les transformations linéaires :
Soit C le nombre total de matrices de transformation W associées aux H modèles STMs. Le nombre total de paramètres (N_W) pour la transformation linéaire est donc : $N_W = C \times L \times (L+1)$.

En pratique, on a $H \approx 35$, $M_{moy} \approx 4$, $L = 65$ ($Q = 5$ et $D = 13$), et $1 \leq C \leq H$. Pour avoir $N_W \ll N_{STM}$, il est nécessaire de réduire le nombre de paramètres de la transformation W.

La matrice de régression transforme un vecteur $\hat{\mathbf{X}}$ de deux façons. Tout d'abord, la dernière colonne de $\mathbf{W}^s$ provoque une translation du vecteur X. Ensuite, le reste de la matrice $\mathbf{W}^s$ effectue une rotation et une homothétie de X. Notons $\mathbf{W}^{s\prime}$ la matrice carrée $L \times L$ correspondant à la matrice $\mathbf{W}^s$ privée de sa dernière colonne. Plusieurs stratégies sont envisageables pour réduire le nombre de paramètres de la transformation : suppression de la translation, utilisation de formes particulières pour la matrice $\mathbf{W}^{s\prime}$.

En forçant à zéro la dernière colonne de la matrice $\mathbf{W}^s$, la translation est supprimée et le nombre de paramètres est diminué de L. Cette réduction n'est cependant pas suffisante et doit être combinée avec des contraintes sur la nature de $\mathbf{W}^{s\prime}$.

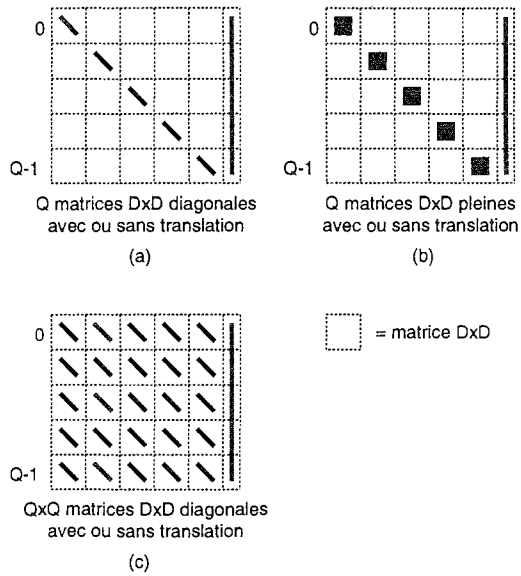

FIG. 8.3 - *Différentes structures des matrices de transformations. (a): indépendance des coefficients cepstraux de des états d'un STM. (b): dépendance des coefficients cepstraux et indépendance des états d'un STM. (c): indépendance des coefficients et dépendance des états d'un STM.*

Différentes configurations quant à la nature de la matrice de transformation sont représentées fig. 8.3, selon que l'on considère que les états d'un STM et les coefficients de l'espace de paramètres sont indépendants entre-eux ou non. Dans le cas (a), les différents états du STM sont indépendants, ainsi que les composantes des vecteurs de paramètres associés à chaque état. La matrice $\mathbf{W}^{s'}$ devient diagonale et le nombre de paramètres N' associés à $\mathbf{W}^{s'}$ est simplement L. Dans le cas (b), les différents états du STMs sont indépendants, mais on suppose que les coefficients des vecteurs de paramètres sont corrélés entre-eux. On obtient alors Q matrices pleines de dimension $D \times D$. Enfin, dans le cas (c), les états des STMs sont dépendants entre-eux, mais les coefficients de l'espace de paramètres sont indépendants. La matrice de transformation comporte $Q \times Q$ matrices diagonales de dimension $D \times D$.

6 Conclusion

Dans ce chapitre, nous avons présenté une approche permettant de compenser des modèles stochastiques de trajectoires, pour prendre en compte les variations entre les conditions d'apprentissage et les conditions de test du système VINICS.

La compensation est déterminée à partir d'un corpus d'adaptation de faible volume, en respectant un critère objectif d'estimation qui maximise la probabilité d'observer le corpus d'adaptation. La compensation mise en œuvre consiste à appliquer des transformations linéaires sur les vecteurs moyennes de STMs, ou sur les données bruitées de test.

La seule hypothèse utilisée suppose que les différences entre les environnements de test

et d'apprentissage peuvent être modélisées par des transformations linéaires par morceaux des espaces de paramètres. Nous n'utilisons pas de connaissances spécifiques à l'origine des perturbations entre conditions de test et d'apprentissage. Une telle approche est donc potentiellement applicable pour l'adaptation au locuteur, au bruit, ou encore aux changements de microphones. Les transformations sont spécifiques à chaque symbole, mais afin d'éviter une estimation non robuste, une même transformation peut être associée à différents symboles. On retrouve alors le compromis habituel entre la précision des transformations et la robustesse des estimateurs.

Cette approche présente néanmoins l'inconvénient de nécessiter l'utilisation d'un corpus d'adaptation étiqueté, et ne peut donc pas être directement appliquée en mode non supervisé. Les différentes expériences réalisées à partir de cette approche en reconnaissance de la parole continue dans le bruit sont exposées au chapitre 9.

Chapitre 9

Expériences et résultats

1 Conditions expérimentales

1.1 Description des corpus de parole et bruit

Les méthodes proposées partie III sont évaluées sur une tâche de reconnaissance de parole continue, en mode monolocuteur. L'évaluation est indépendante de la tâche d'apprentissage, ce qui signifie que le nombre de mots communs au vocabulaire d'apprentissage et de test est très faible. Le texte d'apprentissage est constitué de 79 phrases, riches au niveau phonétique. Le texte du corpus de test comporte 241 phrases issues d'une application de dictée de compte-rendus d'opérations de maintenance d'une centrale nucléaire, soit un total de 1482 mots.

Les corpus de test et d'apprentissage ont été enregistrés au laboratoire par 4 locuteurs masculins, en utilisant un microphone casque SHURE SM10A.

Les signaux ont été échantillonnés à 16 kHz. Une analyse cepstrale en échelle Mel a été effectuée toutes les 10 ms dans des fenêtres de Hamming de 25.6 ms, et 13 coefficients MFCC ont été calculés dans chaque fenêtre d'analyse.

Différents bruits (cf. fig. 9.1) ont été ajoutés aux signaux de parole dans le domaine temporel, pour des rapport signal-à-bruit variant de 0 dB à 36 dB, par pas de 6 dB :

1. un bruit blanc Gaussien, obtenu sous forme numérique à partir d'un générateur aléatoire ;
2. un bruit de sèche-cheveux ;
3. un bruit d'hélicoptère Lynx, issu de la base de données NOISEX [Varga *et al.*, 1992] ;
4. un bruit d'avion de chasse F16, également issu de NOISEX ;
5. un bruit enregistré à l'intérieur d'un autobus en déplacement.

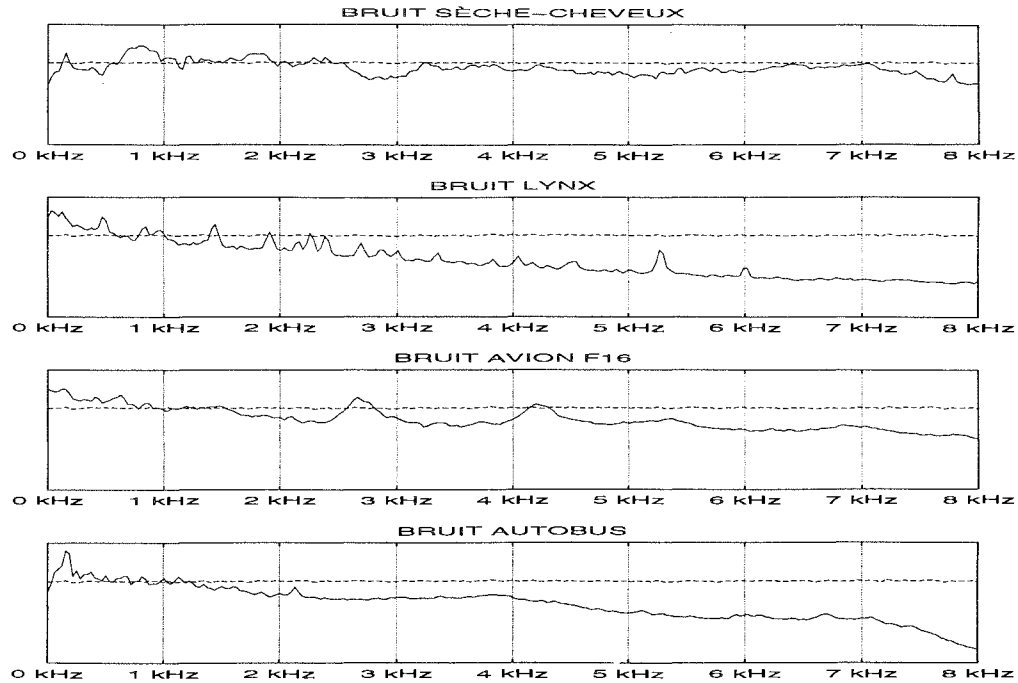

FIG. 9.1 - *Estimation des densités spectrales de puissance des différents bruits. Le bruit Gaussien est représenté en traits interrompus sur chaque figure.*

1.2 Modélisation acoustique

Dans toute cette série d'expériences, 32 modèles de phones indépendants du contexte ($H = 32$) sont utilisés lors de la reconnaissance, dont un modèle de silence. Dans toutes les configurations, les modèles de parole propre sont construits en utilisant l'algorithme EM (cf. partie II, page 69).

La segmentation initiale du corpus d'apprentissage est obtenue par une procédure d'alignement temporel automatique [Gong et Haton, 1991; Gong *et al.*, 1992; Gong et Haton, 1993]. Le corpus d'apprentissage de chaque locuteur comporte 2352 trajectoires de parole, soit en moyenne 70 trajectoires pour chaque modèle de phone.

Le nombre d'états des STMs est identique pour tous les modèles, et est fixé à 5 ($Q = 5$). Les matrices de covariance associées à chaque état des modèles sont diagonales. Le nombre maximal de composantes de trajectoire dans chaque mélange est fixé à 8 ($\text{card}(\mathbb{T}_s) \leq 8$). En définitive, le nombre moyen de densités Gaussiennes de dimension D ($D = 13$) est environ :

$$Q \times E\{\text{card}(\mathbb{T}_s)\} \times H \approx 620.$$

Enfin, dans toutes nos expériences, la probabilité *a posteriori* est utilisée pour normaliser le score $\mu(s, n, d)$ requis par l'étape de décodage, ce qui signifie que l'équation (5.2) est effectivement appliquée, contrairement à une configuration de test généralement utilisée qui supprime le dénominateur de l'équation (5.2) dans le calcul de $\mu(s, n, d)$ [Gong, 1994].

De la même façon, les paramètres heuristiques λ et γ, utilisés pour pondérer respectivement la *pdf* de la durée $p_{\tilde{d}|\tilde{s}}(d|s)$ et la distribution *a priori* des symboles $Pr(\tilde{s} = s)$, ont été fixés une fois pour toute. Ces paramètres n'ont jamais été modifiés par la suite pour tenter d'améliorer les performances dans une configuration de test donnée.

1.3 Décodage des phrases

Le vocabulaire de la grammaire de l'application comporte 1011 mots, et la perplexité en mode *word-pair* de l'automate à états finis représentant les contraintes syntaxiques est environ 28.

L'équation (5.12) est utilisée pour la normalisation de la probabilité de chaque phrase. Le système VINICS effectue une recherche en largeur, et fournit les N meilleures phrases comme résultat de reconnaissance. Pour l'évaluation des performances, seule la meilleure phrase est prise en considération. Dans toutes nos expériences, au maximum 1000 branches de recherche sont développées en parallèle.

L'évaluation des taux de reconnaissance s'effectue avec l'algorithme fourni dans la boîte à outils HTK [Young, 1992b]. Si Cor, Del, Ins, Sub et W désignent respectivement le nombre de mots correctement reconnus, de suppressions, d'insertions, de substitutions, et le nombre total de mots à reconnaître, Cor est défini par :

$$Cor = W - Del - Sub \tag{9.1}$$

Le pourcentage de mots correctement reconnus est :

$$Corr = \frac{Cor}{W} \times 100\% \tag{9.2}$$

Dans tous nos tableaux et courbes de résultats, le taux de reconnaissance, exprimé en pourcent, désigne la précision du système, définie par :

$$Acc = \frac{Cor - Ins}{W} \times 100\% \tag{9.3}$$

Nous faisons également apparaître sur nos courbes l'intervalle de confiance à 95%, calculé selon la méthode présentée dans [Montacié et Chollet, 1988]. Les taux de reconnaissance de deux méthodes sont déclarés comme significativement différents pour une configuration de test donnée, lorsque leurs intervalles de confiance sont disjoints.

2 Configuration des différentes approches

2.1 Configuration pour la combinaison stochastique de modèles

La combinaison stochastique de modèles consiste à générer différents STMs hybrides de parole bruitée à partir des STMs de parole propre et d'un HMM de bruit. La combinaison

est effectuée pour chacun des 5 types de bruit (Gaussien, Lynx, F16, Sèche-cheveux, Autobus) et pour chaque rapport signal-à-bruit (0 dB à 36 dB), ce qui représente un total de 35 configurations de test pour chaque locuteur.

Dans toutes les configurations, le HMM de bruit est construit à partir de 3 secondes de bruit. Étant donné que les niveaux moyens d'énergie de la parole propre d'apprentissage et de la parole propre utilisée pour bruiter les phrases de test sont identiques, le coefficient g des équations (6.16) et (6.17) est positionné à 1.

Dans un premier temps, nous étudions l'influence du nombre d'états du HMM de bruit sur les taux de reconnaissance. Ensuite, nous discutons de la validité de l'hypothèse de log-normalité, utilisée afin de calculer sous forme close la *pdf* du signal de parole bruitée.

2.1.1 Influence du nombre d'états du HMM de bruit

La surcharge de calculs induite par l'utilisation d'un nombre d'états du HMM de bruit supérieur à 1, nous a conduit à évaluer l'influence du nombre d'états sur les performances de la combinaison de modèles, sur une tâche de reconnaissance moins complexe que la précédente.

Pour cette expérience, le corpus d'apprentissage est constitué d'une répétition de 130 phrases choisies aléatoirement dans le corpus TIMIT, et prononcées par un locuteur américain. À partir de ce corpus, 47 modèles de phones indépendants du contexte sont construits en utilisant l'algorithme EM. Le corpus de test regroupe une répétition de 206 mots isolés, également issus du vocabulaire de TIMIT, et prononcés par le même locuteur. Aucune opération particulière n'a été mise en œuvre pour maximiser le taux de recouvrement entre le vocabulaire de test et d'apprentissage.

L'évaluation porte sur la reconnaissance de ces 206 mots isolés, en présence du bruit de sèche-cheveux, d'autobus et d'avion F16, pour des rapports signal-à-bruit variant de 0 dB à 40 dB par pas de 10 dB.

La comparaison porte sur les taux de reconnaissance en fonction du nombre d'états du HMM de bruit. Deux configurations sont testées. Dans la première, le modèle de bruit ne comporte qu'un seul état, les STMs de parole bruitée conservent alors la structure des STMs de parole propre initiaux. Le système VINICS peut être directement utilisé pour la reconnaissance. Dans la seconde, le HMM de bruit est ergodique et comporte 2 états ; les STMs de parole bruitée sont alors des modèles hybrides, et il est nécessaire d'utiliser une extension de VINICS pour prendre en compte la modification de l'algorithme de décodage présenté chap. 6, page 91. Les courbes représentées fig. 9.2 indiquent l'évolution des taux de reconnaissance en fonction du rapport signal-à-bruit pour les différents types de bruits, selon que le modèle de bruit comporte 1 ou 2 états. Il apparaît qu'aucune différence significative ne peut être observée pour les différents types de bruits testés, lorsque le nombre d'états du HMM de bruit varie de 1 à 2. Il est probable que les différents bruits testés ne comportaient pas une non-stationnarité suffisamment marquée pour que leur modélisation par 1 ou 2 états provoque une variation significative des performances. En conséquence, et étant donné la surcharge de calculs provoquée par l'utilisation d'un nombre d'états du HMM de bruit supérieur à 1, nous choisissons d'effectuer tous nos tests ultérieurs de combinaison de modèles avec un modèle

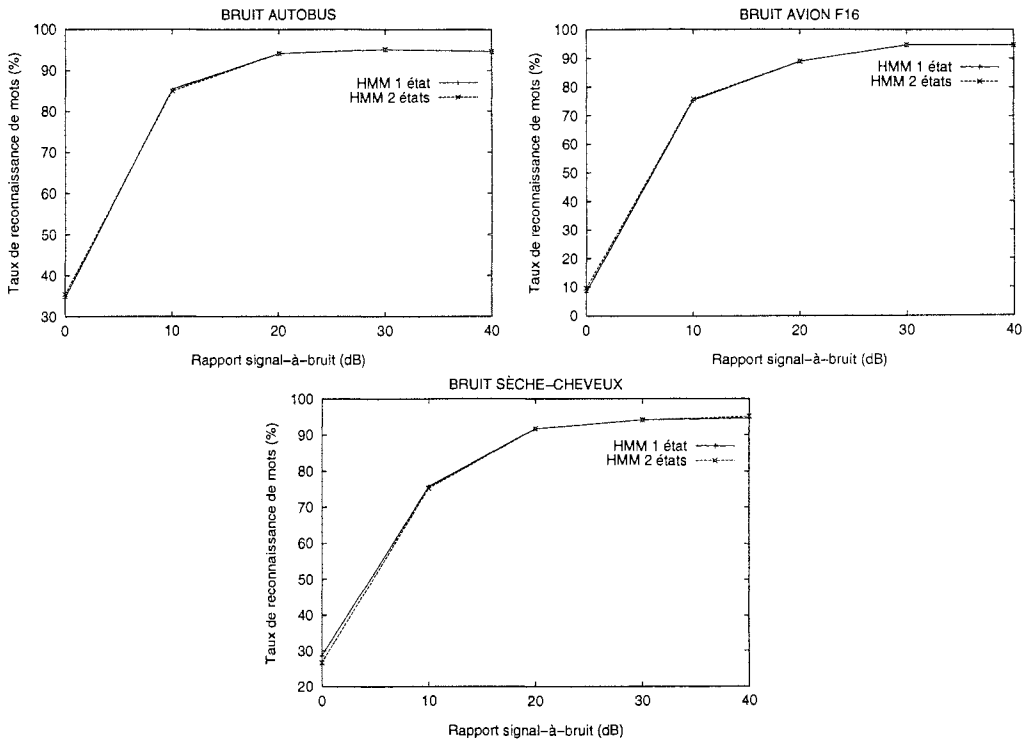

FIG. 9.2 - *Influence du nombre d'états du HMM de bruit sur le taux de reconnaissance de 206 mots isolés.*

de bruit ne comportant qu'un seul état.

2.1.2 Validité de l'hypothèse de log-normalité

L'hypothèse utilisée pour déterminer sous forme close l'expression de la *pdf* du cepstre de la parole bruitée considère que la somme de 2 variables aléatoires de distribution log-normale est également une variable aléatoire log-normale. Sous une telle hypothèse, le cepstre de la parole bruitée suit une distribution normale. Or, d'après [Openshaw et Mason, 1994], il apparaît que la distribution de la parole bruitée dans le domaine cepstral ne respecte plus une distribution Gaussienne (donc unimodale), mais devient au contraire bimodale au fur et à mesure que le SNR diminue. Afin de reproduire l'expérience de Openshaw et Mason, nous avons bruité les 79 phrases du corpus d'apprentissage d'un locuteur avec différents niveaux de bruit blanc Gaussien, puis paramétré l'ensemble du corpus dans l'espace MFCC. L'histogramme de chaque coefficient cepstral a ensuite été déterminé, et nous présentons fig. 9.3 la distribution du coefficient d'ordre 1. Notre expérience confirme que la distribution du cepstre de la parole devient bimodale en présence d'un bruit important.

Les approches de compensation de modèles, qui consistent à déplacer les moyennes et variances d'une loi sans remettre en cause sa nature, peuvent donc sembler inappropriées car elles ne permettent pas de prendre en compte l'évolution d'une distribution telle que celle

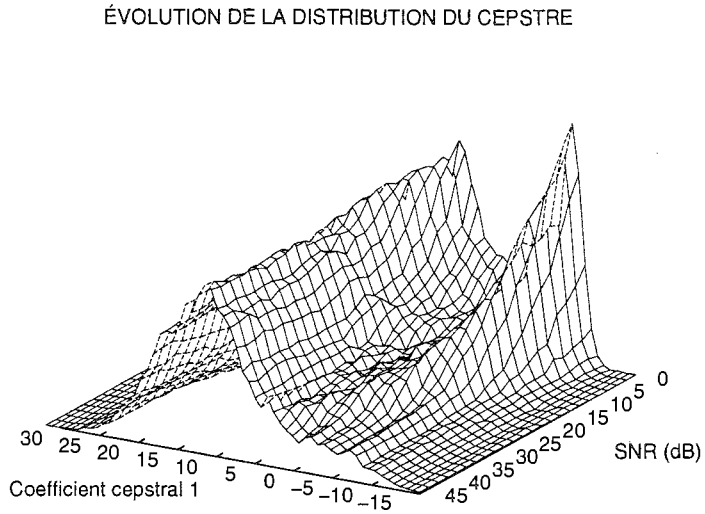

FIG. 9.3 - *Distribution du coefficient cepstral d'ordre 1 en présence de différents niveaux de bruit blanc Gaussien. La distribution est calculée à partir des 79 phrases d'apprentissage prononcées par 1 locuteur.*

de la fig. 9.3. En effet, supposons qu'en absence de bruit, la distribution de la fig. 9.3 soit modélisée par une loi normale. Lorsque le SNR diminue, déplacer la moyenne et la variance de cette loi ne permet pas de modéliser correctement la distribution.

Cependant, il faut noter que l'allure de la distribution de la fig. 9.3 ne caractérise pas celle de chaque symbole. Pour mettre en évidence cette hypothèse, nous avons réparti, à l'aide d'un algorithme de classification automatique, l'ensemble des observations des symboles phonétiques des 79 phrases en 2 classes : l'une pour les trajectoires de faible énergie, l'autre pour les trajectoires de forte énergie. Les histogrammes spécifiques à chaque classe ont ensuite été déterminés pour chaque coefficient cepstral, et nous représentons fig. 9.4 les distributions des 6 premiers coefficients. Il apparaît que les distributions du cepstre des symboles de forte énergie sont peu modifiées par le bruit, alors que le bruit provoque une réduction de la variance des coefficients cepstraux des symboles de faible énergie, la distribution devenant plus étroite et abrupte, ce qui est conforme aux conclusions de [Openshaw et Mason, 1994]. La tendance à la bimodalité qui se manifestait fig. 9.3 provient donc du regroupement de tous les symboles pour calculer les distributions, le mode qui apparaît provenant de la contribution des symboles de faible énergie. Par la suite, en traçant les distributions des coefficients du cepstre, symbole par symbole, nous n'avons jamais observé une tendance à la bimodalité telle que celle de la fig. 9.3. Nous considérons donc que si les coefficients du cepstre d'un symbole de parole propre suivent une *pdf* unimodale, cette *pdf* reste unimodale en présence de bruit, et qu'il est possible d'adapter cette *pdf* au bruit en modifiant ses paramètres. Le calcul des paramètres par la combinaison de modèles est cependant sous optimal, et nous montrons par l'expérience qui suit, qu'en partant des mêmes modèles de parole propre et de bruit, il est

possible de construire des modèles de parole bruitée par génération de données bruitées et réestimation des modèles, qui permettent d'obtenir de meilleurs taux de reconnaissance que ceux provenant de la combinaison des modèles.

À partir des STMs de parole propre, nous pouvons générer artificiellement un grand nombre de trajectoires de parole qui respectent les modèles. De la même façon, des vecteurs de bruit peuvent être artificiellement générés à partir du modèle de bruit. Étant donné ces trajectoires de parole propre et ces observations de bruit, obtenues dans le domaine cepstral, il est possible de calculer un grand nombre de trajectoires de parole bruitée dans le domaine cepstral, la combinaison des observations étant effectuée après un va-et-vient vers le domaine spectral linéaire où parole et bruit s'ajoutent. En utilisant les observations de parole bruitée ainsi synthétisées, il est alors possible d'entraîner complètement des STMs de parole bruitée par application de l'algorithme EM.

Les performances obtenues par cette réestimation à partir des données artificielles sont comparées à celles de la combinaison de modèles sur la tâche de reconnaissance décrite au paragraphe 1, en présence d'un bruit blanc Gaussien. Il faut noter que les 2 approches utilisent exactement les mêmes informations, qui sont celles fournies par les modèles de parole propre et le modèle de bruit. Les différences se situent uniquement dans la façon de construire les modèles de parole bruitée. Dans un cas, seules les moyennes et variances sont modifiées (combinaison des modèles) ; dans l'autre cas, les modèles entiers sont réestimés. Les taux de reconnaissance des 2 expériences sont présentés fig. 9.5. Il apparaît que la réestimation complète donne des résultats équivalents à ceux de la combinaison de modèles lorsque le SNR est élevé. Par contre, lorsque le SNR est faible, la combinaison des modèles s'avère moins efficace que la réestimation à partir des données artificielles. Cette expérience confirme donc que l'hypothèse de log-normalité, qui conduit à modéliser la distribution de la parole bruitée dans le domaine de la densité spectrale de puissance par une loi log-normale dont la moyenne et la variance sont la somme des moyennes et variances des distributions de parole et de bruit, devient sous optimale lorsque le niveau de bruit est important. Cependant, étant donnée la quantité de données à générer, ainsi que la charge en calculs, la réestimation à partir de données synthétisées n'est pas une approche viable en pratique. Notre expérience n'avait pour seul objectif que de mettre en évidence la sous-optimalité de l'hypothèse de log-normalité en présence d'un bruit important. La combinaison de modèles s'avère au contraire très rapide, et ne nécessite que quelques secondes de calculs. Cette méthode est utilisée dans nos expériences ultérieures, et est comparée avec les autres approches développées. Ces résultats sont donnés paragraphe 3.

2.2 Configuration pour le filtrage par états

Le filtrage non linéaire par états se décompose en 2 étapes.

La première consiste à déterminer pour chaque état de chaque STM une table de filtrage dans le domaine cepstral, à partir des modèles STMs de parole propre et d'un modèle de bruit. Le modèle de bruit utilisé est le même que pour la combinaison de modèles, et est déterminé à partir des mêmes séquences de bruit. Les tables de compensation sont déterminées pour les 35 configurations de types et niveau de bruit. Les tables sont discrétisées en 256 points, et le calcul d'une table pour un locuteur s'effectue en environ 10 heures sur une machine

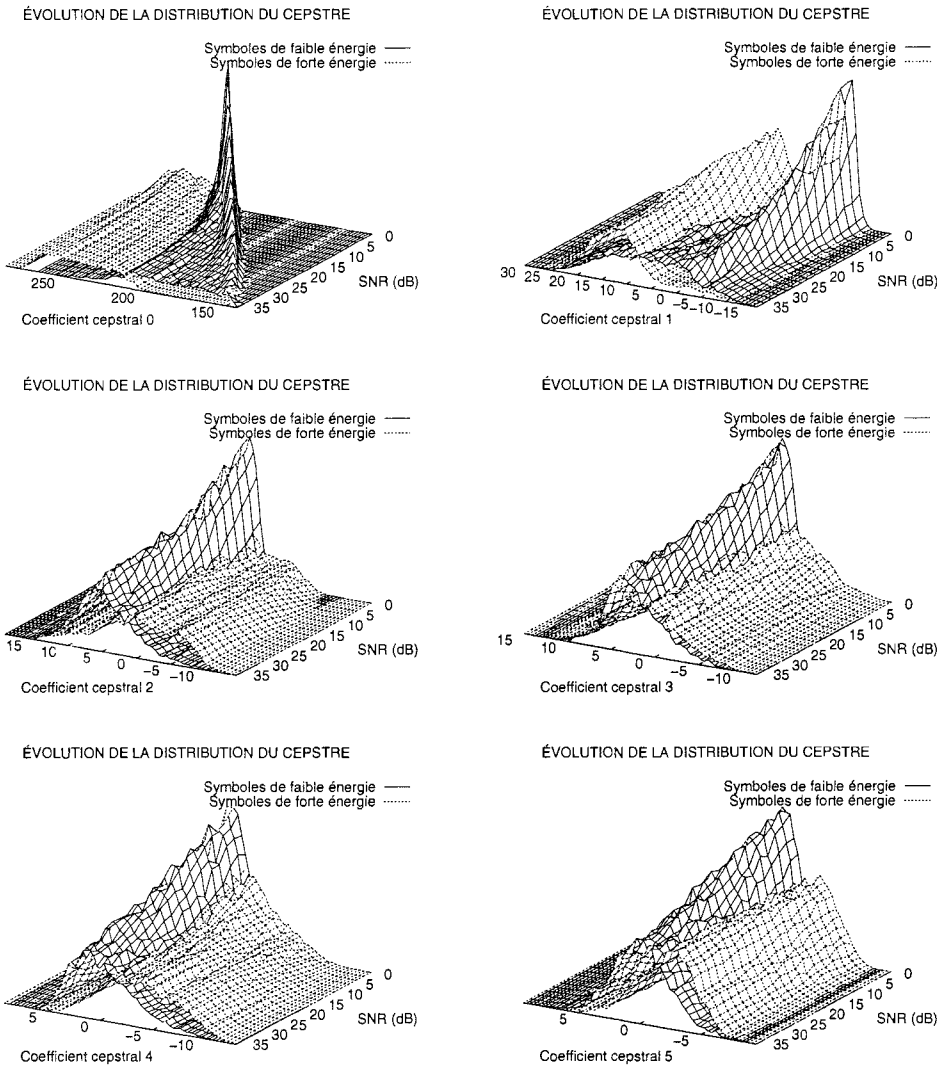

FIG. 9.4 - *Distributions des coefficients cepstraux 0 à 5 en présence de différents niveaux de bruit blanc Gaussien. Les distributions sont calculées à partir des 79 phrases d'apprentissage prononcées par 1 locuteur, après classification automatique de l'ensemble des symboles en 2 classes : 1 classe pour les symboles de forte énergie, 1 classe pour les symboles de faible énergie.*

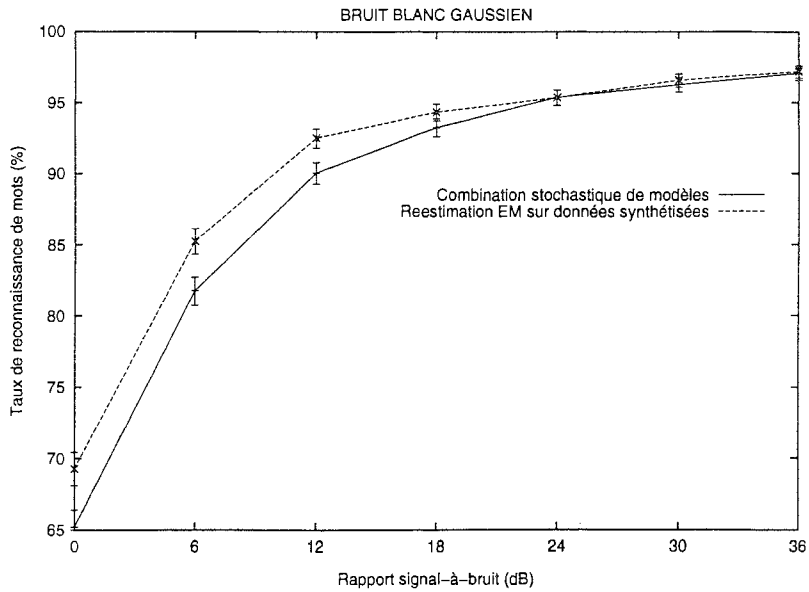

FIG. 9.5 - *Comparaison entre la combinaison stochastique de modèles et la réestimation des modèles par l'algorithme EM à partir de données de parole bruitée artificiellement générées. Bruit blanc Gaussien. Moyenne sur tous les locuteurs.*

type Sun Sparc 10/40. Des expériences non reportées ici ont montré qu'une approximation de chaque table de filtrage par une fonction polynomiale ne permet pas d'améliorer de façon significative les taux de reconnaissance. Aussi, les tables brutes ont été conservées pour la reconnaissance. À titre d'exemple, nous représentons fig. 9.6 les tables de compensation pour chaque composante du vecteur de cepstre, en l'état 0 du modèle du symbole /R/, en présence d'un bruit blanc Gaussien pour un SNR de 0 dB.

Les tables de compensation étant déterminées, celles-ci sont, dans une seconde étape, associées à chaque état de chaque STM. Le signal de parole bruitée est alors filtré pendant la reconnaissance, de façon spécifique à chaque état, lors du calcul de la probabilité d'une trajectoire bruitée $p(\mathbf{Y}|t_k, s)$, par application de la relation (7.11). Les taux de reconnaissance obtenus sont présentés paragraphe 3.

2.3 Configuration pour la régression linéaire

L'approche d'adaptation des modèles par régression linéaire nécessite d'utiliser un corpus d'adaptation étiqueté de parole bruitée. L'objectif de l'adaptation est de rechercher un ensemble de transformations linéaires, afin de transformer soit la parole bruitée de test, soit les modèles propres de référence. Les transformations sont déterminées selon un critère objectif consistant à maximiser la probabilité d'observer le corpus d'adaptation. Nous utilisons comme corpus d'adaptation un sous-ensemble du corpus d'apprentissage, constitué de 10 phrases, soit environ 20 secondes de parole. Aucune sélection des phrases du corpus d'adaptation n'a été effectuée dans le but de maximiser les taux de reconnaissance.

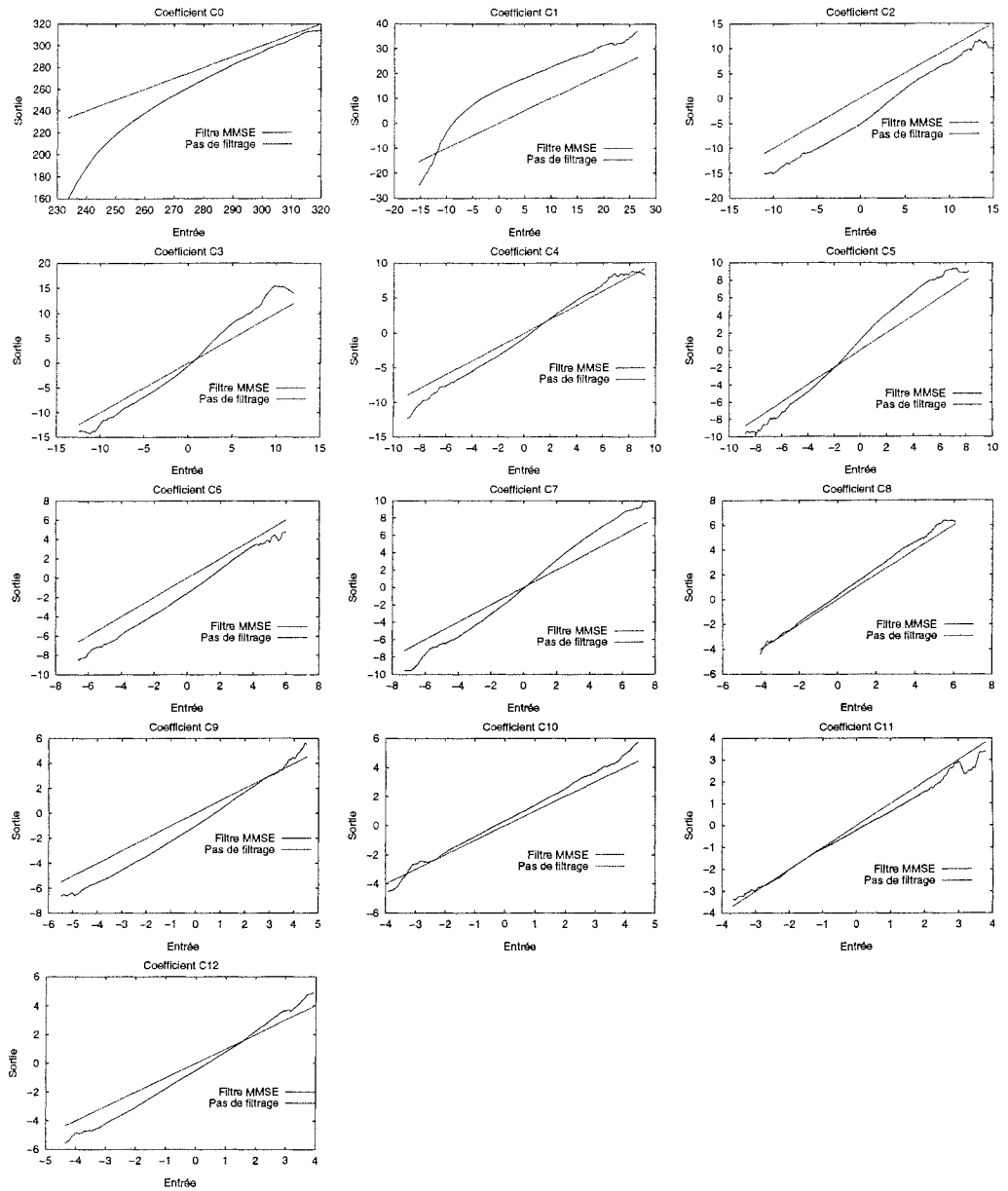

FIG. 9.6 - *Tables de filtrage MMSE pour les 13 coefficients cepstraux. Symbole /R/, état 0 du modèle, bruit blanc Gaussien, SNR 0 dB.*

Nous avons choisi d'utiliser 3 transformations linéaires au total. Cela signifie que l'ensemble des symboles est divisé en 3 classes, et qu'une transformation est associée à chacune d'entres-elles. La première classe est constituée du symbole silence et des 2 symboles correspondant aux silences qui précèdent les occlusives sourdes et sonores. La seconde classe regroupe l'ensemble des sons voyelles, et la dernière contient les symboles restant. Cette répartition a été fixée une fois pour toute, et aucune tentative d'optimisation des performances, par des modifications du nombre de classes ou de la répartition des symboles dans chaque classe, n'a été effectuée.

Comme précédemment, les transformations linéaires sont déterminées pour chacune des 35 configurations de bruit possible (5 types de bruit $\times$ 7 SNRs différents). Nous discutons au paragraphe suivant du choix de la structure des matrices de transformations.

2.3.1 Choix de la transformation de la régression linéaire

Nous avons noté chapitre 8, page 112, qu'il est nécessaire de trouver un compromis entre le nombre de paramètres à estimer et la précision des transformations à mettre en œuvre.

La modification du nombre des paramètres à estimer pour l'adaptation peut porter soit sur le nombre total de transformations (fixé à 3), soit sur la nature des matrices de transformations (matrices pleines, diagonales, avec ou sans translation, etc.). Étant donné le nombre de degrés de liberté, nous avons choisi uniquement 2 configurations de test quant à la structure des matrices de transformation. La première consiste à utiliser des matrices de transformation diagonales, sans la partie spécifique à la translation. La seconde configuration utilise une matrice diagonale avec la translation. Ces 2 configurations sont évaluées uniquement en présence du bruit blanc Gaussien, sur la tâche de reconnaissance décrite paragraphe 1, et portent sur la transformation des modèles de parole propre. Les résultats sont présentés fig. 9.7.

Il apparaît clairement que l'utilisation de la translation permet d'améliorer de façon significative les taux de reconnaissance, en particulier lorsque le niveau de bruit est important. La transformation mise en œuvre est donc une translation associée à une homothétie, et cette configuration est adoptée pour nos tests ultérieurs.

2.4 Configuration pour la transformation de base

La transformation de base est une méthode d'adaptation à l'environnement [Gong, 1993; Treurniet et Gong, 1994]. Étant donnée que cette approche a été développée au CRIN par Gong, les outils pour mettre en œuvre cette méthode étaient directement disponibles. Pour cette raison, nous avons également évalué cette approche sur notre application de test, afin de la comparer à nos méthodes.

La transformation de base est une méthode de transformation d'espace, dont l'objectif est de projeter l'espace spectral de test sur l'espace spectral de référence. Dans un espace donné, un vecteur de paramètre s'écrit comme une combinaison linéaire des éléments d'une base. La transformation consiste à remplacer les vecteurs de la base de parole bruitée par ceux de la base de parole propre. La correspondance entre les éléments des différentes bases provient d'un corpus d'adaptation, disponible à la fois en parole propre et bruitée. Le corpus

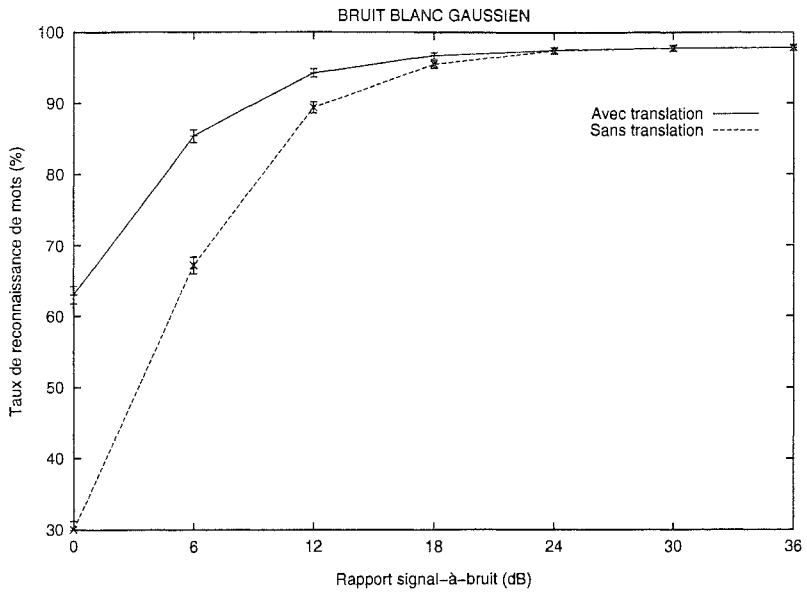

FIG. 9.7 - *Comparaison entre une adaptation par régression linéaire avec ou sans translation. Bruit blanc Gaussien. Moyenne sur tous les locuteurs.*

d'adaptation utilisé est le même que pour l'adaptation par régression linéaire, soit 20 secondes de parole au total. Pour plus de précision sur la transformation de base, le lecteur peut se reporter à l'annexe B.

La transformation de base existe sous deux configurations, qui se différencient par la façon de bruiter le corpus d'adaptation. Dans la configuration d'origine [Gong, 1993; Treurniet et Gong, 1994], le signal de bruit est ajouté au signal de parole propre du corpus d'adaptation, et le corpus ainsi bruité est paramétré. Une démarche plus récente consiste à moyenner des versions paramétrées du corpus d'adaptation bruité par différentes occurrences de bruits. La transformation de base n'utilise pas de critères mathématiques d'optimisation, et cette heuristique, proposée par Gong, permet de prendre en compte les effets «moyens» du bruit. En effet, dans le premier cas, les éléments de la base sont constitués à partir d'une occurrence du corpus d'adaptation bruité ; dans le second cas, les éléments de la base sont représentés par une moyenne de différentes occurrences bruitées du corpus d'adaptation. Nous comparons ces 2 variantes sur l'application de reconnaissance présentée paragraphe 1, pour les 35 configurations de types et niveaux de bruit. Les résultats sont disponibles fig. 9.8.

Il apparaît que pour les bruits blancs (bruit Gaussien et sèche-cheveux), la configuration avec moyenne de différentes occurrences bruitées du corpus d'adaptation permet d'obtenir de meilleurs taux de reconnaissance lorsque le niveau de bruit est important. Cette tendance se retrouve également pour les bruit d'avion F16 et d'hélicoptère Lynx. Par contre, en présence d'un bruit d'autobus, l'absence de moyenne s'avère plus efficace aux SNRs faibles. Cependant, pour tous les types de bruit, les deux configurations sont équivalentes lorsque le niveau du bruit est faible.

Il est également possible d'utiliser ces deux configurations du corpus d'adaptation pour

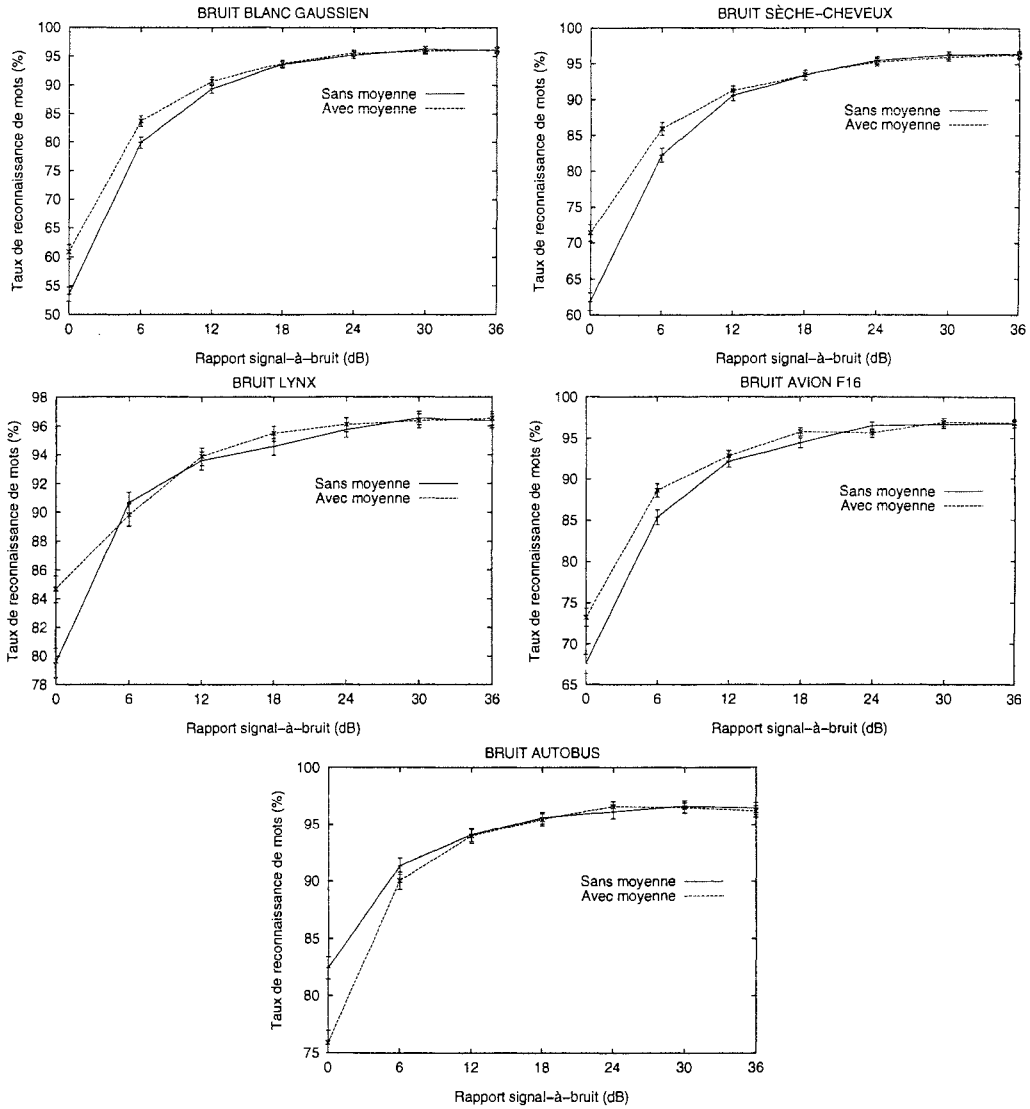

FIG. 9.8 - *Comparaison des performances de la transformation de base. Première configuration : les éléments de la base sont obtenus à partir une seule occurrence du corpus d'adaptation (Sans moyenne). Seconde configuration : les éléments de la base sont obtenus à partir d'une moyenne de différentes versions bruitées du corpus d'adaptation (Avec moyenne). Moyenne des taux de reconnaissance sur tous les locuteurs.*

l'adaptation par régression linéaire. Les taux de reconnaissance pour ces deux variantes sont présentés fig. 9.9. La transformation utilisée pour la régression linéaire est celle décrite paragraphe 2.3.1.

Contrairement à la transformation de base, il apparaît que le moyennage du corpus d'adaptation ne conduit pas à une amélioration sensible des performances. Dans le cas de la transformation de base, aucun critère probabiliste n'est employé ; le moyennage des occurrences bruitées du corpus d'adaptation permet alors de minimiser l'influence d'une réalisation d'un élément de la base fortement éloignée de la moyenne. Par contre, l'adaptation par régression linéaire se fonde sur un critère probabiliste du maximum de la vraisemblance, et le moyennage du corpus d'adaptation n'a pas d'influence sur les performances.

En définitive, nous choisissons d'utiliser pour la transformation de base, la configuration avec moyenne de plusieurs occurrences du corpus d'adaptation. Pour l'approche par transformation linéaire, nous conservons la configuration présentée paragraphe 2.3.1, c.-à-d. sans moyenne de différentes occurrences bruitées du corpus d'adaptation.

3 Comparaison des performances

Dans ce paragraphe, nous comparons les performances respectives de la combinaison de modèles, du filtrage non linéaire par états, de l'adaptation par régression linéaire et de la transformation de base sur l'application de reconnaissance de parole continue décrite paragraphe 1. Les configurations des différentes approches sont celles des paragraphes 2.1 à 2.4. Les tests sont effectués pour les 35 configurations de bruit déjà présentées (5 types de bruits et 7 rapports signal-à-bruit). Dans les 2 approches nécessitant l'estimation de modèles de bruit (filtrage par états et combinaison de modèles), nous rappelons que ces modèles sont estimés à partir du signal de bruit seul. En pratique, il serait nécessaire d'estimer ces modèles à partir des zones du signal où la parole est absente, afin de prendre en compte les variabilités à long terme du bruit. Pour cela, il faut disposer d'un détecteur d'activité vocale robuste au bruit. Nous n'abordons pas ici ce problème, et nous nous plaçons donc dans le cas où cette détection d'activité vocale est idéale.

Les résultats obtenus sont représentés fig. 9.10 à 9.12, pour chaque type de bruit. Nous indiquons également sur les courbes, les performances obtenues lorsque les modèles acoustiques sont construits en utilisant un corpus d'apprentissage bruité dans les mêmes conditions que celles du test (noté Apprentissage dans le bruit). Ces résultats sont également rapportés sous forme de tables dans l'annexe C.

Au regard des courbes, il apparaît que pour un rapport signal-à-bruit de 0 dB, l'apprentissage dans le bruit permet d'obtenir les meilleurs taux de reconnaissance. Une seule exception se produit pour le bruit de sèche-cheveux, qui correspond au bruit le plus pénalisant. Il faut cependant noter que la stratégie utilisée lors du décodage pour élaguer l'arbre de recherche conduit à couper prématurément certaines branches, lorsque le score associé est faible. Cette situation se produit plus particulièrement lorsque le bruit est important. Par conséquent, les taux de reconnaissance à 0 dB sont biaisés par ce phénomène, et doivent être interprétés avec précautions, notamment pour les bruits très pénalisants comme le sèche-cheveux, qui conduisent à des taux de reconnaissance très faibles à 0 dB. Les tendances sont néanmoins

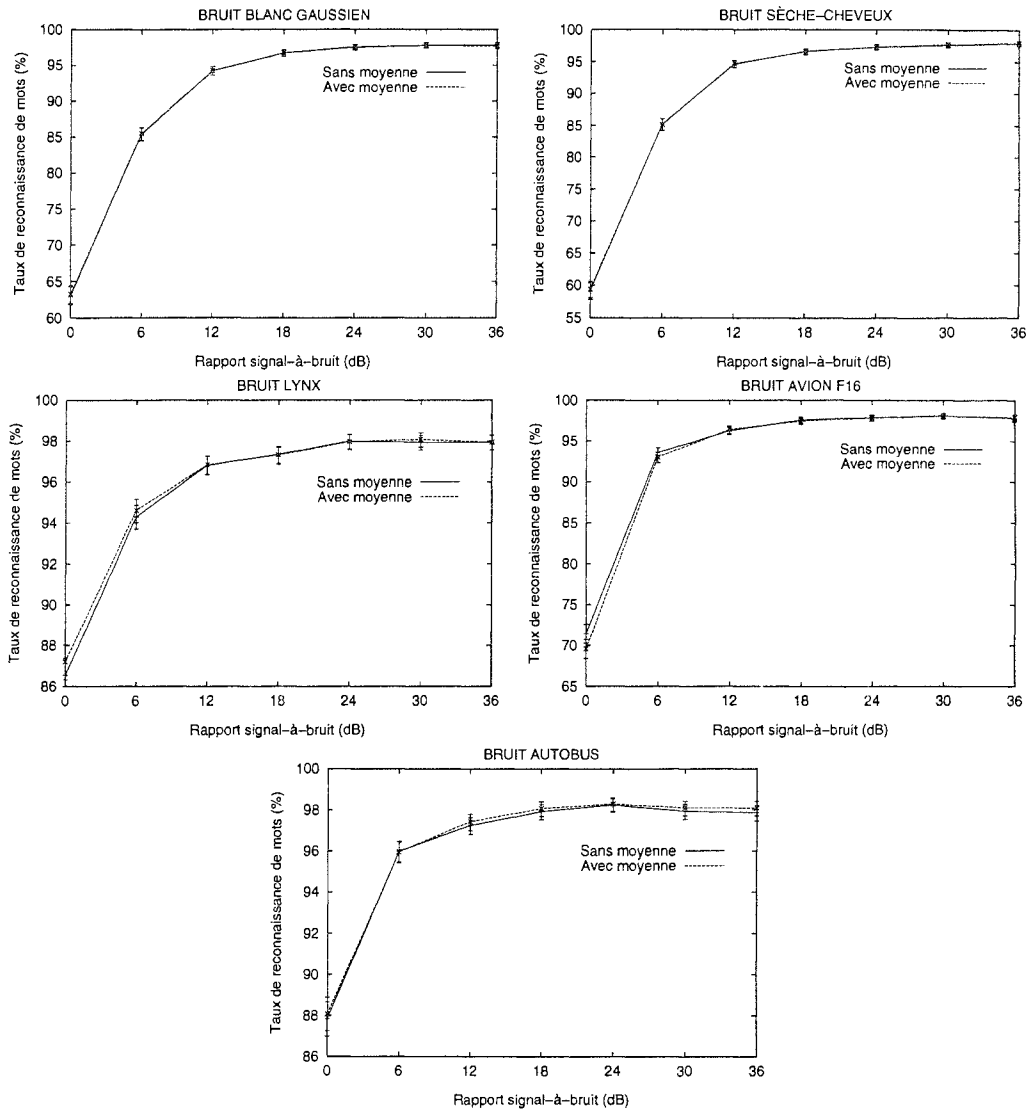

FIG. 9.9 - *Comparaison des performances de la régression linéaire. Première configuration : le corpus d'adaptation est constitué d'une seule occurrence bruitée des phrases d'adaptation (Sans moyenne). Seconde configuration : le corpus d'adaptation est obtenus à partir d'une moyenne de différentes versions bruitées des phrases d'adaptation (Avec moyenne). Moyenne sur tous les locuteurs.*

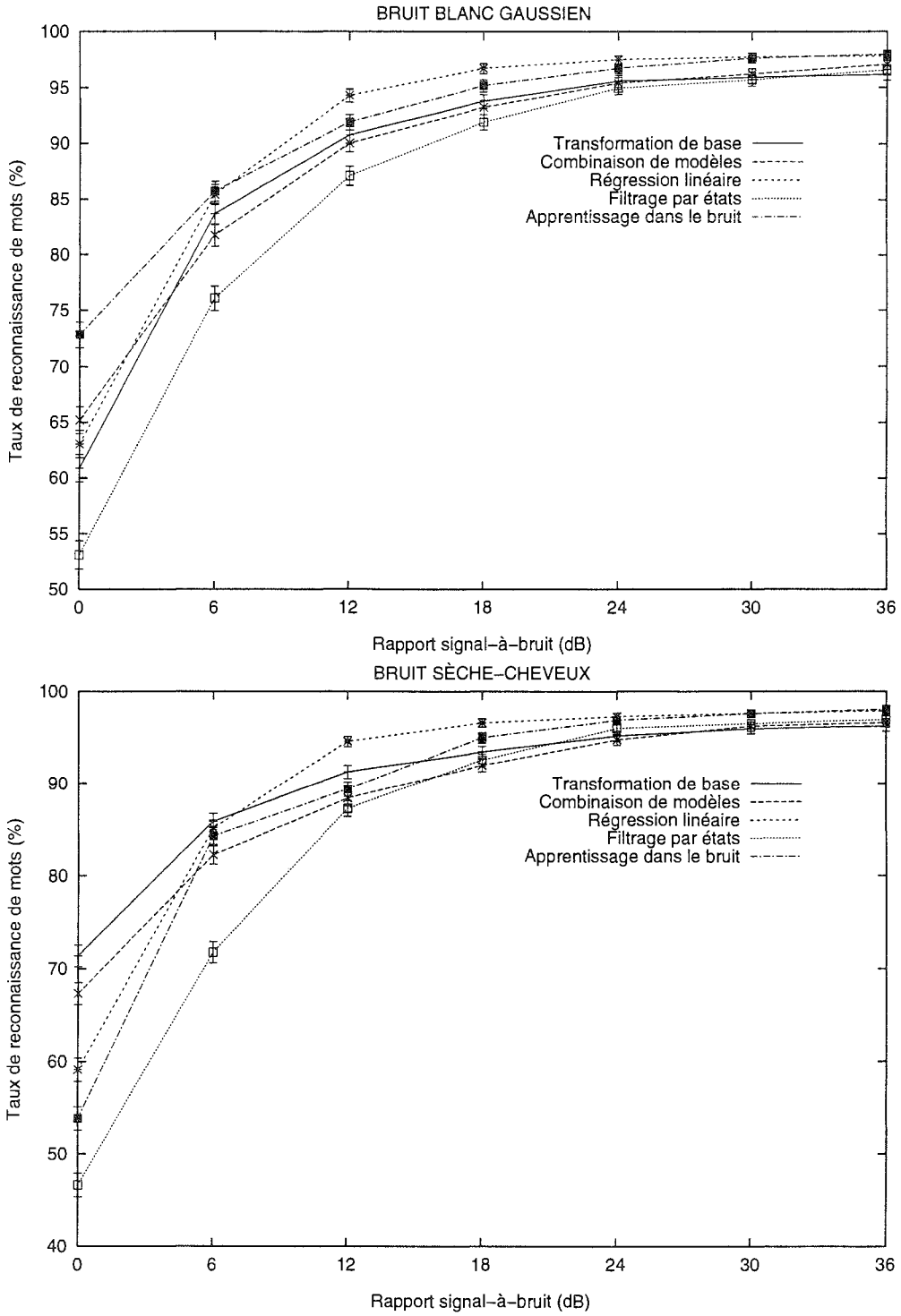

FIG. 9.10 - *Comparaison des performances des différentes approches. Bruit blanc et bruit de sèche-cheveux. Moyenne sur tous les locuteurs (cf. tables C.1 et C.2).*

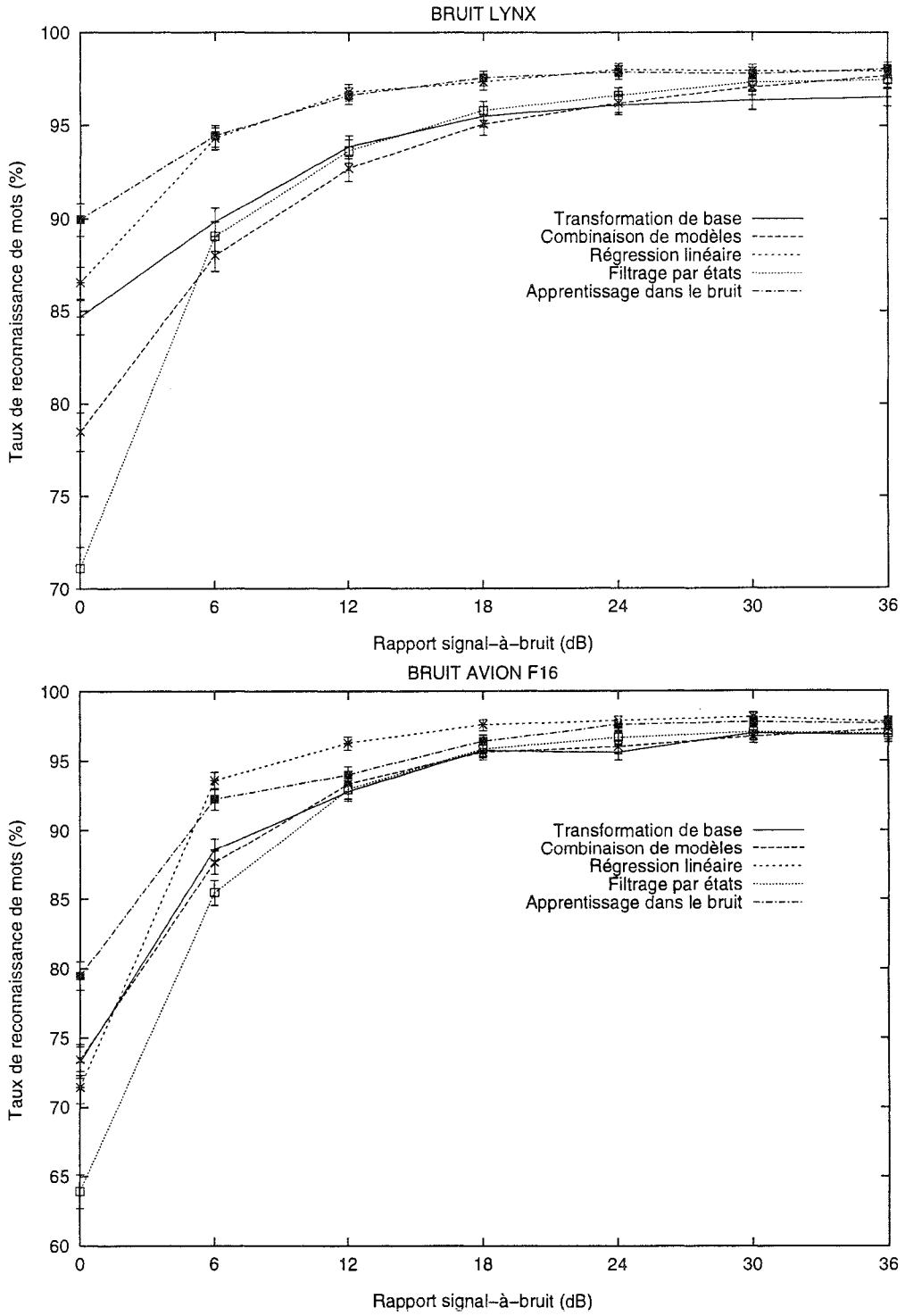

FIG. 9.11 - *Comparaison des performances des différentes approches. Bruit d'avion F16 et d'hélicoptère Lynx. Moyenne sur tous les locuteurs (cf. tables C.3 et C.4).*

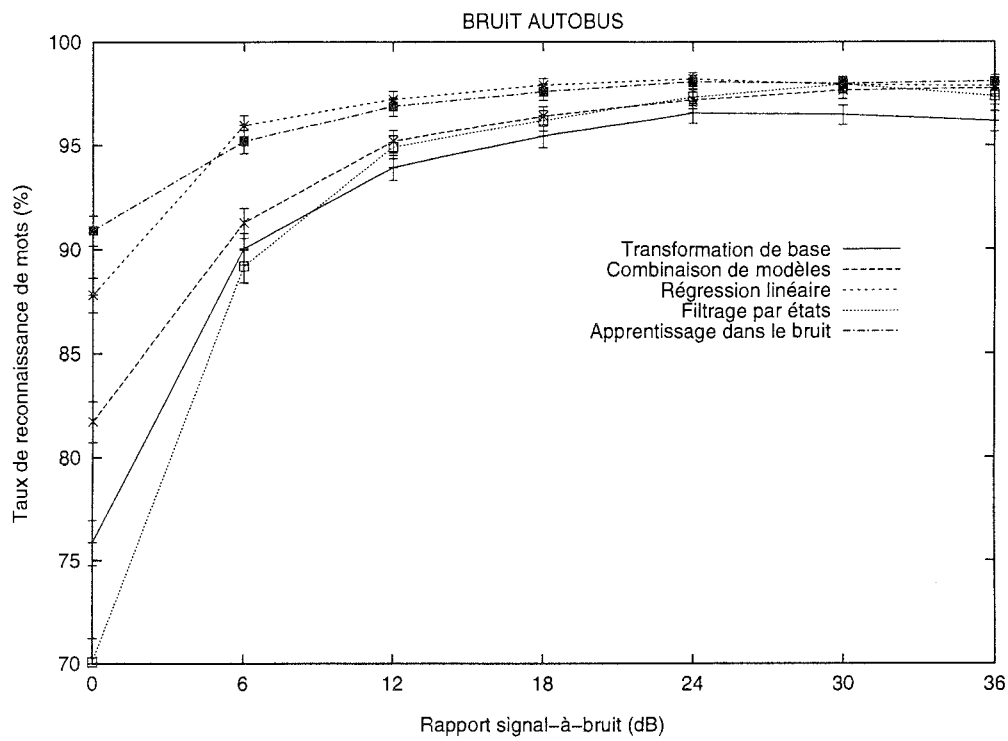

FIG. 9.12 - *Comparaison des performances des différentes approches. Bruit d'autobus. Moyenne sur tous les locuteurs (cf. table C.5).*

conservées, et on peut remarquer en particulier que le filtrage par états s'avère toujours moins performant que les autres approches à 0 dB.

De façon générale, pour tous les types de bruit, le filtrage par états correspond à la configuration la moins efficace lorsque le SNR est faible ($\leq$ 6 dB). En présence d'un SNR élevé ($\geq$ 24 dB), le filtrage par états et la combinaison de modèles sont équivalents, ce qui est un résultat attendu : le bruit étant faible, les modèles propres initiaux sont peu modifiés par le filtrage et par la combinaison avec le modèle de bruit (la densité spectrale de puissance du bruit perturbe peu la densité spectrale de puissance du signal de parole). Pour les SNRs intermédiaires, cette tendance persiste pour les bruits de sèche-cheveux, de Lynx, d'autobus et d'avion F16, mais n'est plus observée pour le bruit Gaussien où la combinaison de modèles s'avère plus efficace. Les inconvénients majeurs du filtrage par états sont d'une part l'utilisation de tables discrétisées de filtrage (256 points), qui nécessitent de plus une capacité de stockage importante, d'autre part le temps prohibitif de calcul des estimateurs. Pour ces raisons, cette méthode est inapplicable en pratique. Au contraire, la combinaison de modèles s'effectue en quelques secondes, dès que le modèle de bruit est disponible. De plus, pour les SNRs très élevés, les performances de la combinaison de modèles convergent vers celles de l'apprentissage dans le bruit. Il faut cependant garder à l'esprit que la combinaison est sous optimale pour les SNRs faibles, comme nous l'avons noté paragraphe 2.1.2.

Les performances de la transformation de base dépendent fortement du type et du niveau de bruit. Une tendance particulièrement nette apparaît aux SNRs élevés ($\geq$ 24 dB) où la méthode s'avère significativement moins efficace que la régression linéaire et l'apprentissage dans le bruit. Cela se distingue clairement en présence du bruit d'autobus. Avec le bruit de sèche-cheveux, la transformation de base est la méthode la plus performante pour les SNRs faibles ($\leq$ 6 dB). Pour les SNRs intermédiaires, les performances obtenues relativement aux autres approches sont très variables : équivalence par rapport au filtrage par états pour le bruit de Lynx et d'avion F16, significativement meilleur que la combinaison de modèles pour le bruit de sèche-cheveux. Nous expliquons les fluctuations des performances par l'absence d'un critère objectif pour définir la transformation, qui ne permet pas de garantir une convergence vers la condition d'apprentissage dans le bruit lorsque le bruit est faible, et conduit à de grandes fluctuations des performances à 0 dB. Cependant, comme cette méthode n'utilise aucune connaissance explicite sur la nature de la perturbation, elle présente l'intérêt de pouvoir s'appliquer pour des tâches d'adaptation au locuteur ou aux variations du canal d'acquisition du signal.

Enfin, il faut noter que pour les SNRs moyens et élevés ($\geq$ 12 dB), et pour tous les bruits testés, la régression linéaire est significativement plus efficace que toutes les autres approches de reconnaissance dans le bruit. Pour les SNRs supérieurs à 6 dB, la régression linéaire est toujours au moins équivalente à la condition d'apprentissage dans le bruit, et à 0 dB, cette méthode est au pire équivalente aux autres approches de reconnaissance dans le bruit. À titre d'exemple, pour un SNR de 12 dB, la régression linéaire permet une réduction des taux d'erreurs de reconnaissance, tous types de bruits confondus, de :

- 38% à 54% par rapport à la transformation de base ;

- 42% à 56% par rapport à la combinaison de modèles ;

– 45% à 57% par rapport au filtrage par états.

Tout comme l'approche par transformation de base, l'adaptation par régression linéaire n'utilise pas de connaissance spécifique sur les variations entre conditions de test et d'apprentissage, et est donc potentiellement applicable à des tâches d'adaptation au locuteur ou à la ligne de transmission.

En définitive, pour mettre en évidence l'intérêt de l'adaptation par transformation linéaire, nous traçons fig. 9.13 les taux de reconnaissance en fonction du rapport signal-à-bruit pour les configurations de test suivantes. La première, notée Modèles de parole propre, correspond aux cas où les modèles, construits à partir de parole propre, sont utilisés pour reconnaître de la parole bruitée. La seconde configuration, notée Apprentissage dans le bruit, correspond à la situation où l'apprentissage et le test s'effectuent dans les mêmes conditions. Enfin, dans la configuration Régression linéaire, les modèles de parole propre sont adaptés par régression linéaire, pour reconnaître la parole bruitée. Il apparaît que cette dernière approche permet d'obtenir des résultats équivalents à ceux obtenus lorsque test et apprentissage sont effectués dans les mêmes conditions, pour les SNRs supérieurs à 6 dB, pour tous les types de bruit testés. Cela signifie qu'en utilisant 20 secondes de parole bruitée, nous sommes capable de transformer les modèles de parole propre pour construire des modèles qui permettent d'obtenir des taux de reconnaissance équivalents à ceux obtenus par les modèles entraînés à partir de 79 phrases de parole bruitée.

Dans certaines configurations de test, les taux de reconnaissance obtenus par l'adaptation par régression linéaire sont supérieurs à ceux obtenus lorsque les environnements de test et d'apprentissage sont semblables, ce qui est habituellement considéré comme la configuration optimale de test. Ce phénomène se produit principalement pour les SNRs de 12 et 18 dB, pour les bruits blanc Gaussien, bruit d'avion F16, et est particulièrement net pour le bruit de sèche-cheveux, où la régression linéaire permet de réduire d'environ 50% le taux d'erreur pour un SNR de 12 dB, par rapport à la condition d'apprentissage et de test dans le bruit. Des observations analogues sont également rapportées dans la littérature. Ainsi, sur une tâche de localisation de mots dans un flot de parole bruitée, une transformation probabiliste des paramètres des HMMs permet d'obtenir des performances supérieures à celles obtenues lorsque test et apprentissage sont effectués dans le bruit [Gish *et al.*, 1990]. De même, en utilisant une approche similaire, [Ng *et al.*, 1992] obtiennent des résultats équivalents.

Dans notre application, nous expliquons que l'adaptation par régression linéaire conduise à l'obtention de taux de reconnaissance supérieurs à la condition d'apprentissage dans le bruit, par l'absence de compensation des variances des modèles de parole propre. L'expérience suivante tente de mettre en évidence le phénomène. À partir du corpus d'apprentissage bruité, nous estimons tout d'abord par MLE les modèles acoustiques. Puis, nous modifions artificiellement les estimations des variances, en leur ajoutant et en leur retranchant 15%, et nous effectuons enfin la reconnaissance du corpus de test bruité avec ces différents modèles. La figure 9.14 indique les taux de reconnaissance obtenus pour un bruit blanc Gaussien pour des rapports signal-à-bruit variant de 0 à 36 dB, à partir des modèles suivants :

– modèles MLE initiaux (entraînés dans le bruit) ;

– modèles MLE initiaux où toutes les variances sont diminuées de 15% ;

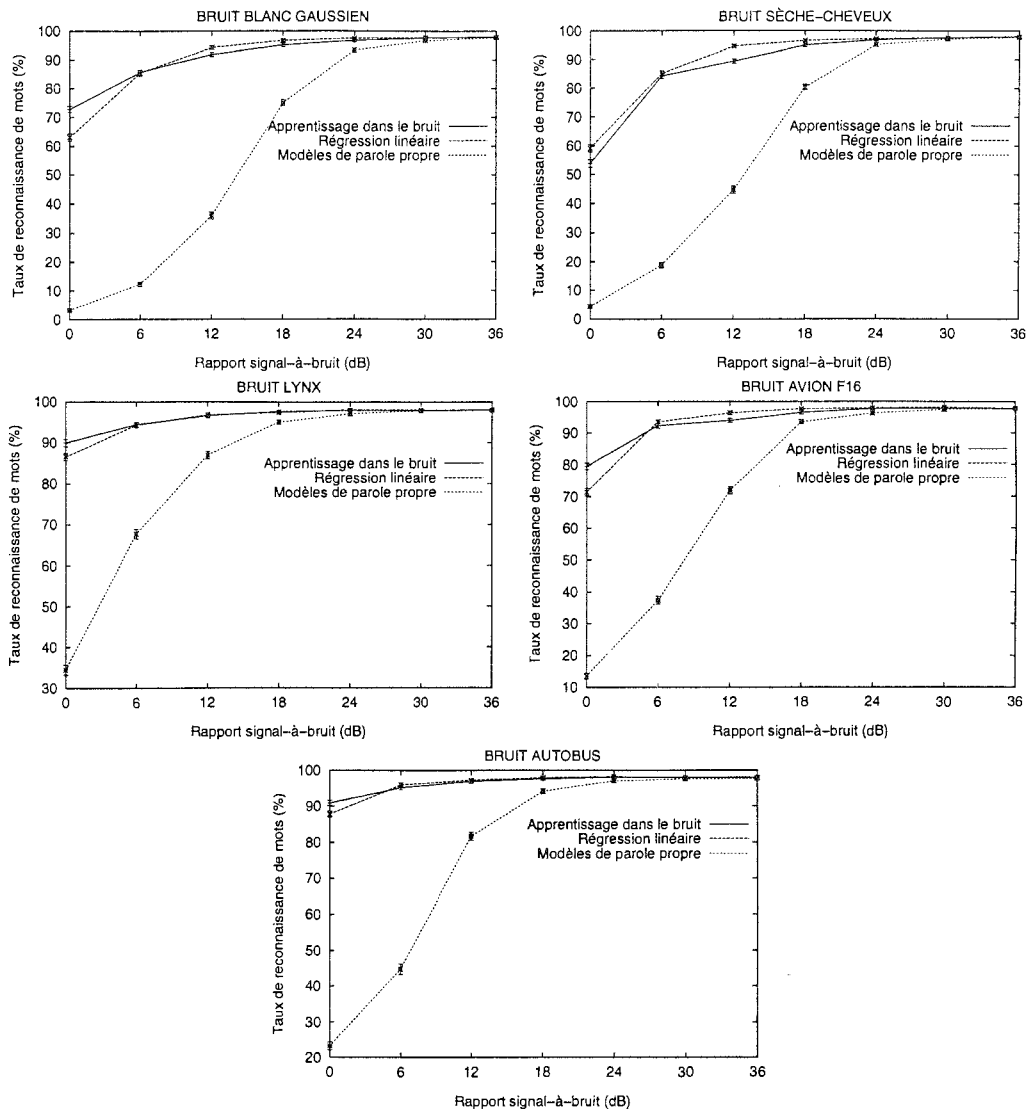

FIG. 9.13 - *Taux de reconnaissance* vs *SNR. 3 conditions de test : pas de compensation (Modèles de parole propre), apprentissage dans les conditions de test (Apprentissage dans le bruit), adaptation au bruit par régression linéaire (Régression linéaire). Moyenne sur tous les locuteurs.*

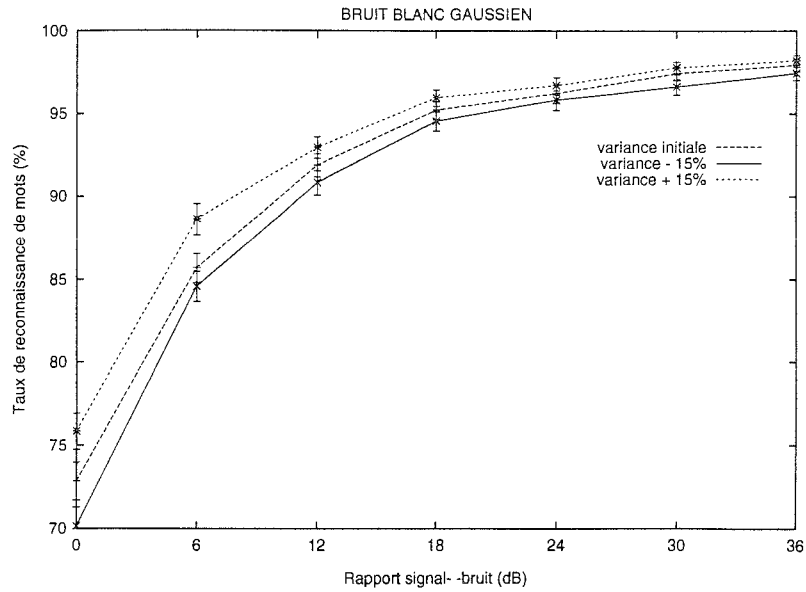

FIG. 9.14 - *Évolution des taux de reconnaissance en fonction de la modification de la variance des modèles MLE bruités. Moyenne sur tous les locuteurs.*

– modèles MLE initiaux où toutes les variances sont augmentées de 15%.

Il apparaît qu'augmenter les variances par rapport à leur estimation MLE dans le bruit permet d'améliorer les taux de reconnaissance. Il est possible que cela provienne d'un sous-entraînement des modèles. Cela signifie que pour notre application, utiliser des modèles dont les moyennes sont proches de celles des modèles de parole bruitée et dont les variances sont plus grandes que celles des modèles de parole bruitée, peut permettre d'améliorer les taux de reconnaissance par rapport aux modèles bruités estimés selon le critère MLE. C'est ce qui se produit lorsqu'on effectue l'adaptation par régression linéaire : les modèles adaptés ont une moyenne proche de celles de la parole bruitée, mais conservent les variances de la parole propre, qui sont plus grandes que celles de la parole bruitée (dans l'espace MFCC).

4 Sensibilité de la régression linéaire aux variations du SNR

Dans le paragraphe précédent, nous avons observé que l'approche par régression linéaire s'avère la plus performante par rapport aux autres méthodes développées. La régression linéaire nécessite de disposer d'un corpus d'adaptation de petite dimension (20 secondes), dans des conditions de bruit semblables à celles présentes lors du test. Il est alors intéressant d'étudier la robustesse de cette approche lorsque le rapport signal-à-bruit moyen du signal de test varie par rapport à celui du corpus d'apprentissage.

Dans ce paragraphe, nous décrivons les résultats de l'expérience suivante. Notons SNR_{ref}

le rapport signal-à-bruit de référence du corpus d'adaptation, et SNR_{test} le rapport signal-à-bruit lors du test. Les transformations linéaires sont tout d'abord déterminées à partir du corpus d'adaptation à SNR_{ref}, et sont utilisées pour transformer les modèles de parole propre. Puis, la reconnaissance du corpus de test est effectuée pour différentes valeurs de SNR_{test}. Cette configuration de test correspond donc à la situation où le rapport signal-à-bruit moyen évolue entre le moment où les modèles sont adaptés et l'instant où la reconnaissance est effectuée. En pratique, et afin de limiter le nombre de configurations de test, nous utilisons uniquement 3 valeurs différentes pour SNR_{test}, égales à : $\mathrm{SNR}_{ref} - 6$ dB, SNR_{ref}, $\mathrm{SNR}_{ref} + 6$ dB. La tâche de reconnaissance est la même que précédemment, et l'évaluation est effectuée pour tous les types et niveau de bruit, ce qui signifie que SNR_{ref} varie de 0 à 36 dB par pas de 6 dB, pour les 5 types de bruits considérés.

Les taux de reconnaissance obtenus sont présentés fig. 9.15. Au regard de ces courbes, il apparaît que l'adaptation par régression linéaire est peu sensible aux variations du SNR entre le corpus d'adaptation et le corpus de test, quelque soit le bruit considéré. Pour un SNR de référence donné SNR_{ref}, la condition optimale de test est généralement celle où $\mathrm{SNR}_{test} = \mathrm{SNR}_{ref}$.

5 Conclusion

Dans ce chapitre, nous avons comparé expérimentalement différentes méthodes permettant d'améliorer la robustesse du système VINICS sur une tâche de reconnaissance de parole continue en présence de différents types et niveaux de bruit. L'application considérée portait sur une reconnaissance en mode dépendant du locuteur, avec une grammaire comportant environ un millier de mots.

La méthode appelée combinaison de modèles (cf. chap. 6) permet de construire un STM de parole bruitée, à partir d'un HMM de bruit et d'un STM de parole propre. Cette approche présente l'intérêt de conduire à l'obtention de résultats qui convergent vers ceux obtenus lorsque les environnements de test et d'apprentissage sont semblables, pour les rapports signal-à-bruit élevés. Nous avons noté que l'hypothèse de base utilisée par la combinaison, qui consiste à supposer que la somme de deux variables aléatoires log-normales est également une variable aléatoire log-normale, conduit à la génération de STMs de parole bruitée sous optimaux (au sens du critère du maximum de la vraisemblance), lorsque le SNR est faible. Les avantages de cette méthode résident dans son absence de supervision, c.-à-d. que la compensation des modèles peut être appliquée dès que l'on dispose d'un modèle de bruit et des modèles de parole propre. De plus, la combinaison des modèles est peu coûteuse en calculs, et s'effectue en quelques secondes. La méthode peut potentiellement s'appliquer à des bruits non stationnaires, modélisés par des HMMs ergodiques, mais provoque dans ce cas une forte augmentation de la complexité. Pour cette raison, nous n'avons pas évalué ici cette approche dans de telles configurations, ce qui peut être une perspective de travail.

La seconde méthode consiste à appliquer un filtrage non linéaire, spécifique à chaque état de chaque STM, lors de la reconnaissance de la parole bruitée (cf. chap. 7). L'intérêt initial était d'appliquer une transformation du signal spécifique à chaque son, selon un critère significatif au niveau perceptif. Pour les rapports signal-à-bruit faibles, cette approche s'avère

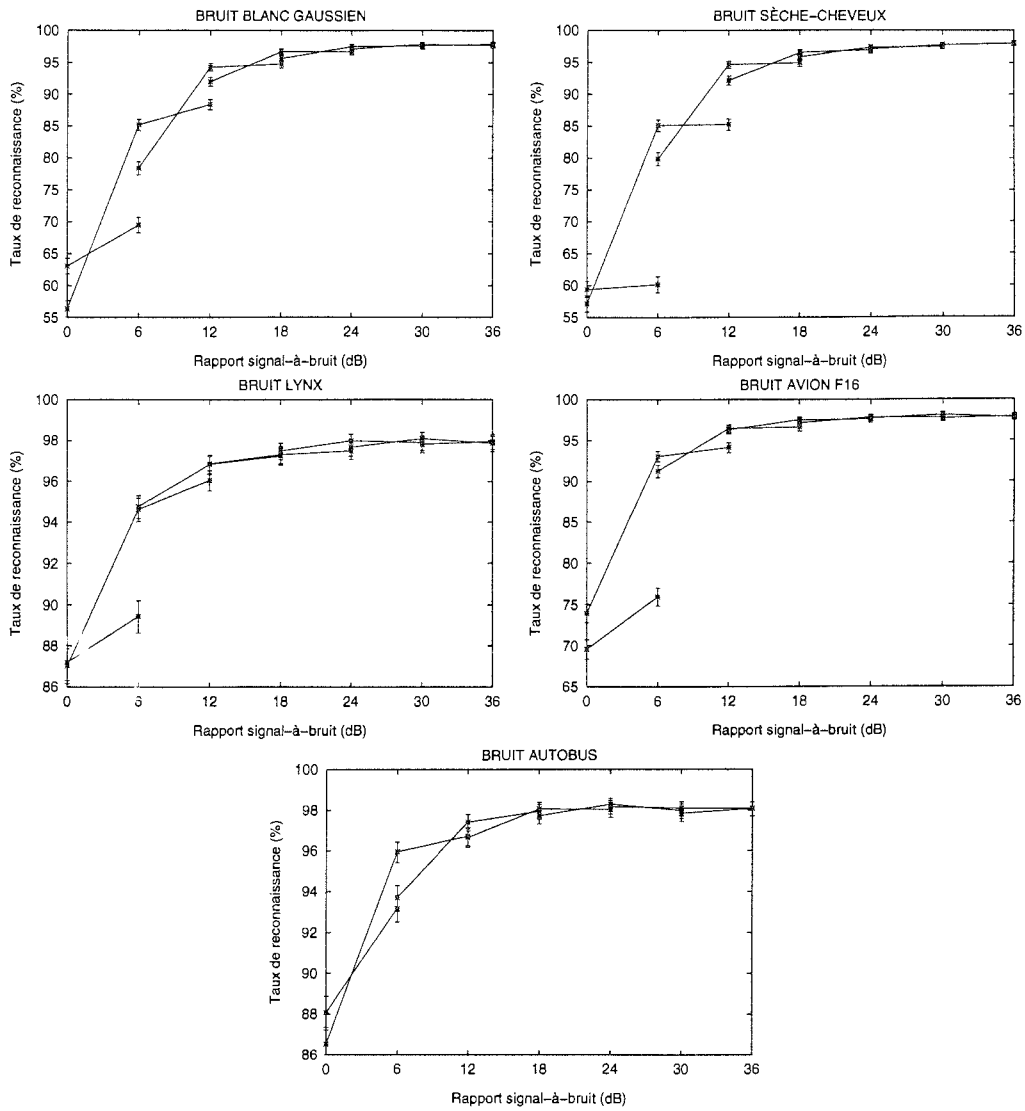

FIG. 9.15 - *Sensibilité de l'adaptation par régression linéaire aux variations du rapport signal-à-bruit entre le corpus d'adaptation et le corpus de test. Moyenne sur tous les locuteurs.*

la moins efficace, mais reste équivalente à la combinaison de modèles pour les SNRs très élevés. Étant donné la charge de calculs nécessaire pour déterminer les estimateurs du cepstre débruité, cette méthode est en pratique sans intérêt, bien qu'elle soit non supervisée comme la combinaison de modèles.

Une troisième approche, appelée transformation de base (cf. annexe B), que nous n'avons pas développée mais uniquement évaluée sur notre tâche de reconnaissance, consiste à transformer l'espace de paramètres de parole bruitée, pour le projeter sur l'espace de paramètres de parole propre. La transformation s'effectue par un changement de base : les vecteurs de la base bruitée sont remplacés par les éléments associés de la base propre. Cette méthode est supervisée dans la mesure où il est nécessaire de disposer d'un corpus d'adaptation de parole bruitée, étiqueté mais de taille réduite (20 secondes), afin de déterminer la correspondance entre les bases et donc d'effectuer la transformation de la parole bruitée. Cette approche s'avère la moins efficace de toutes lorsque le rapport signal-à-bruit est élevé, et les performances que l'on obtient fluctuent par rapport aux autres approches, pour les SNRs modérés. Nous expliquons ces phénomènes par l'absence d'un critère objectif utilisé pour définir la transformation. Par contre, cette approche n'utilise aucune connaissance sur la nature des variations entre les conditions de test et d'apprentissage, et peut donc être appliquée pour effectuer une adaptation au locuteur ou encore une adaptation à la ligne de transmission du signal.

Enfin, nous avons proposé une méthode pour adapter les paramètres des STMs de parole propre aux variations du bruit d'environnement (cf. chap. 8). L'adaptation consiste à transformer linéairement les vecteurs moyennes des modèles, de façon spécifique à différentes classes de sons. Ces transformations sont déterminées selon un critère objectif qui consiste à maximiser la vraisemblance d'un corpus d'adaptation de parole bruitée, étiqueté et de taille réduite (20 secondes). Cette méthode est celle qui fournit les meilleurs résultats. Pour les rapports signal-à-bruit supérieurs à 6 dB, les performances sont toujours au moins équivalentes à celles obtenues lorsque les conditions de test et d'apprentissage sont semblables. Nous avons également constaté que la méthode autorise une fluctuation du SNR moyen entre l'étape d'estimation des transformations et l'étape de test, ce qui constitue un intérêt pratique évident. Notons encore que comme pour la transformation de base, la nature des variations d'environnement entre test et apprentissage n'est pas exploitée, et la méthode est donc potentiellement applicable pour compenser des variations de locuteur ou bien de canal d'acquisition du signal. Par contre, cette approche présente l'inconvénient d'être supervisée, ce qui limite fortement son intérêt pratique, bien que la quantité nécessaire de données d'adaptation soit réduite. Comme perspective d'étude, nous essayons de nous affranchir de cette supervision en recherchant simultanément un étiquetage et un ensemble de transformations qui maximisent la vraisemblance du corpus d'adaptation.

Quatrième partie

Prétraitement par LDA et adaptation des modèles de durée

Introduction

Dans cette partie, deux approches indépendantes sont proposées, l'une concernant la reconnaissance de la parole dans le bruit, l'autre la reconnaissance de la parole en présence d'effet Lombard.

Dans le chapitre 10, nous proposons d'utiliser l'analyse linéaire discriminante (LDA) pour tenter de définir un paramétrage du signal de parole robuste au bruit. Cette approche s'inscrit dans le cadre des travaux sur la recherche d'un paramétrage robuste du signal de parole. Ici, nous mettons en œuvre une transformation de l'espace cepstral (MFCC + ΔMFCC) fondée sur la LDA, afin de définir un nouvel espace de paramètres. Dans un premier temps, nous étudions expérimentalement la précision du paramétrage construit par LDA, par rapport au paramétrage MFCC. Ensuite, nous étudions la sensibilité de ce paramétrage aux variations du rapport signal-à-bruit. Ce paramétrage est évaluée sur une tâche de reconnaissance de parole continue dans le bruit, identique à celle de la partie III. Cependant, cette méthode ne peut être comparée avec les approches précédentes, car ici l'apprentissage et le test s'effectuent dans les mêmes conditions d'environnement. Il ne s'agit donc pas de transformer un système de reconnaissance entraîné dans un environnement calme pour reconnaître de la parole dans le bruit, ni de filtrer le signal pour atténuer les effets du bruit, mais plutôt de rechercher à définir un paramétrage insensible au bruit.

Le chapitre 11 se focalise sur l'adaptation de modèles de parole propre pour la reconnaissance de la parole Lombard. L'adaptation mise en œuvre ne concerne que les paramètres des modèles de durée utilisés dans le formalisme du STM, et contrairement aux méthodes proposées dans la partie III, les modèles acoustiques ne sont pas modifiés. La démarche utilisée pour adapter les modèles s'inscrit dans le cadre de l'adaptation Bayesienne, et utilise un corpus d'adaptation étiqueté de parole Lombard pour transformer les modèles. L'approche est évaluée sur une tâche de reconnaissance de mots isolés en langue anglaise, en utilisant un corpus de parole avec effet Lombard.

Chapitre 10

Robustesse du prétraitement par analyse linéaire discriminante

1 Introduction

Dans la partie I, chapitre 3, nous avons présenté l'utilisation d'un paramétrage robuste du signal de parole comme une approche possible pour la reconnaissance automatique de la parole dans le bruit. En particulier, nous avons noté que l'analyse linéaire discriminante fournit un cadre propice pour définir un paramétrage précis du signal de parole, pour la reconnaissance de parole propre. Appliquée à des tâches de reconnaissance de petits et grands vocabulaires, l'utilisation de la LDA permet de réduire significativement les taux d'erreurs [Doddington, 1989; Haeb-Umbach et Ney, 1992; Hunt *et al.*, 1991].

Un autre avantage du paramétrage du signal par LDA réside dans son utilisation pour réduire la dimension de l'espace de paramètres. Les systèmes actuels de RAP utilisent de plus en plus des vecteurs de paramètres constitués d'éléments non homogènes (p.ex. cepstre avec Δcepstre). Cette hétérogénéité peut provoquer des difficultés pour définir une mesure de distance. De plus, les vecteurs de paramètres sont souvent de grande dimension. Cela nécessite des capacités de calcul et de stockage de données importantes. D'autre part, plus la dimension des vecteurs augmente, plus il est nécessaire d'utiliser une quantité importante de données pour entraîner les modèles [Kanal et Chandrasekaran, 1971]. La LDA présente donc l'avantage de permettre de réduire la dimension en projetant l'espace original de dimension D sur un sous-espace de dimension d ($d \leq D$) [Paliwal, 1992]. Jusqu'à présent, aucun travail ne reporte, à notre connaissance, d'expériences concernant la robustesse du prétraitement par LDA aux variations du niveau de bruit, pour la reconnaissance de la parole dans le bruit. Ainsi, dans les différentes expériences autour de IMELDA [Hunt et Lefèbvre, 1988], les variations du SNR ne sont pas prises en compte.

Dans ce chapitre, nous étudions expérimentalement la robustesse d'un paramétrage du signal par LDA, pour une reconnaissance de la parole dans le bruit, lorsque la transformation LDA, calculée pour un SNR donné, est appliquée pour transformer un espace de paramètres ayant un SNR différent. Le paragraphe 2 présente le cadre théorique de l'analyse linéaire discriminante. Les conditions expérimentales et les résultats obtenus sont décrit paragraphe 3.

Le paragraphe 4 conclut ce chapitre.

2 Analyse linéaire discriminante

L'analyse linéaire discriminante est une technique employée en classification, permettant d'améliorer la discrimination et de compresser l'information utile pour la classification, contenue dans un vecteur de paramètres. L'objectif est de rechercher la transformation linéaire d'un espace de paramètres de dimension D, permettant de projeter cet espace sur un espace de dimension d ($d \leq D$), dans lequel la séparation entre les classes est maximale, selon un certain critère.

Notons $\mathbf{X}_{kn}$ le n-ième vecteur de dimension D, appartenant à la classe k, et N_k le nombre de vecteurs appartenant à la classe k. La moyenne des observations dans la classe k, $\boldsymbol{\mu}_k$, et la moyenne totale des observations $\boldsymbol{\mu}$, s'écrivent, si K désigne le nombre total de classes et N le nombre total de vecteurs d'observations :

$$\boldsymbol{\mu}_k = \frac{1}{N_k} \sum_{n=1}^{N_k} \mathbf{X}_{kn} \tag{10.1}$$

$$\boldsymbol{\mu} = \frac{1}{N} \sum_{k=1}^{K} N_k \boldsymbol{\mu}_k \tag{10.2}$$

Notons respectivement $\mathbf{W}$ et $\mathbf{B}$ les matrices de covariances intra et inter-classes, définies comme suit, où $(\cdot)^{\#}$ désigne l'opérateur de transposition :

$$\mathbf{B} = \frac{1}{N} \sum_{k=1}^{K} N_k (\boldsymbol{\mu}_k - \boldsymbol{\mu})(\boldsymbol{\mu}_k - \boldsymbol{\mu})^{\#} \tag{10.3}$$

$$\mathbf{W} = \frac{1}{N} \sum_{k=1}^{K} \sum_{n=1}^{N_k} (\mathbf{X}_{kn} - \boldsymbol{\mu}_k)(\mathbf{X}_{kn} - \boldsymbol{\mu}_k)^{\#} \tag{10.4}$$

L'objectif est de déterminer une matrice de transformation $\mathbf{U}$, de dimension $D \times d$, utilisée pour projeter une observation $\mathbf{X}$ de dimension D en un vecteur $\mathbf{Y}$ de dimension d, avec $\mathbf{Y} = \mathbf{U}^{\#}\mathbf{X}$. Un critère de calcul couramment utilisé est celui qui maximise la trace de la matrice $\mathbf{W}^{-1}\mathbf{B}$. Intuitivement, cela signifie que dans l'espace transformé, la variance intra-classes est minimale (les données spécifiques à une classe sont « regroupées »), et que la variance inter-classes est maximale (les différentes classes sont « éloignées » les unes des autres).

En utilisant ce critère de calcul, il est possible de montrer que les d colonnes de la matrice de transformation $\mathbf{U}$ correspondent aux d vecteurs propres de la matrice $\mathbf{W}^{-1}\mathbf{B}$ dont les valeurs propres sont les plus grandes. Étant donné que la matrice $\mathbf{W}^{-1}\mathbf{B}$ n'est pas symétrique, la détermination de toutes les valeurs propres et de tous les vecteurs propres n'est pas triviale. On peut montrer que la matrice $\mathbf{U}$ s'obtient par [Atal, 1972] :

$$\mathbf{U} = \mathbf{C}\mathbf{L}^{-1/2}\mathbf{V} \tag{10.5}$$

où :

- C est une matrice unitaire ($C^{-1} = C^{\#}$) diagonalisant la matrice W, c.-à-d. que les vecteurs colonnes de C sont le vecteurs propres de la matrice symétrique W,
- L est une matrice diagonale constituée des valeurs propres de la matrice W, $C^{\#}WC = L$,
- V est une matrice unitaire dont les colonnes sont les d vecteurs propres associés aux d plus grandes valeurs propres de la matrice symétrique S, avec $S = L^{-1/2}C^{\#}BCL^{1/2}$.

Lorsque la matrice de transformation U est déterminée, toutes les observations de l'espace initial de dimension D peuvent être projetées sur l'espace discriminant de dimension d.

3 Expériences et résultats

3.1 Conditions expérimentales

Les expériences concernant le prétraitement par LDA sont effectuées sur la même tâche de reconnaissance que dans le chap. 9. L'espace initial que nous transformons par application de la LDA est constitué de l'espace des paramètres MFCC+ΔMFCC, ce qui signifie que dans cet espace, un vecteur de paramètres est formé de la concaténation de 13 coefficients MFCC avec leurs 13 dérivées premières ΔMFCC, soit un espace de dimension $D = 26$. Les paramètres dynamiques sont calculées par régression [Furui, 1986b], en utilisant 2 vecteurs de part et d'autre du vecteur courant.

Habituellement, le prétraitement du signal par LDA est appliqué de la façon suivante :

1. Les corpus d'apprentissage et de test sont paramétrés dans l'espace d'origine (p.ex. MFCC+ΔMFCC).
2. Une transformation linéaire est déterminée selon le critère présenté § 2, en utilisant le corpus d'apprentissage.
3. Les corpus d'apprentissage et de test sont projetés dans l'espace LDA avec la transformation déterminée en 2.
4. Les modèles du système de RAP sont construits en utilisant le corpus d'apprentissage paramétré dans l'espace LDA.
5. La reconnaissance du corpus de test paramétré dans l'espace LDA s'effectue avec les modèles entraînés en 4.

Dans une telle procédure, les conditions de test et d'apprentissage sont identiques, la LDA est utilisée pour déterminer un espace de paramètres plus discriminant que l'espace initial MFCC+ΔMFCC.

Ici, nous souhaitons d'une part mettre en évidence l'augmentation de la discrimination entre les différentes classes provoquée par le paramétrage LDA par rapport aux paramétrage initial, et d'autre part étudier la robustesse de la transformation LDA lorsque le rapport signal-à-bruit du corpus utilisé pour définir la transformation est différent de celui du corpus projeté par cette transformation. Pour cela, nous définissons la procédure de test suivante, appelée *mode croisé* :

1. Le corpus d'apprentissage de parole bruitée à SNR_{ref} dB est paramétré dans l'espace MFCC+ΔMFCC.
2. La matrice de transformation LDA, notée $\mathrm{U}_{\mathrm{SNR}_{ref}}$ est déterminée à partir du corpus d'apprentissage paramétré en 1.
3. Le corpus d'apprentissage de l'étape 1 est projeté dans l'espace LDA avec la matrice $\mathrm{U}_{\mathrm{SNR}_{ref}}$.
4. Les modèles acoustiques sont construits à partir du corpus d'apprentissage paramétré dans l'espace LDA obtenu en 3.
5. Le corpus de test de parole bruitée à SNR_{test} dB est paramétré dans l'espace MFCC+ΔMFCC.
6. Le corpus de test de l'étape 5 est projeté dans l'espace LDA avec la matrice $\mathrm{U}_{\mathrm{SNR}_{ref}}$.
7. Le corpus de test obtenu en 6 est reconnu en utilisant les modèles construits en 3.

Cette procédure expérimentale, résumée sur le schéma fig. 10.1, correspond à la situation pratique où la matrice de transformation et les modèles sont construits dans une condition d'environnement donnée (SNR_{ref}), puis utilisés dans une condition différente (SNR_{test}). Afin de limiter le nombre d'expériences, nous conservons les mêmes configurations pour SNR_{ref} et SNR_{test} que celles appliquées dans le chap. 9, § 4 : SNR_{ref} varie de 0 à 36 dB par pas de 6 dB, et SNR_{test} prend les valeurs $\mathrm{SNR}_{ref} - 6$ dB, SNR_{ref}, $\mathrm{SNR}_{ref} + 6$ dB.

3.2 Résultats

3.2.1 Comparaison de la robustesse du paramétrage LDA et du paramétrage MFCC

Lorsque la LDA est mise en œuvre, il est nécessaire de définir la notion de classe, utilisée dans le critère d'estimation de la transformation. Étant donné que le système de reconnaissance que nous utilisons est fondé sur la modélisation des sons associés à chaque phonème, nous choisissons dans un premier temps de regrouper dans une classe les sons spécifiques à un phonème.

Une première série d'expériences consiste à comparer le pouvoir de discrimination de l'espace généré par la LDA à celui de l'espace MFCC. Pour cela, nous effectuons une évaluation de la précision des modèles obtenus dans l'espace LDA et dans l'espace MFCC, en comparant les taux de reconnaissance phonétique du corpus d'apprentissage. L'évaluation s'effectue en mode croisé, défini paragraphe 3.1. Bien évidemment, les modèles définis dans

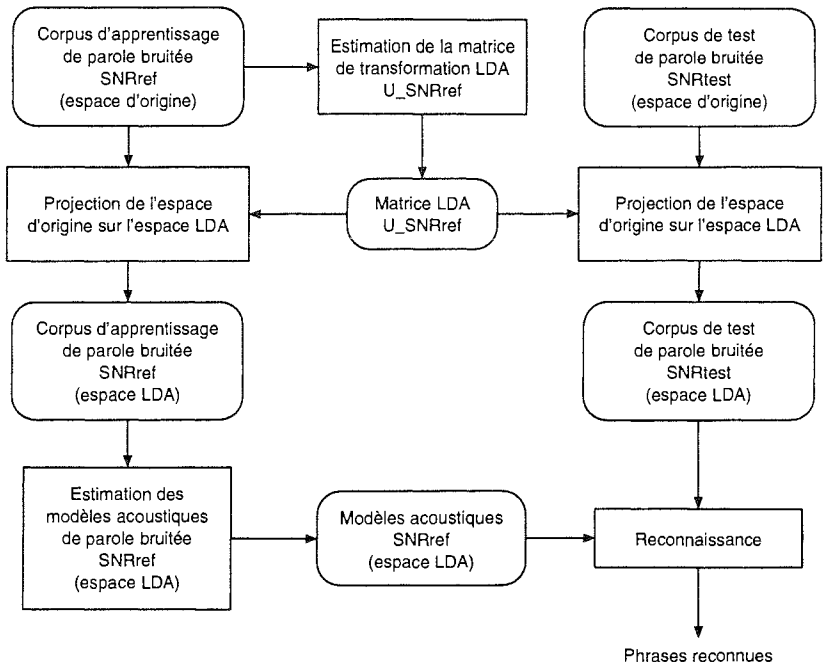

FIG. 10.1 - *Procédure de test de la robustesse de la LDA aux variations du rapport signal-à-bruit entre le corpus d'apprentissage et le corpus de test.*

l'espace MFCC et ceux définis dans l'espace LDA doivent comporter le même nombre de paramètres, sans quoi la comparaison n'a pas de sens. Ici, nous utilisons un paramétrage MFCC d'ordre 13, et la LDA effectue une projection de l'espace MFCC+ΔMFCC sur un espace de dimension 13. La comparaison porte donc sur la précision des modèles, c'est à dire sur leur capacité à différencier les classes du corpus d'apprentissage. La figure 10.2 représente les résultats obtenus, qui correspondent au pourcentage des symboles du corpus d'apprentissage correctement reconnus, lorsque les modèles sont construits dans l'espace LDA et dans l'espace MFCC.

Au regard de ces courbes, on peut distinguer plusieurs phénomènes. Tout d'abord, lorsque les SNRs de référence et de test sont égaux ($\mathrm{SNR}_{ref} = \mathrm{SNR}_{test}$), les modèles déterminés dans l'espace LDA sont toujours plus précis que les modèles déterminés dans l'espace MFCC, pour toutes les configurations de type et niveau de bruit: l'utilisation de la LDA permet d'améliorer le taux d'évaluation phonétique de 4 à 9%. Lorsque les SNRs de référence et de test sont différents ($\mathrm{SNR}_{ref} \neq \mathrm{SNR}_{test}$), les résultats varient selon la nature du bruit. Pour les bruits d'autobus et de Lynx, le paramétrage par LDA reste toujours plus précis que le paramétrage par MFCC, et peut conduire jusqu'à 30% d'amélioration du taux d'évaluation phonétique, par rapport au paramétrage MFCC. En présence de bruits blancs (Gaussien et sèche-cheveux), les tendances sont opposées. Les modèles fondés sur la LDA, construits pour le SNR_{ref} sont impropres à la reconnaissance du corpus de test à SNR_{test}, lorsque les SNRs sont faibles. Par exemple, le paramétrage LDA provoque une chute d'environ 60% des taux de reconnaissance phonétique par rapport au paramétrage MFCC lorsque SNR_{ref}

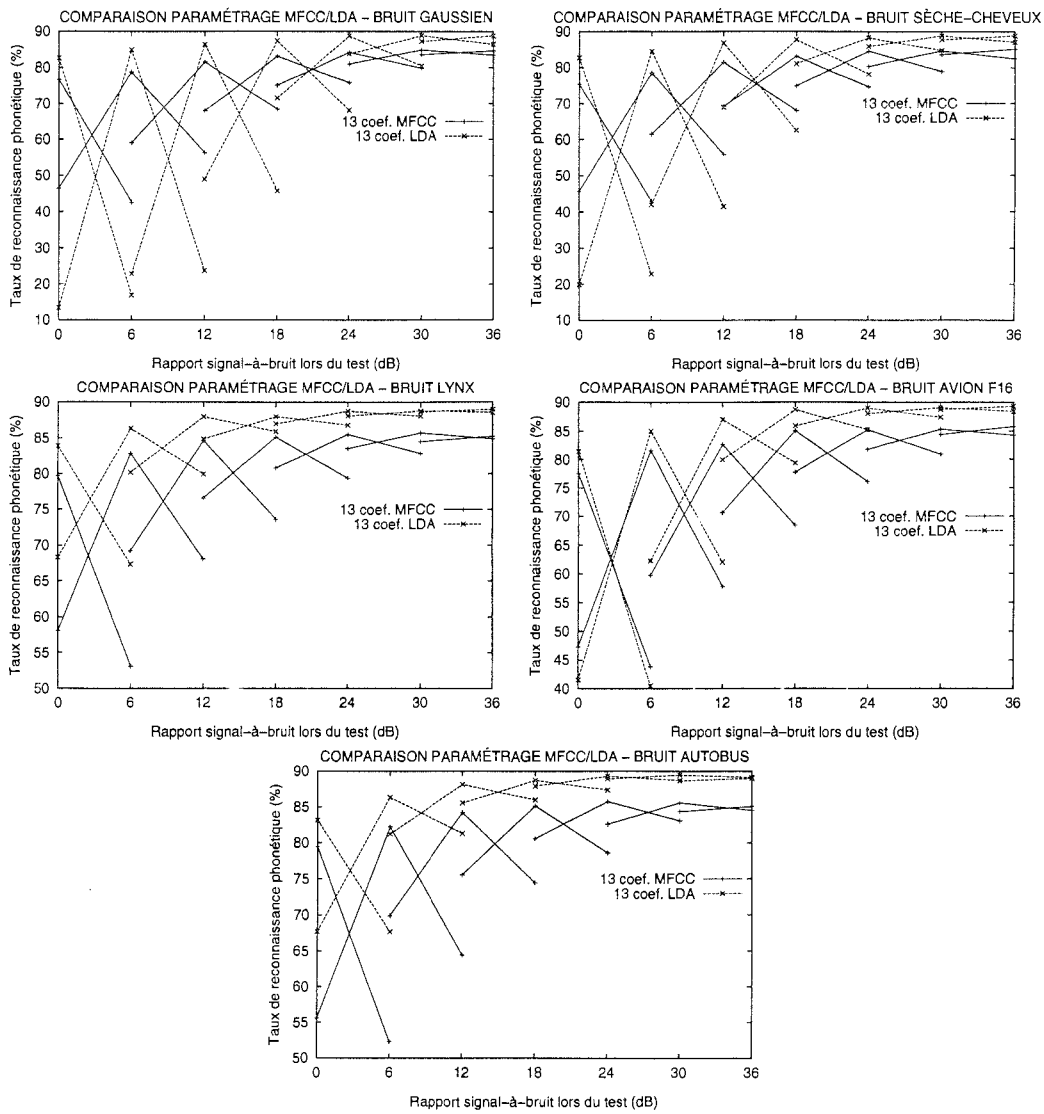

FIG. 10.2 - *Comparaison de la robustesse du paramétrage LDA / MFCC aux variations du rapport signal-à-bruit entre le corpus utilisé pour déterminer la matrice de projection LDA et le corpus d'évaluation. Les courbes indiquent le taux de reconnaissance phonétique du corpus utilisé pour l'apprentissage des modèles. Moyenne sur tous les locuteurs.*

= 12 dB et SNR_{test} = 6 dB. Lorsque le SNR augmente, le paramétrage LDA redevient plus précis que le paramétrage MFCC.

Si on s'intéresse à la robustesse des paramétrages MFCC et LDA aux variations du SNR, on peut tracer les courbes précédentes en les normalisant à 0 pour $SNR_{ref} = SNR_{test}$ (cf. fig. 10.3), ce qui met en évidence la chute relative des performances par rapport à la condition $SNR_{ref} = SNR_{test}$. La dépendance au type de bruit apparaît clairement. En présence du bruit Gaussien et du bruit de sèche-cheveux, le paramétrage MFCC est beaucoup plus robuste que LDA. Par contre la LDA s'avère plus robuste que les paramètres MFCC pour les bruits d'autobus et de Lynx. Enfin, pour le bruit d'avion F16, les tendances évoluent en fonction du SNR. Lorsque le SNR est élevé ($\geq$ 18 dB) la LDA est plus robuste que les paramètres MFCC, alors qu'il est difficile de dégager une tendance pour les SNRs faibles.

Dans une seconde série d'expériences, toujours en mode croisé, nous évaluons les taux de reconnaissance sur le corpus de test. Cette évaluation porte sur le paramétrage dans l'espace LDA, et la dimension de l'espace LDA est cette fois fixée à 20. Le taux de reconnaissance obtenu avec un paramétrage MFCC d'ordre 13 est également représenté, mais uniquement pour des SNRs de test et d'apprentissage semblables ($SNR_{ref} = SNR_{test}$). Les taux de reconnaissance sont présentés fig. 10.4, pour toutes les configurations de types et niveaux de bruit.

On retrouve les tendances qui apparaissaient lors de l'évaluation de la précision des modèles. Lorsque SNR_{ref} et SNR_{test} sont égaux, les taux de reconnaissance obtenus par le paramétrage LDA sont toujours supérieurs à ceux utilisant le paramétrage MFCC. La table 10.1 met en évidence la diminution du taux d'erreur de reconnaissance lorsque l'on passe d'un paramétrage MFCC à un paramétrage par LDA. La réduction la plus importante est obtenue pour les SNRs faibles, mais varie néanmoins d'environ 11 % à 40 % selon le type du bruit, pour un SNR de 36 dB. On peut donc conclure que les paramètres calculés par LDA sont bien adaptés à la reconnaissance dans le bruit, lorsque les SNRs de test et d'apprentissage sont identiques.

SNR (dB)	Gaussien	Autobus	Lynx	Sèche-cheveux	F 16
0	-58.3 %	-4.3 %	-36.8 %	-72.9 %	-36.5 %
6	-65.7 %	-44.3 %	-44.1 %	-60.9 %	-56.1 %
12	-57.7 %	-31.7 %	-49.6 %	-65.9 %	-47.1 %
18	-34.2 %	-42.3 %	-42.4 %	-28.6 %	-49.3 %
24	-18.9 %	-30.9 %	-25.9 %	-32.8 %	-32.5 %
30	-9.4 %	-38.8 %	-21.2 %	-29.5 %	-18.4 %
36	-11.5 %	-30.7 %	-32.1 %	-20.8 %	-40.8 %

TAB. 10.1 - *Évolution du taux d'erreur de reconnaissance entre le paramétrage MFCC et LDA, lorsque* $SNR_{test} = SNR_{ref}$.

En ce qui concerne la robustesse aux variations du SNR entre le test et l'apprentissage, les comportements sont très variables selon la nature du bruit. Pour les bruits d'autobus, de

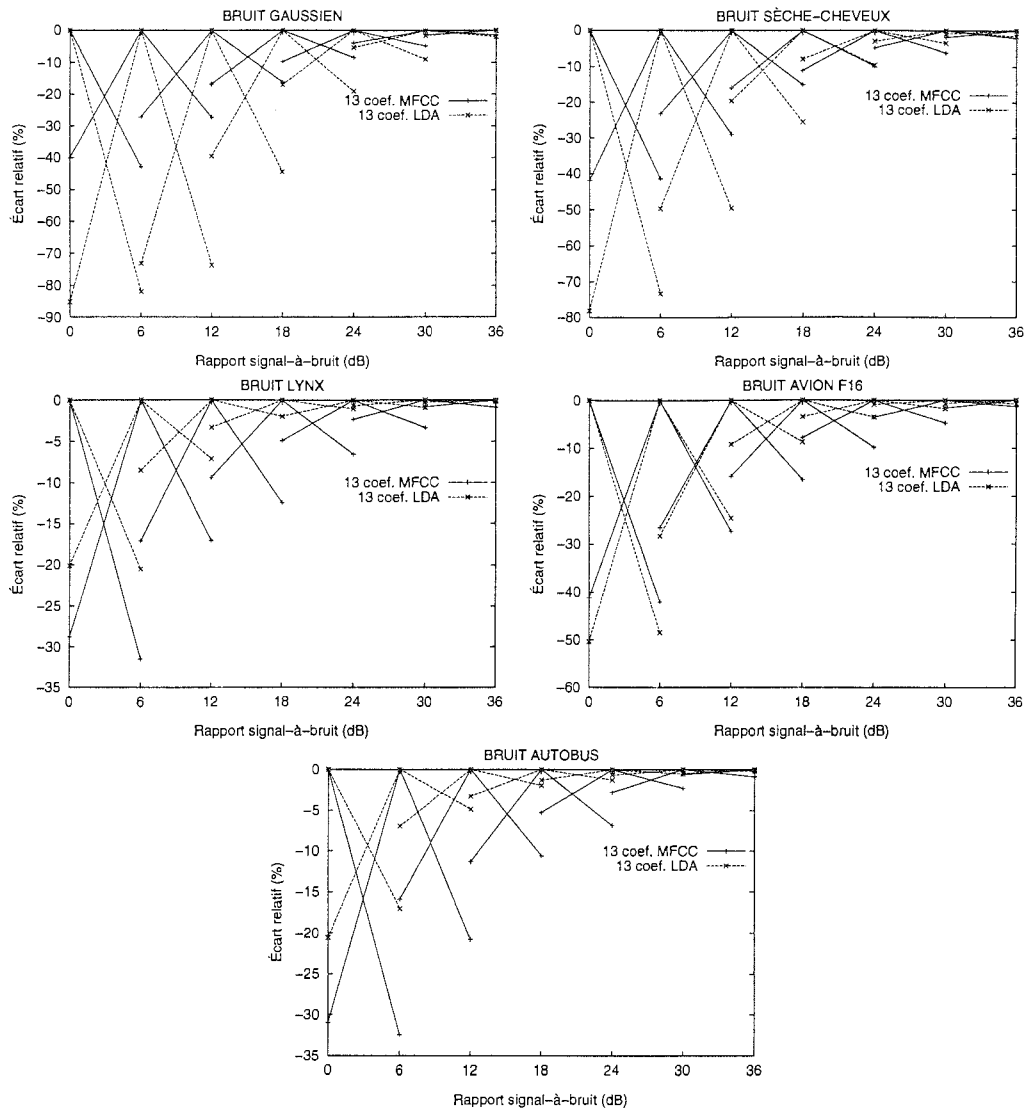

FIG. 10.3 - *Robustesse relative des paramétrage LDA / MFCC aux variations du rapport signal-à-bruit* SNR_{test} *et* SNR_{ref}. *Écart relatif des taux d'évaluation phonétique par rapport à la configuration où* $\mathrm{SNR}_{ref} = \mathrm{SNR}_{test}$. *Moyenne sur tous les locuteurs.*

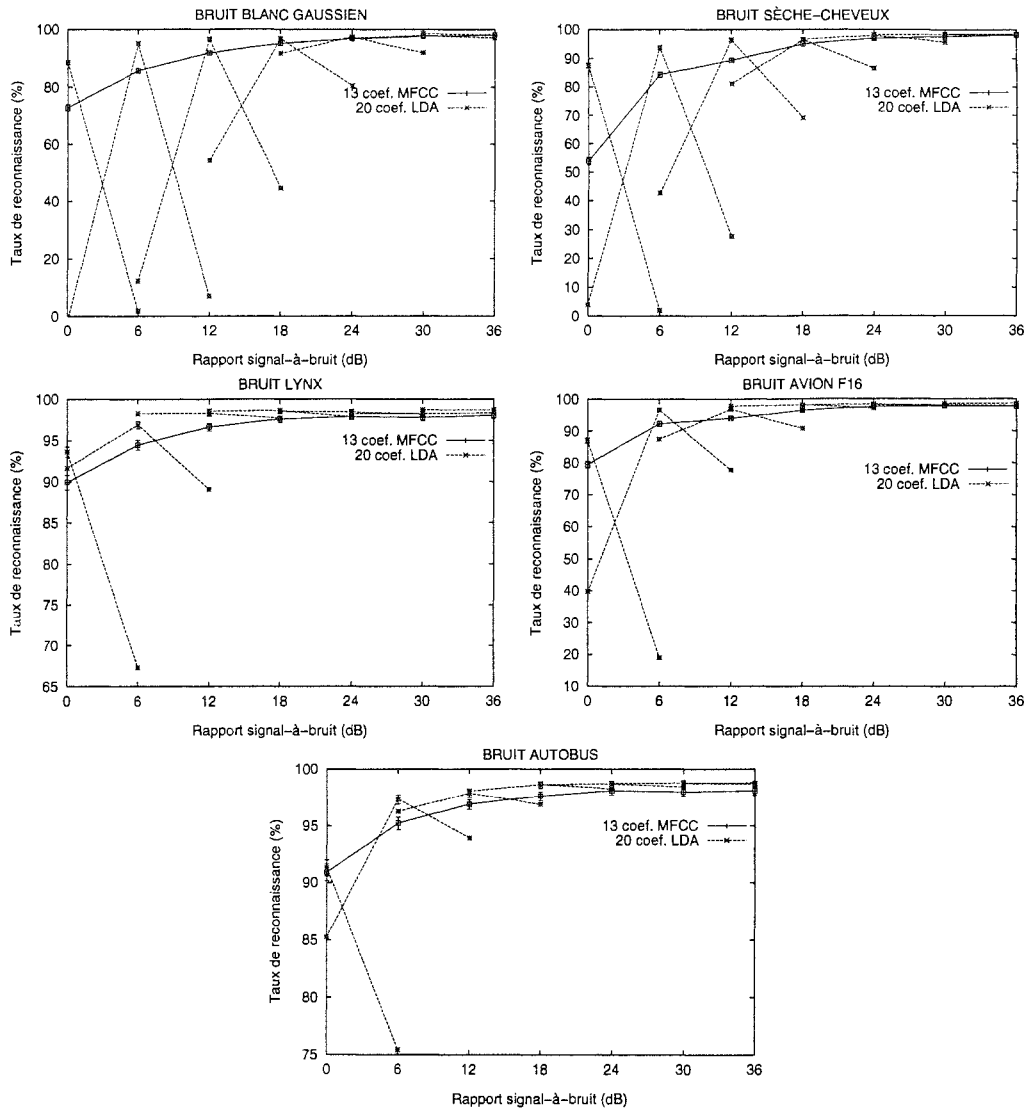

FIG. 10.4 - *Robustesse du prétraitement par LDA aux variations du rapport signal-à-bruit entre le corpus utilisé pour déterminer la matrice de projection LDA et le corpus de test. Moyenne sur tous les locuteurs.*

Lynx et d'avion F16, on observe que le paramétrage LDA est peu sensible aux variations entre SNR_{ref} et SNR_{test} : les courbes restent horizontales de 36 à 12 dB. Par contre, pour les bruits blancs comme le bruit Gaussien et le bruit de sèche-cheveux, les taux de reconnaissance s'effondrent en mode croisé, ce qui est conforme aux résultats de l'évaluation phonétique. On note enfin le comportement particulier en présence du bruit d'avion F16, où les taux de reconnaissance en mode croisé s'écroulent lorsque le SNR est faible, tandis qu'ils restent très robustes aux variations du SNR pour les SNRs modérés et élevés.

3.2.2 Influence du nombre d'observations associées à chaque classe

Lors du calcul de la matrice de transformation LDA, il est important d'éviter qu'une classe ait une influence prépondérante par rapport aux autres. Dans notre formulation, une classe est associée à un symbole phonétique. Or, les nombres d'occurrences des différents symboles dans le corpus d'apprentissage sont très variables : domination des symboles de silences, peu de /S/, /Z/, /f/, etc. Il convient donc, lors du calcul des matrices $\mathbf{W}$ et $\mathbf{B}$, de rééquilibrer l'histogramme de répartition des observations par classe, en limitant par exemple le nombre d'observations affectées à chaque classe. Dans le paragraphe précédent, cette limite était fixée au nombre moyen d'observations par classe, soit 70, c.-à-d. que seules les 70 premières observations associées à chaque classe étaient utilisées. Le nombre total d'observations dans le corpus d'apprentissage est d'environ 2300. La limitation à 70 symboles maximum par classe conduit à utiliser 1560 symboles pour calculer la matrice LDA.

Dans un premier temps, nous diminuons cette limite à 40, afin d'équilibrer plus encore l'histogramme des distributions. Cela revient à utiliser uniquement 1098 observations du corpus d'apprentissage. Une série d'expériences est menée pour comparer les taux de reconnaissance en mode croisé pour les 35 configurations de bruits, pour des limites de 40 et 70 observations par classe. Les résultats sont représentés fig. 10.5.

Lorsque $\mathrm{SNR}_{ref} = \mathrm{SNR}_{test}$, les deux configurations de test sont équivalentes, aucun écart significatif ne peut être mis en évidence. En ce qui concerne la robustesse aux variations du SNR entre le test et l'apprentissage, les tendances varient en fonction du bruit. Pour le bruit de Lynx et d'autobus, la configuration avec 40 observations au maximum par classe semble la plus robuste pour les SNRs faibles. Pour les autres bruits, les deux limitations à 40 et 70 sont équivalentes. Il n'est donc pas nécessaire d'utiliser la totalité du corpus d'apprentissage (soit en moyenne 70 symboles par classe) pour déterminer la matrice de transformation LDA.

Dans un second temps, nous choisissons de ne pas utiliser les symboles de silences (silences de début et fin de phrases, silences des plosives voisées et non voisées) lors du calcul de la matrice de transformation LDA. En effet, ces symboles sont très rarement confondus avec les symboles associés aux phonèmes, et notre objectif est d'améliorer la discrimination entre ces phonèmes. Nous effectuons une série de tests en mode croisé pour évaluer l'influence de l'utilisation des symboles de silences dans le calcul de la matrice de transformation LDA. La figure 10.6 reporte les résultats obtenus.

Une tendance générale est que la configuration avec suppression des silences permet d'améliorer la robustesse aux SNRs élevés, quelque soit le type de bruit. Pour les SNRs faibles, seul le bruit de Lynx conduit à une dégradation significative des performances lorsqu'on conserve les symboles de silences. Avec les autres bruits, la configuration avec rejet

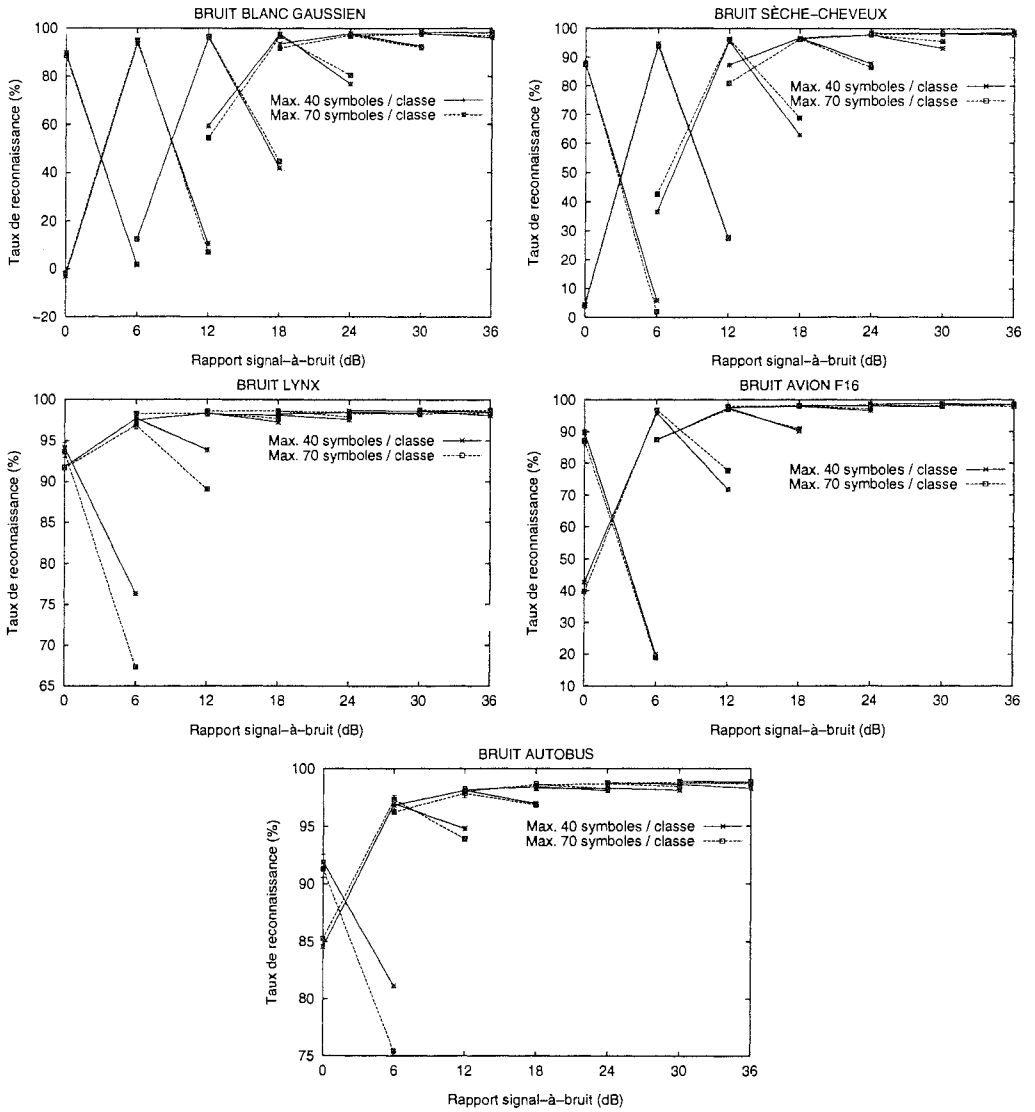

FIG. 10.5 - *Influence de la limitation du nombre maximum de symboles par classe lors du calcul de la matrice LDA sur la robustesse aux variations du SNR entre le test et l'apprentissage. Limitation à 40 et 70 symboles par classe. Moyenne sur tous les locuteurs.*

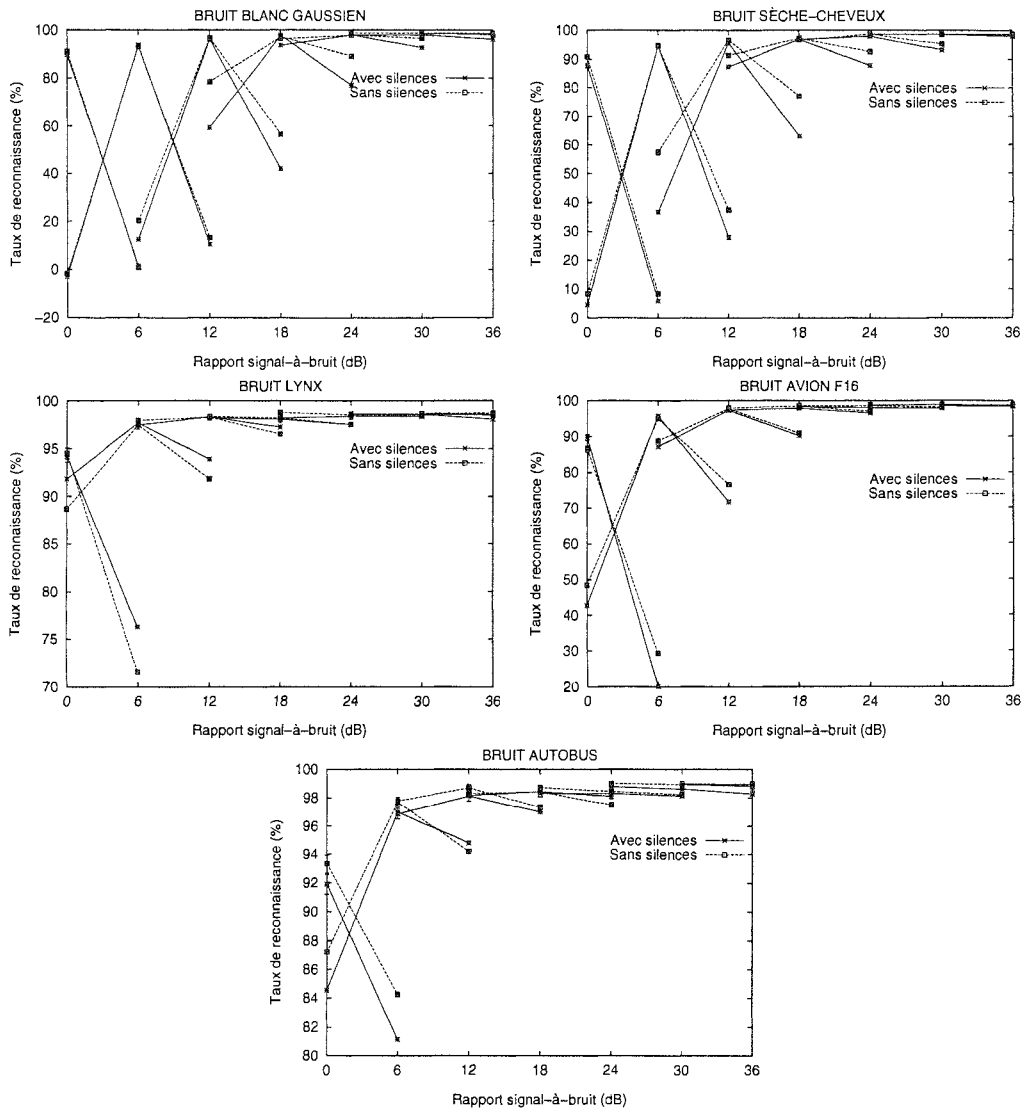

FIG. 10.6 - *Influence de l'utilisation ou non des symboles de silences lors du calcul de la matrice LDA sur la robustesse aux variations du SNR entre le test et l'apprentissage. Moyenne sur tous les locuteurs.*

des silences est toujours au moins équivalente à la configuration avec utilisation des silences.

3.2.3 Choix des classes à discriminer

L'utilisation de la LDA nécessite de définir la notion de classes que l'on souhaite discriminer. Comme nous utilisons un système de reconnaissance à base de modèles de phonèmes, nous avons jusqu'à présent choisi d'associer à une classe un phonème. D'autres choix sont cependant possibles, mais la complexité des interactions entre la LDA et les modèles utilisés est telle que seule une réponse expérimentale peut être apportée sur les avantages respectifs de telle ou telle association entre les classes utilisées pour la LDA et les unités acoustiques représentées par les modèles.

Nous choisissons dans ce paragraphe d'associer une classe à chaque état de chaque modèle. Si H désigne le nombre de modèles ($H = 32$), et Q le nombre d'états de chaque modèle ($Q = 5$), le nombre total de classes à discriminer est donc $H \times Q$. Nous effectuons un ensemble d'expériences en mode croisé pour les 35 configurations de bruits, en utilisant 40 observations au maximum par classe, à la fois dans le cas où une classe est associée à un symbole (c.-à-d. à un modèle), et dans le cas où une classe est associée à chaque état de chaque modèle. Les résultats sont représentés fig. 10.7.

Les courbes indiquent que pour tous les types de bruits, et pour les rapports signal-à-bruit faibles (< 18 dB), la configuration où un symbole phonétique est associé à chaque classe conduit aux meilleurs résultats, à la fois lorsque $\mathrm{SNR}_{test} = \mathrm{SNR}_{ref}$ et lorsque $\mathrm{SNR}_{test} \neq \mathrm{SNR}_{ref}$. Lorsque le SNR est plus élevé, une tendance commune à tous les types de bruit ne se dégage pas. Dans [Haeb-Umbach et Ney, 1992], l'utilisation d'un symbole par classe s'avérait moins efficace qu'une unité sous-phonétique par classe, sur une tâche de reconnaissance de parole propre. Ici, en présence du bruit Gaussien, l'utilisation d'un symbole par classe conduit au système le plus robuste alors que la conclusion est opposée en présence d'un bruit d'avion F16. Néanmoins, tous types de bruit confondus, l'utilisation d'un symbole par classe permet d'obtenir les résultats les plus robustes et les meilleurs.

4 Conclusion

Dans ce chapitre, nous avons étudié expérimentalement la robustesse d'un paramétrage du signal obtenu par LDA à partir d'un espace MFCC, pour la reconnaissance de la parole en présence de différents types et niveaux de bruits.

Nous avons tout d'abord mis en évidence que les paramètres LDA permettent de construire des modèles plus précis que les paramètres MFCC, pour un nombre de variables libres des modèles identiques. Pour tous les types et niveaux de bruits testés, l'utilisation du paramétrage par LDA permet de diminuer de façon significative les taux d'erreurs par rapport au paramétrage MFCC, lorsque le rapport signal-à-bruit n'évolue pas entre l'apprentissage et le test.

De plus, nous constatons que la robustesse du prétraitement par LDA est fortement conditionnée par la nature du bruit perturbateur. En présence des bruits Gaussien et de sèche-cheveux, le paramétrage LDA s'avère très sensible aux différences entre SNR_{test} et SNR_{ref},

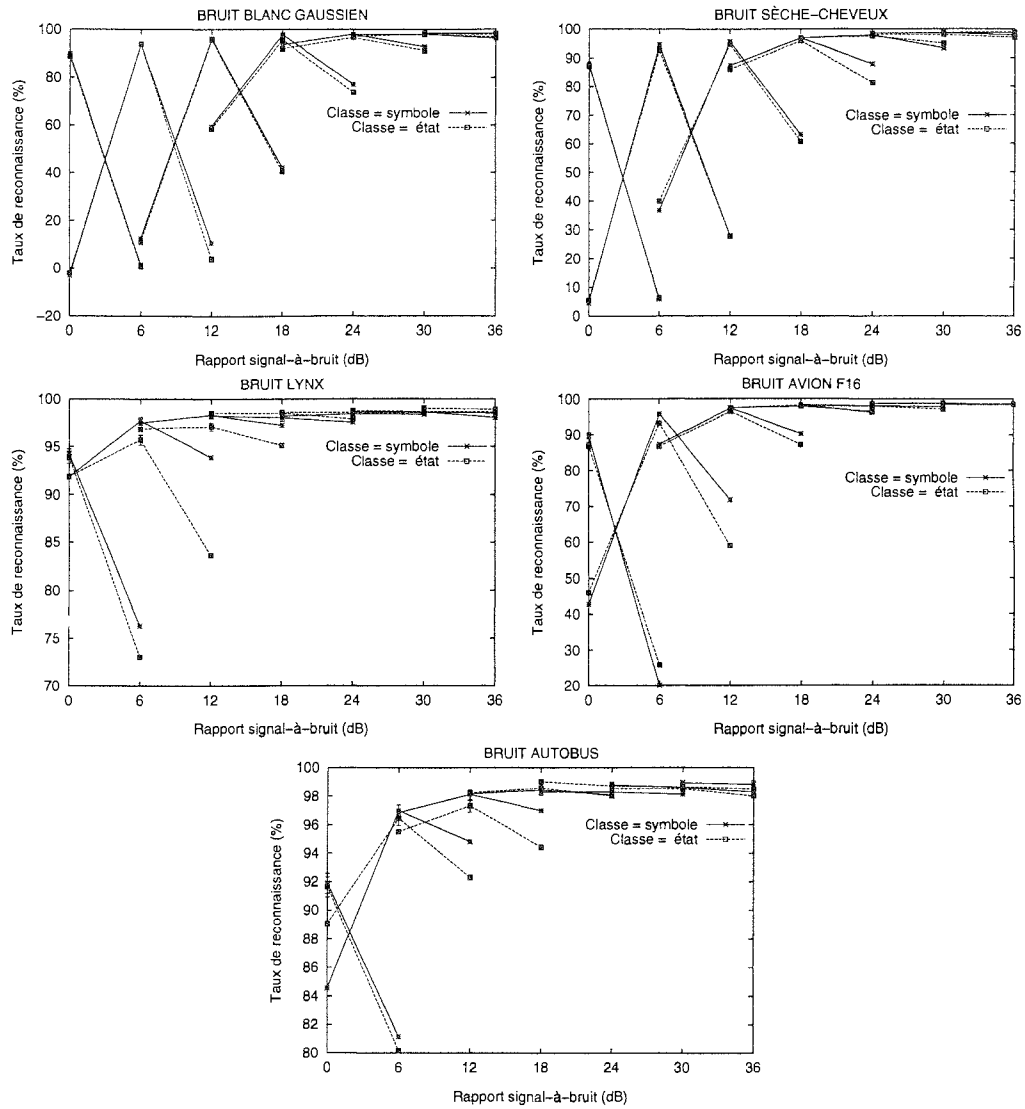

FIG. 10.7 - *Influence de la notion de classe sur la robustesse aux variations du SNR entre le test et l'apprentissage. Première configuration : classe = symbole. Deuxième configuration : classe = état. Moyenne sur tous les locuteurs.*

alors qu'une bonne robustesse aux variations du SNR est conservée en présence des autres types de bruits.

L'utilisation de la LDA nécessite de limiter l'influence dominante d'une classe par rapport aux autres, en équilibrant les distributions des observations dans chaque classe. Nous montrons que sur notre application, l'utilisation de 40 observations au maximum par classe, soit la moitié du corpus d'apprentissage, est suffisante pour déterminer la matrice de transformation LDA. De plus, la suppression des symboles de silences (silences des extrémités de phrases, des plosives) lors du calcul de la matrice de transformation LDA permet d'améliorer la robustesse, particulièrement pour les SNRs élevés.

Enfin, le choix d'unités sous-phonétiques (état des STMs) comme classe conduit, sur notre application et pour les rapports signal-à-bruit faibles, à une dégradation des performances par rapport à l'association d'un phonème à chaque classe.

Chapitre 11

Adaptation Bayésienne des modèles de durée des phonèmes

1 Introduction

En présence d'un bruit d'environnement important, le locuteur modifie de façon réflexe son mode articulatoire afin que ses propos restent intelligibles. Ce phénomène, appelé effet Lombard [Lombard, 1911], se traduit par la production d'un signal de parole dont les propriétés sont très différentes de celles d'un signal prononcé en environnement calme. Dans la partie I, nous avons décrit diverses approches permettant de réduire les variabilités provoquées par cet effet. Un point commun de ces méthodes est qu'elles se focalisent sur la réduction des distorsions au niveau spectral, et négligent la modification du rythme d'élocution. Cette modification est pourtant sévère, comme l'illustrent les études de [Junqua et Anglade, 1990; Junqua, 1993; Junqua, 1994].

De plus, nous avons noté en partie I, que l'utilisation d'un modèle de durée permet d'améliorer de façon significative les taux de reconnaissance. Aussi, nous nous intéressons dans ce chapitre à la prise en compte des fluctuations du rythme d'élocution, afin d'adapter les modèles de durée des phonèmes, entraînés à partir de parole propre, dans le but de reconnaître de la parole Lombard. Les modèles de durée sont adaptés sous le cadre de l'apprentissage Bayésien, présenté paragraphe 2. Les expériences réalisées, et les résultats obtenus sont exposés paragraphe 3. Le paragraphe 4 conclut ce chapitre.

2 Estimation Bayésienne

2.1 Principe

Soit X une variable aléatoire donc la fonction de densité de probabilité est notée $f_X(\cdot)$, de paramètre θ inconnu. Soit $\{x_1, \ldots x_N\}$ N échantillons indépendants, réalisations de la variable aléatoire X. Estimer θ à partir des N observations x_i, selon le critère du maximum

a posteriori revient à déterminer θ_{MAP} tel que :

$$\theta_{MAP} = \underset{\theta}{\operatorname{argmax}}\, f_{\Theta|X_1,\ldots X_N}(\theta|x_1,\ldots x_N) \tag{11.1}$$

En appliquant la règle de Bayes, l'équation (11.1) se réécrit :

$$\theta_{MAP} = \underset{\theta}{\operatorname{argmax}}\, \frac{f_{X_1,\ldots X_N|\Theta}(x_1,\ldots x_N|\theta) f_\Theta(\theta)}{f_{X_1,\ldots X_N}(x_1,\ldots x_N)} \tag{11.2}$$

où $f_{X_1,\ldots X_N|\Theta}(\cdot|\theta)$ désigne la *pdf* conjointe de $X_1,\ldots X_N$ étant donné θ, $f_\Theta(\cdot)$ représente la *pdf a priori* de Θ, et $f_{X_1,\ldots X_N}(\cdot)$ représente la *pdf* conjointe de $X_1,\ldots X_N$. Comme $f_{X_1,\ldots X_N}(\cdot)$ ne dépend pas de θ, et n'intervient donc pas dans la maximisation, l'équation (11.2) devient :

$$\theta_{MAP} = \underset{\theta}{\operatorname{argmax}}\, f_{X_1,\ldots X_N|\Theta}(x_1,\ldots x_N|\theta) f_\Theta(\theta) \tag{11.3}$$

Si Θ est supposé fixe mais inconnu, on ne dispose d'aucune connaissance sur Θ, ce qui revient à considérer que $f_\Theta(\cdot) =$ constante. Dans ces conditions, l'équation (11.3) se réduit alors à l'estimation habituelle au sens du maximum de la vraisemblance, soit $\theta_{MAP} = \theta_{MLE} = \operatorname{argmax}_\theta f_{X_1,\ldots X_N|\Theta}(x_1,\ldots x_N|\theta)$.

Si au contraire, Θ est aléatoire, $f_\Theta(\cdot)$ décrit la connaissance *a priori* du paramètre Θ. Dans ce cas, l'équation (11.3) permet de réestimer le paramètre θ en exploitant à la fois l'information *a priori* sur θ, $f_\Theta(\theta)$, et l'information apportée par les N observations x_i.

La résolution de l'équation (11.3) n'est généralement pas triviale, et différents problèmes doivent être considérés. D'abord, étant donné $f_{X|\Theta}(\cdot)$, comment faut-il choisir la loi $f_\Theta(\cdot)$? Ainsi, par exemple, si $f_{X|\Theta}(\cdot)$ est une loi normale dont les paramètres moyenne et variance constituent le vecteur aléatoire θ, quelles lois de probabilité doit-on choisir pour caractériser la moyenne et la variance ? D'autre part, comment estimer les paramètres de la loi *a priori* $f_\Theta(\cdot)$? Enfin, comment évaluer le maximum *a posteriori* ? Ces problèmes sont reliés entre-eux, et un choix judicieux de $f_\Theta(\cdot)$ permet d'obtenir une estimation MAP d'une façon relativement simple, semblable à une estimation MLE.

Ce choix se fonde sur la notion de famille conjuguée d'une distribution [DeGroot, 1970]. La famille de distributions conjuguée du paramètre Θ a la propriété suivante : si la distribution *a priori* de Θ, $f_\Theta(\cdot)$ appartient à cette famille, alors, pour n'importe quelle taille N des échantillons x_i, la distribution *a posteriori* de Θ, $f_{\Theta|\mathbf{X}_1,\ldots\mathbf{X}}(\cdot)$, appartient également à cette famille. Cela signifie que $f_{\Theta|\mathbf{X}_1,\ldots\mathbf{X}_N}(\theta|x_1,\ldots x_N) \propto f_{\mathbf{X}_1,\ldots\mathbf{X}_N|\Theta}(x_1,\ldots x_N|\theta) f_\Theta(\theta)$, où $\propto$ désigne la relation de proportionnalité, et où $f_\Theta(\theta)$ et $f_{\Theta|\mathbf{X}_1,\ldots\mathbf{X}_N}(\theta|x_1,\ldots x_N)$ appartiennent à la même famille de lois de probabilité. En d'autres termes, lorsque $f_\Theta(\theta)$ est connue, si de nouvelles observations x_i sont disponibles, la loi *a posteriori* à maximiser est de la même famille que la loi *a priori* $f_\Theta(\theta)$. Bien évidemment, d'autres choix seraient possibles pour la loi *a priori* $f_\Theta(\cdot)$, en se basant par exemple sur la signification physique du paramètre Θ.

2.2 Application à la réestimation des modèles de durée des phonèmes

En partie II, nous avons noté que la durée d'un phonème est considérée comme une variable aléatoire, modélisée dans VINICS par une loi Gamma. Malheureusement, la réestima-

tion des 2 paramètres d'une loi Gamma selon le critère du maximum *a posteriori* ne s'effectue pas sous forme close. Aussi, nous avons choisi de modéliser la durée par une loi normale, dont l'estimation MAP des paramètres ne pose pas de difficultés.

Le problème abordé est le suivant. On dispose au départ d'un modèle de durée pour chaque symbole phonétique, déterminé à partir de parole propre. On dispose également d'un petit corpus d'adaptation de parole avec effet Lombard. L'objectif est de modifier les paramètres des lois de probabilités des modèles de durée des symboles en parole propre, afin de les rendre spécifiques aux durées des symboles de la parole Lombard. Lorsque la *pdf* représentant la durée est une loi normale, de moyenne m et de *précision*[1] r, plusieurs types de réestimations peuvent être effectuées :

- réestimation de la moyenne seule, $\theta = m$, la précision r étant connue et fixe,
- réestimation de la précision seule, $\theta = r$, la moyenne m étant connue et fixe,
- réestimation de la moyenne et de la précision $\theta = (m, r)$.

L'adaptation de la précision seule n'est pas intéressante, car l'effet Lombard se manifeste principalement par un déplacement de la durée moyenne des phonèmes. Aussi, nous ne présentons que la réestimation de la moyenne seule, paragraphe 2.2.1, et la réestimation conjointe de la moyenne et la précision, paragraphe 2.2.2, qui n'est possible que lorsqu'un grand volume de données d'adaptation est disponible.

2.2.1 Adaptation de la moyenne

Supposons que l'on dispose de N observations de durée notées $x_1, \ldots x_N$, spécifiques à un symbole et issues du corpus d'adaptation. Ces données sont des réalisations d'une variable aléatoire modélisée par une loi normale de moyenne θ inconnue et de précision r, notée $\mathcal{N}(x; \theta, r)$. Notons $\bar{x}$ la moyenne empirique des N observations x_i, $\bar{x} = 1/N \cdot \sum_{i=1}^{N} x_i$.

La loi *a priori* conjuguée de θ est une loi normale de moyenne μ et de précision τ fixes et connues, notée $\mathcal{N}(\theta; \mu, \tau)$ [DeGroot, 1970]. Cette loi caractérise l'information dont on dispose *a priori*, concernant la moyenne θ de la durée des symboles. Sous ces hypothèses, il est possible de montrer que la distribution *a posteriori* de θ, connaissant les x_i, est une loi normale de moyenne μ' et précision τ' avec [DeGroot, 1970] :

$$\mu' = \frac{\tau\mu + Nr\bar{x}}{\tau + Nr} \tag{11.4}$$

$$\tau' = \tau + Nr \tag{11.5}$$

Il apparaît donc que la moyenne réestimée $\theta_{MAP} = \mu'$ est une somme pondérée de la moyenne des données d'adaptations $\bar{x}$, et de la moyenne *a priori* de la distribution de θ. Plus le nombre N de données d'adaptation est grand, plus le poids affecté à la moyenne empirique du corpus d'adaptation est important. Lorsque le nombre de données d'adaptation est nul, seule l'information *a priori* est prise en compte ; lorsque ce nombre tend vers l'infini, seule

[1] Par définition, la *précision* est l'inverse de la variance

la moyenne des données d'adaptation intervient. De plus, il faut noter que la précision de θ connaissant les x_i augmente de façon prédéfinie avec le nombre de données d'adaptation, ce qui signifie que la distribution de θ devient plus concentrée autour de sa moyenne.

2.2.2 Adaptation de la moyenne et de la précision

Comme précédemment, supposons que l'on dispose de N observations de durée spécifiques à un symbole, issues du corpus d'adaptation. Ces données sont des réalisations d'une variable aléatoire modélisée par une loi normale de moyenne M inconnue et de précision R inconnue, notée $\mathcal{N}(x; M, R)$. Notons $\theta = (M, R)$ le couple des paramètres inconnus à estimer.

La loi *a priori* conjuguée de θ, c.-à-d. la loi conjointe de M et R, correspond au produit de la loi conditionnelle de M lorsque R est fixé, par la loi marginale de R, définies comme suit. La loi conditionnelle de M, lorsque R est fixé ($R = r$), est une loi normale de moyenne μ et de précision τr fixes et connues ($\tau > 0$). La loi marginale de R est une loi Gamma de paramètres α et β, avec $\alpha > 0$, $\beta > 0$ [DeGroot, 1970].

Dans ces conditions, la loi conjointe *a posteriori* de M et R, connaissant les x_i, est le produit de la loi conditionnelle de M, lorsque $R = r$ est connu, par la loi marginale de R. La loi conditionnelle de M, R étant fixé à r, est une loi normale de moyenne μ' et de précision τ', avec [DeGroot, 1970] :

$$\mu' = \frac{\tau\mu + N\bar{x}}{\tau + N} \tag{11.6}$$

$$\tau' = (\tau + N)r \tag{11.7}$$

La loi marginale de R est une loi Gamma de paramètres α' et β', avec :

$$\alpha' = \alpha + \frac{N}{2} \tag{11.8}$$

$$\beta' = \beta + \frac{1}{2}\sum_{i=1}^{N}(x_i - \bar{x})^2 + \frac{\tau N(\bar{x} - \mu)^2}{2(\tau + N)} \tag{11.9}$$

Ce qui nous intéresse est d'obtenir une réestimation MAP pour le couple $\theta = (M, R)$. On a [Lee *et al.*, 1991] :

$$\theta_{MAP} = (\mu', \alpha'/\beta') \tag{11.10}$$

3 Expériences et résultats

3.1 Conditions expérimentales

Cette série d'expériences sur l'adaptation des modèles de durée utilise un corpus différent de celui de la partie III. L'application considérée est une tâche de reconnaissance de mots isolés en mode dépendant du locuteur, constitués de l'alphabet américain, des chiffres 0 à 9,

et de quelques mots de contrôle (*enter, erase, go, help, no, off, on, repeat, right, rubout, start, stop, yes*). Le corpus de parole est prononcé par 4 locutrices et 4 locuteurs américains. Les données d'apprentissage sont obtenues à partir de 2 répétitions de la totalité du vocabulaire dans des conditions calmes. Le corpus de test est constitué de 2 répétitions du vocabulaire, prononcées par chacun des locuteurs écoutant un bruit blanc de 85 dB SPL[2] par l'intermédiaire d'un casque audio. Les données de test sont ainsi perturbées par l'effet Lombard, mais le bruit utilisé pour contraindre le locuteur n'est pas enregistré. La totalité de la base de données, propre et Lombard, a été étiquetée manuellement par un expert, au niveau phonétique, ce qui fournit la durée de chaque symbole phonétique.

Contrairement à toutes les autres expériences présentées dans ce document, les tests de reconnaissance ont été effectués en utilisant la version 2 du système VINICS (février 93). Les signaux de parole ont été échantillonnés à 10 kHz, et un paramétrage MFCC d'ordre 12 a été appliqué toutes les 10 ms, dans des fenêtres de Hamming de 25.6 ms. Les modèles acoustiques de phones sont construits à partir du corpus d'apprentissage de parole propre, en utilisant l'algorithme LBG [Linde *et al.*, 1980].

Le problème associé à ces expériences réside dans la faible quantité de données disponibles. Les modèles étant sous-entraînés, nous considérons dans cette évaluation qu'un mot est correctement reconnu s'il se situe dans les 3 premières propositions fournies par VINICS. De la même façon, nous ne disposons pas d'un corpus spécifique de parole Lombard pour effectuer l'adaptation des modèles de durée. Par conséquent, nous utilisons comme données d'adaptation différentes fractions du corpus de parole Lombard. Bien évidemment cela biaise le test, car l'information de durée utilisée pour effectuer l'adaptation correspond aux durées rencontrées lors du test. Cette configuration reflète la situation idéale où les durées moyennes des symboles d'adaptation correspondent à la durée moyenne des durées des symboles du corpus de test, et fournit donc une borne supérieure de l'amélioration des performances que l'on peut espérer obtenir uniquement en adaptant la durée. Il convient de se souvenir que la différence entre parole Lombard et normale se situe principalement au niveau spectral ; ces tests ont donc pour objectif de montrer si une durée correctement modélisée permet d'améliorer les taux de reconnaissance en parole Lombard, et si l'approche Bayésienne est appropriée pour effectuer la réestimation des paramètres des modèles de durée.

Aux paragraphes 2.2.1 et 2.2.2, nous avons considéré que les paramètres des lois *a priori* de θ étaient connus. En pratique, il est possible d'estimer ces paramètres à partir d'un ensemble de modèles spécifiques à chaque locuteur. Traitons le cas de l'adaptation de la moyenne seule. Pour chaque symbole, trois paramètres sont inconnus mais fixes : d'une part les 2 paramètres de la loi *a priori* de la moyenne θ, c.-à-d. la moyenne μ et la précision τ des moyennes des durées du symbole, d'autre part la précision r de la durée du symbole. Supposons que le corpus d'apprentissage de chaque locuteur soit divisé en M parties. Il est alors possible de calculer par MLE et pour chaque symbole, la moyenne μ_k de la durée du symbole dans chaque partie k. Soit ω_k le poids associé à chaque partie k : $\omega_k = 1/M$ si le corpus d'apprentissage de chaque symbole est décomposé en M parties égales. Nous choisissons d'estimer la moyenne μ et la variance $1/\tau$ de la moyenne des durées d'un symbole, par la

[2] *Sound Pressure Level*

moyenne et la variance empirique des durées moyennes :

$$\mu = \sum_{k=1}^{M} \omega_k \mu_k \tag{11.11}$$

$$1/\tau = \sum_{k=1}^{M} \omega_k (\mu_k - \mu)^2 \tag{11.12}$$

Pour l'estimation de la variance de la durée r, nous avons utilisé deux approches distinctes. Dans la première, appelée Exp-1, $1/r$ est estimé de façon classique selon le critère MLE, par la variance des durées du symbole dans tout le corpus d'apprentissage. Dans la deuxième approche, appelée Exp-2, $1/r$ est estimé par la moyenne des variances $1/\tau_k$ des durées dans chaque classe :

$$1/r = \sum_{k=1}^{M} \omega_k \frac{1}{\tau_k} \tag{11.13}$$

Nous n'avons pas jugé opportun de tester l'adaptation simultanée de la moyenne et de la variance des durées. Le problème réside dans l'estimation des paramètres des lois *a priori*. En effet, il est nécessaire de diviser le corpus d'apprentissage en M parties, afin d'estimer les statistiques sur la moyenne et la variance des durées. Il faut donc disposer de suffisamment de données pour que les estimations dans une classe donnée soient robustes, et il est nécessaire de définir suffisamment de classes pour calculer les moyennes et variances de ces estimations. Étant donné la faible quantité de données d'adaptation dont nous disposons, il n'est pas possible d'assurer une estimation de bonne qualité des paramètres *a priori* de la variance des durées, ce qui explique ce choix.

3.2 Résultats

Les résultats présentés dans ce paragraphe ne concernent que l'adaptation de la moyenne de la loi normale modélisant les durées des symboles. L'adaptation est effectuée en utilisant différentes quantités de données d'adaptation. Le corpus d'adaptation est divisé en 4 parties, et les tests couvrent donc 5 conditions différentes : pas d'adaptation (utilisation de 0% du corpus d'adaptation), utilisation de 25, 50, 75 et 100% du corpus d'adaptation.

Les taux de reconnaissance moyennés sur les 4 locuteurs sont représentés fig. 11.1, en fonction de la quantité de données d'adaptation utilisée, pour les 2 configuration de test Exp-1 et Exp-2. La même courbe, pour les 4 locutrices est présentés fig. 11.2. Les tableaux 11.1 et 11.2 donnent le détail des taux de reconnaissance pour chacun des locuteurs et locutrices, pour les configurations Exp-1 et Exp-2. Au regard des courbes fig. 11.1 et 11.2, il apparaît que la condition expérimentale Exp-2, qui est celle utilisée dans [Lee *et al.*, 1991] pour adapter les paramètres de HMMs continus, s'avère moins efficace que la condition Exp-1. Il est probable que le mauvais comportement de Exp-2 provienne du nombre insuffisant de données pour estimer la variance de la durée dans chacune des M classes. Dans Exp-1, la variance de la durée est estimée en calculant la variance empirique des durées du corpus d'apprentissage, où le nombre de données disponibles garantit une estimation de qualité suffisante.

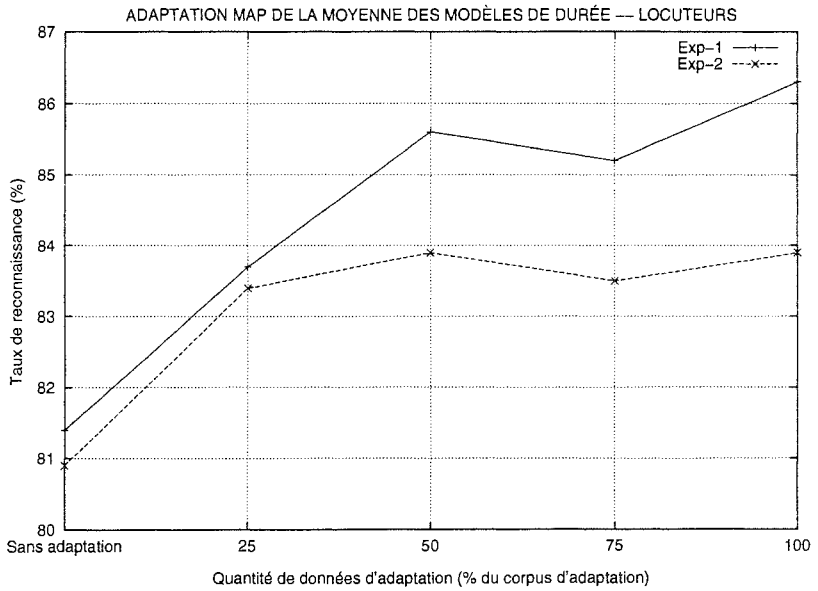

FIG. 11.1 - *Taux de reconnaissance* vs *quantité de données d'adaptation. Moyenne des 4 locuteurs.*

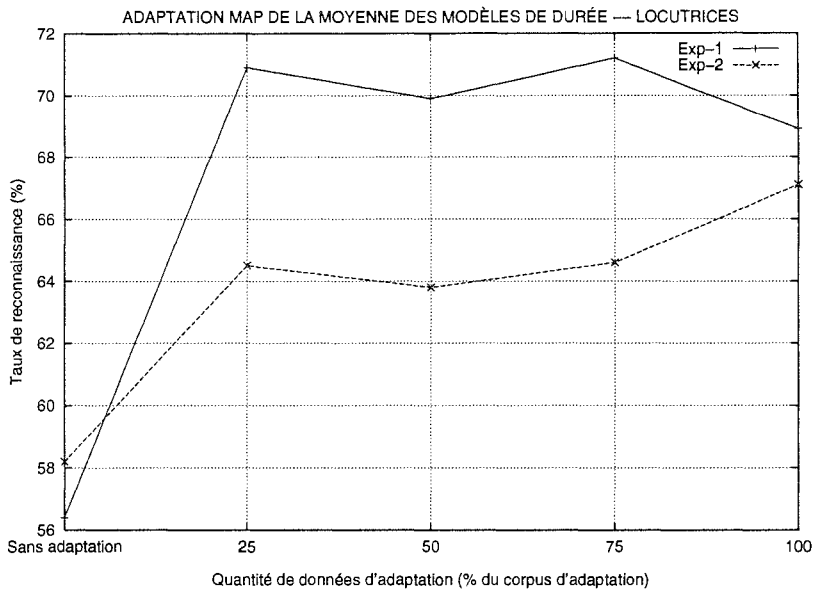

FIG. 11.2 - *Taux de reconnaissance* vs *quantité de données d'adaptation. Moyenne des 4 locutrices.*

Hommes	bah	dz	pf	sp
0%	78.6	79.6	85.7	81.6
25%	79.6	77.6	83.7	93.9
50%	81.6	80.6	87.8	92.9
75%	80.6	78.6	87.8	93.9
100%	82.7	79.6	88.8	93.9
Femmes	ac	ak	na	th
0%	38.8	77.6	58.2	51.0
25%	59.2	79.6	65.3	79.6
50%	58.2	79.6	64.3	77.6
75%	59.2	78.6	66.3	80.6
100%	58.2	75.5	64.3	77.6

TAB. 11.1 - *Taux de reconnaissance (%)* vs *quantité de données d'adaptation. Expérience Exp-1.*

Hommes	bah	dz	pf	sp
0%	82.7	77.6	78.6	84.7
25%	80.6	75.5	85.7	91.8
50%	82.7	74.5	85.7	92.9
75%	82.7	68.4	88.8	93.9
100%	82.7	69.4	89.8	93.9
Femmes	ac	ak	na	th
0%	44.9	72.4	54.1	61.2
25%	56.1	71.4	59.2	71.4
50%	55.1	71.4	59.2	69.4
75%	57.1	73.5	58.2	69.4
100%	60.2	75.5	57.1	75.5

TAB. 11.2 - *Taux de reconnaissance (%)* vs *quantité de données d'adaptation. Expérience Exp-2.*

Sans adaptation des durées, les scores de reconnaissance des locutrices sont nettement inférieurs à ceux des locuteurs. L'adaptation de la durée provoque une forte remontée des taux de reconnaissance chez les femmes. Au mieux, le taux de reconnaissance des locutrices augmente d'environ 20%, alors que le taux de reconnaissance des locuteurs n'est amélioré que d'environ 6%. Lorsque le taux de reconnaissance est faible au départ (cas des locutrices), l'adaptation de la durée semble beaucoup plus efficace que lorsque le score est initialement élevé. De même, il faut noter que l'amélioration des performances en fonction du volume de données d'adaptation est beaucoup plus régulière chez les hommes que chez les femmes : le gain de performance est obtenu chez les locutrices dès le début de l'adaptation (utilisation de 25% du corpus d'adaptation). En moyenne, l'adaptation des durées permet d'améliorer de façon significative les taux de reconnaissance de la parole Lombard, lorsque les modèles sont construits à partir de parole propre.

Cependant, d'après les tableaux 11.1 et 11.2, nous observons une forte variabilité des résultats entre les différents locuteurs, hommes et femmes confondus. L'étude de la variation de la durée des phonèmes en présence d'effet Lombard, réalisée dans [Junqua, 1993], montre que la modification de la durée est plus dépendante du locuteur chez les femmes que chez les hommes. Nous constatons que la variabilité des taux de reconnaissance est également plus marquée chez les femmes que chez les hommes. En définitive, il faut néanmoins constater que pour tous les locuteurs testés, l'adaptation de la durée permet d'améliorer les taux de reconnaissance, ce qui est également le cas pour 3 des 4 locutrices.

4 Conclusion

Dans ce chapitre, nous avons proposé d'adapter les paramètres des modèles de durée des phonèmes, entraînés à partir de parole propre, dans le but de reconnaître de la parole Lombard. Cette adaptation est motivée par le fait que l'effet Lombard provoque une modification de la durée des sons, par rapport à des conditions classiques d'élocution. Cette variation étant dépendante du locuteur, il paraît intéressant d'adopter une approche de compensation spécifique à chaque locuteur, qui exploite les données issues d'un corpus d'adaptation de parole Lombard. La méthode d'adaptation choisie se fonde sur l'apprentissage Bayésien, qui utilise à la fois une connaissance *a priori* sur la parole propre, ainsi que l'information provenant du corpus d'adaptation de parole Lombard.

Il est important de noter que dans les expériences réalisées, les modèles acoustiques utilisés lors de la reconnaissance de parole Lombard étaient des modèles de parole propre. Les résultats des expériences montrent d'une part que l'utilisation d'un modèle de durée correctement entraîné, c.-à-d. représentatif du phénomène considéré, permet d'améliorer de façon significative les taux de reconnaissance. Prendre en compte la durée des symbole est donc essentiel en RAP, et s'avère efficace lorsque les espaces acoustiques de test et de référence sont distants (cas des locutrices dans nos expériences). D'autre part, afin de lutter contre les variations de durée entre parole propre et Lombard, il apparaît que l'apprentissage Bayésien fournit un cadre approprié pour la réestimation des paramètres des modèles.

Cependant, l'estimation Bayésienne présente plusieurs contraintes. Il est nécessaire de modéliser correctement la connaissance *a priori*, ce qui nécessite un corpus d'apprentissage

de grande taille. De plus, le processus d'adaptation lui-même nécessite l'utilisation d'un corpus d'adaptation volumineux ; il n'est en particulier pas question d'effectuer une réestimation Bayésienne à partir de quelques secondes de parole. Cette raison explique que nous n'ayons pas recherché à adapter également les modèles acoustiques des symboles dans cette application. Ainsi, si l'adaptation des modèles acoustiques par la régression linéaire présentée partie III, pouvait être effectuée à partir de 20 secondes de parole. Il est improbable qu'une telle quantité de données d'adaptation puisse être utilisée dans le cadre d'une réestimation Bayésienne.

Conclusion et perspectives

Les variations de l'environnement acoustique entre l'apprentissage et le test d'un système de reconnaissance automatique de la parole constituent une source de dégradation des taux de reconnaissance, et sont le principal obstacle au développement d'applications de reconnaissance de parole. Ce problème représente donc un thème privilégié d'études dans la communauté « reconnaissance de parole ». Dans ce travail, nous nous sommes intéressés plus particulièrement à l'amélioration de la robustesse d'un système de reconnaissance de la parole continue, lorsque les perturbations de l'environnement sont provoquées par un bruit acoustique, modélisable de façon additive.

Dans une première partie, nous avons dressé un bilan des principales approches permettant d'améliorer la robustesse au bruit des systèmes de RAP, et identifié trois grandes familles de méthodes. Il est tout d'abord possible de traiter le signal de parole bruitée afin d'atténuer les effets du bruit, pour utiliser par la suite un système de parole entraîné dans des conditions calmes. En second lieu, un système de reconnaissance de parole propre peut être modifié pour tenir compte de la présence d'un bruit lors de la reconnaissance. Enfin, définir un paramétrage du signal de parole robuste au bruit permet de minimiser l'influence d'une variation du niveau du bruit entre les phases de test et d'apprentissage des modèles. Au regard de cette étude, nous avons développé certaines voies de recherches.

Dans la partie III, nous avons proposé puis comparé expérimentalement trois approches sur une même application, en utilisant comme plate-forme de test le système VINICS, décrit partie II. Les tests ont été effectués à partir d'un corpus de parole propre, artificiellement bruité avec différents types et niveaux de bruits, sur une tâche de reconnaissance de parole continue.

Dans un premier temps, nous avons travaillé sur une méthode de combinaison de modèles, dont l'objectif est de fournir une approximation de modèles stochastiques de trajectoires de parole bruitée, à partir de STMs de parole propre et d'un HMM de bruit. L'introduction de plusieurs hypothèses a été nécessaire pour mettre en œuvre cette approche. D'une part, nous avons supposé que les signaux de parole et de bruit sont indépendants et additifs dans le domaine des densités spectrales de puissance, hypothèse qui modélise un grand nombre d'environnements de parole bruitée. D'autre part, nous avons considéré que la somme de deux variables aléatoires de distribution log-normale est également une variable aléatoire log-normale, ce qui s'est avéré satisfaisant lors des expériences. Une telle approche de combinaison de modèles présente, selon notre thèse, plusieurs intérêts. D'une part, elle permet d'éviter d'entraîner un modèle à partir d'un corpus d'apprentissage de parole bruitée, solution lourde et rarement applicable en pratique. Ensuite, d'après son principe, appliquer cette méthode en absence de bruit ne provoque pas une dégradation des performances par rapport à l'utilisation des modèles de parole propre initiaux. Les expériences systématiques effectuées en présence de différents types de bruits ont confirmé cette thèse. De plus, dès que les modèles de parole propre et de bruit sont disponibles, cette approche s'avère peu coûteuse en

calculs. Il faut cependant remarquer que cette méthode, qui se base sur l'utilisation d'un modèle de bruit, nécessite de pouvoir détecter les zones d'activité vocale dans le signal bruité, par exemple dans le but d'estimer le modèle de bruit pendant les pauses du locuteur. En pratique, nous n'avons pas considéré ce problème et avons toujours estimé les modèles de bruit à partir de signaux de bruit correctement localisés et ne comportant pas de parole. La qualité du modèle de bruit contribuant fortement à la qualité des modèles de parole bruitée obtenus, développer des détecteurs robustes d'activité vocale est une perspective de travail. Enfin, cette méthode peut potentiellement être appliquée pour reconnaître de la parole corrompue par des bruits non stationnaires, tâche particulièrement difficile lorsque l'acquisition du signal est effectuée avec un seul microphone. Nous n'avons pas procédé à une évaluation systématique de la combinaison de modèles sur de tels bruits, et ce travail devra être réalisé dans le futur.

L'objectif de la deuxième méthode proposée était de calculer des estimateurs de la parole propre dans le domaine cepstral, avec minimisation des distorsions dans le domaine du logarithme de la densité spectrale de puissance. Les motivations étaient d'une part d'appliquer un filtrage du signal spécifique aux différents sons, car un bruit stationnaire ne perturbe pas le signal de parole de façon consistante, d'autre part de minimiser les distorsions dans un domaine relié à celui de la perception humaine. Comme dans l'approche précédente, nous nous sommes heurté aux problèmes posés par la nature complexe de la combinaison entre la parole et le bruit dans le domaine du logarithme du spectre. C'est pourquoi, la détermination des estimateurs a été effectuée selon une approche numérique, très coûteuse en calcul, qui rend l'approche difficilement applicable en pratique. La seule hypothèse utilisée portait sur la nature de la combinaison entre parole et bruit, mais aucune hypothèse ne conditionnait la nature des distributions des signaux de parole ou de bruit, contrairement à la méthode précédente. Étant donné les performances obtenues et la charge de calculs nécessaire pour déterminer les estimations, nous ne poursuivrons pas de travaux sur cette approche.

La dernière méthode de la partie III consistait à rechercher un ensemble de transformations linéaires définies de façon optimale selon un critère objectif, dans le but de transformer soit les modèles de parole propre, soit la parole bruitée. Une de nos motivations était encore d'appliquer des transformations spécifiques à des classes de sons. Cette méthode a été mise en œuvre selon une approche supervisée, qui consiste a exploiter les informations sur l'environnement de test contenues dans un corpus d'adaptation étiqueté de taille réduite. Un compromis a été adopté entre la précision des transformations estimées et la quantité de données d'adaptation qui permet de déterminer de façon robuste ces transformations. La seule hypothèse introduite supposait que les différences d'environnement entre le test et l'apprentissage pouvaient être modélisées par des transformations linéaires par parties, ce qui couvre un grand nombre de situations. D'après notre thèse, cette méthode s'avère plus efficace que les deux approches précédentes et que la transformation de base de [Gong, 1993]. En exploitant un corpus d'adaptation de 20 s de parole bruitée, nous sommes parvenus à construire des modèles de parole bruitée permettant d'obtenir des performances équivalentes, voire supérieures, à celles obtenues par des modèles directement entraînés à partir de parole bruitée, pour des rapports signal-à-bruit supérieurs ou égaux à 6 dB. De plus, l'avantage de cette méthode par rapport aux précédentes est qu'elle est potentiellement applicable pour prendre en compte des types de perturbations autres que le bruit additif, comme par exemple des variations de locuteurs ou de la ligne de transmission du signal. Évaluer cette approche sur de

telles applications est à court terme une perspective de travail.

Malheureusement, cette approche présente l'inconvénient de nécessiter un corpus d'adaptation étiqueté au niveau phonétique. Nous souhaitons donc porter nos efforts sur la suppression de cette contrainte, afin de développer une approche d'adaptation de modèles totalement non supervisée. Pour cela, nous pensons qu'il est possible de modifier l'algorithme d'estimation des matrices de transformation, afin de rechercher non plus les transformations qui maximisent la probabilité d'observer le corpus d'adaptation, mais plutôt les transformations et la segmentation qui maximisent la probabilité du corpus d'adaptation étiqueté. Il est envisageable d'effectuer une telle maximisation sous une forme itérative proche de l'algorithme de *segmental k-means* [Lee *et al.*, 1991], où une transformation est d'abord estimée en supposant une segmentation connue, puis une segmentation est déterminée en utilisant la transformation estimée, et ainsi de suite jusqu'à convergence. De plus, nous pensons qu'il est possible d'utiliser les scores de reconnaissance des N premières phrases, fournis par le système, afin de prédire si la première phrase proposé par le système est correctement reconnue ou non. Une telle information peut être utilisée, à terme, pour modifier les paramètres des modèles acoustiques dès qu'une phrase est présumée correctement reconnue, et cela sans intervention du locuteur pour valider ou non le résultat de la reconnaissance.

Dans la partie IV, nous avons proposé deux approches, l'une visant à définir un paramétrage du signal de parole robuste au bruit, l'autre à adapter des modèles de durée de phonème par un processus de réestimation Bayésienne, afin de lutter contre les modifications du rythme d'élocution provoqué par l'effet Lombard.

Selon nos travaux, l'utilisation d'une procédure d'analyse linéaire discriminante permet de définir un paramétrage du signal plus efficace que le paramétrage MFCC, lorsque les conditions bruitées de test et d'apprentissage sont semblables. Cependant, notre thèse est que la robustesse de ce paramétrage aux variations du rapport signal-à-bruit entre les conditions de test et d'apprentissage, est fortement conditionnée par la nature du bruit perturbateur. De plus, nous mettons en évidence l'influence des choix de mise en œuvre sur les performances, tels que la définition de la notion de classe utilisée pour le calcul des matrices de covariances inter et intra classes, ou encore l'utilisation ou non des modèles de silences pour déterminer la transformation discriminante. Étant donné le bon comportement du paramétrage par LDA lorsque les environnements de test et d'apprentissage sont semblables, il semble intéressant d'associer la méthode d'adaptation de modèles à un paramétrage LDA, dans le but de minimiser la sensibilité aux variations du rapport signal-à-bruit.

Enfin, nous avons proposé d'appliquer le cadre de l'estimation Bayésienne pour adapter des modèles de durée des phonèmes, initialement entraînés à partir de parole propre, pour permettre leur utilisation en parole Lombard. Les modèles de durée correctement estimés permettent d'améliorer les taux de reconnaissance en parole Lombard. Ici encore, il est envisageable d'associer cette méthode d'adaptation des modèles de durée avec une adaptation des modèles acoustiques par régression linéaire.

Dans ce travail, nous nous sommes plus particulièrement intéressé aux distorsions provoquées par un bruit additif, dans des conditions simulées. À l'heure actuelle, il est important de développer des travaux dans un cadre moins restrictif, à partir de parole enregistrée dans des conditions réelles, afin de considérer un ensemble plus large de variabilités du signal. Ainsi, on considère généralement qu'un utilisateur est réticent à l'emploi d'un microphone casque,

et souhaite éviter de tenir en main un microphone. Les problèmes liées à l'utilisation d'un microphone distant, tels que les bruits de fond (stationnaires ou non), les réverbérations, les interférences provoquées par d'autres locuteurs doivent donc être considérés. De même, la demande de systèmes de reconnaissance de parole au travers le réseau téléphonique est très forte, et nécessite de traiter les problèmes d'échos, de bruits, et des distorsions non linéaires provenant de la ligne téléphonique. L'étude de tels problèmes constitue notre thématique de recherche à long terme.

Bibliographie

[Acero et Stern, 1990] A. Acero et R. M. Stern. Environmental robustness in automatic speech recognition. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pages 849–852, Albuquerque, New Mexico, Avril 1990. ICASSP'90.

[Acero et Stern, 1991] A. Acero et R. M. Stern. Robust speech recognition by normalisation of the acoustic space. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pages 893–896, Toronto, Canada, 1991. ICASSP'91.

[Acero et Stern, 1992] A. Acero et R. M. Stern. Cepstral normalization for speech recognition. Dans *ESCA Workshop Proceedings of Speech Processing in Adverse Conditions*, pages 89–92, Cannes-Mandelieu, France, 1992.

[Acero, 1992] A. Acero. *Acoustical and environmental robustness in automatic speech recognition.* Kluver Academic Publishers, 1992.

[Afify *et al.*, 1994] M. Afify, Y. Gong, et J.-P. Haton. Non-linear time alignment in stochastic trajectory models for speech recognition. Dans *Proc. Int. Conf. on Spoken Language Processing*, Yokohama, Japan, 1994. ICSLP'94.

[Ahmed, 1989] M. S. Ahmed. Comparison of noisy speech enhancement algorithms in terms of LPC perturbation. *IEEE trans. on ASSP*, 37(1):121–125, Janvier 1989.

[Aikawa et Saito, 1994] K. Aikawa et T. Saito. Noise robust speech recognition using a dynamic-cepstrum. Dans *Proc. Int. Conf. on Spoken Language Processing*, volume 3, pages 1579–1582, Yokohama, Japan, Septembre 1994.

[Alexandre *et al.*, 1993] P. Alexandre, J. Boudy, et P. Lockwood. Root homomorphic deconvolution schemes for speech processing in car noise environments. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, volume 2, pages 99–102, Minneapolis, Minnesota, USA, 1993. ICASSP'93.

[Alexandre et Lockwood, 1993] P. Alexandre et P. Lockwood. Root-cepstral Analysis: A unified view. *Speech Communication*, 12, 1993.

[Alexandre, 1993] P. Alexandre. *Reconnaissance de la parole en milieu bruité. Application à la commande vocale d'un radio-téléphone de voiture.* Thèse de doctorat, Université de Rennes I, Décembre 1993.

[Anastasakos *et al.*, 1994] A. Anastasakos, F. Kubala, J. Makhoul, et R. Schwartz. Adaptation to new microphones using tied-mixture normalization. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, volume 1, pages 433–436, Adelaide, Australia, 1994. ICASSP'94.

[Anglade *et al.*, 1993] Y. Anglade, D. Fohr, et J.-C. Junqua. Speech Discrimination in Adverse Conditions Using Acoustic Knowledge and Selectively Trained Neural Networks. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, volume 2, pages 279–282, Minneapolis, Minnesota, USA, 1993. ICASSP'93.

[Applebaum et Hanson, 1991] T. H. Applebaum et B. A. Hanson. Regression features for recognition of speech in quiet and in noise. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pages 985–988, Toronto, Canada, 1991. ICASSP'91.

[Arslan et Hansen, 1994] L. M. Arslan et J. H. L. Hansen. Minimum cost based phoneme class detection for improved iterative speech enhacement. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, volume 2, pages 45–48, Adelaide, Australia, 1994. ICASSP'94.

[Atal, 1972] B. S. Atal. Automatic Speaker Recognition Based on Pitch Contours. *Journal of the Acoustical Society of America*, 52(6):1687–1697, 1972.

[Atal, 1974] B. Atal. Effectiveness of Linear Prediction Characteristics of the Speech Wave for Automatic Speaker Identifition and Verification. *Journal of the Acoustical Society of America*, 55:1304–1312, 1974.

[Aubert *et al.*, 1994] X. Aubert, C. Dugast, H. Ney, et V. Steinbiss. Large Vocabulary, Continuous Speech Recognition of Wall Street Journal Corpus. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, volume 2, pages 129–132, Adelaide, Australia, Avril 1994.

[Bahl *et al.*, 1983] L. R. Bahl, F. Jelinek, et R. L. Mercer. A Maximum Likelihood approach to continuous speech recognition. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 5(2):179–190, Mars 1983.

[Bahl *et al.*, 1987] L. R. Bahl, P. F. Brown, P. V. de Souza, et R. L. Mercer. Speech Recognition with Continuous-Parameter Hidden Markov Model. *Computer Speech and Language*, 2(3–4), 1987.

[Barbier et Chollet, 1991] L. Barbier et G. Chollet. Robust speech parameters extraction for word recognition in noise using neural networks. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pages 145–148, Toronto, Canada, Mai 1991. ICASSP'91.

[Barbier, 1992] L. Barbier. *Techniques neuronales pour l'adaptation à l'environnement d'un système de reconnaissance automatique de la parole.* Thèse de doctorat, Télécom Paris, 1992.

[Baudois *et al.*, 1989] D. Baudois, C. Serviere, et A. Silvent. Soustraction de bruit. Analyse et synthèse bibliographique. *Traitement du Signal*, 6(5), 1989.

[Beattie et Young, 1991] V. L. Beattie et S. Young. Noisy speech recognition using hidden Markov model state-based filtering. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pages 917–920, Toronto, Canada, 1991. ICASSP'91.

[Beattie et Young, 1992a] V. L. Beattie et S. J. Young. Hidden Markov Model State-Based Noise Cancellation. Rapport Technique F-INFENG/TR 92, Cambridge University Engineering Department, Février 1992.

[Beattie et Young, 1992b] V. L. Beattie et S. J. Young. Hidden Markov Model stated-based cepstral noise compensation. Dans *Proc. Int. Conf. on Spoken Language Processing*, pages 519–522, Banff, Alberta, Canada, 1992. ICSLP'92.

[Bellagarda *et al.*, 1992] J. R. Bellagarda, P. V. De Souza, A. J. Nadas, D. Nahamoo, M. A. Picheny, et L. R. Bahl. Robust speaker adaptation using a piecewise linear acoustic mapping. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, volume 1, pages 445–448, San Francisco, California, Mars 1992. ICASSP'92.

[Berouti *et al.*, 1979] M. Berouti, B. Schwartz, et J. Makhoul. Enhancement of speech corrupted by acoustic noise. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pages 208–211. ICASSP'79, Avril 1979.

[Berstein et Shallom, 1991] A. D. Berstein et I. D. Shallom. An Hypothesized Wiener Filtering Approach to Noisy Speech Recognition. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pages 913–916, Toronto, Canada, 1991.

[Boll et Pulsipher, 1980] S. F. Boll et D. C. Pulsipher. Suppression of acoustic noise in speech using two microphone adaptive noise cancellation. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 28:752–753, Décembre 1980.

[Boll, 1979] S. F. Boll. Suppression of acoustic noise in speech using spectral subtraction. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 27:113–120, 1979.

[Botte *et al.*, 1989] M. C. Botte, G. Canévet, L. Demany, et C. Sorin. *Psychoacoustique et perception auditive.* INSERM Editions midicales internationales, Paris Cachan, 1989.

[Bou-Ghazale et Hansen, 1994] S. E. Bou-Ghazale et J. H. L. Hansen. Duration and spectral based stress token generation for HMM speech recognition under stress. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, volume 1, pages 413–416, Adelaide, Australia, 1994. ICASSP'94.

[Bridle *et al.*, 1984] J. S. Bridle, K. M. Ponting, M. D. Brown, et A. W. Borrett. A noise compensating spectrum distance measure applied to automatic speech recognition. Dans *Proc. of the Institute of Acoustics*, Windermere, UK, Novembre 1984. Autumn conference.

[Cairns et Hansen, 1992] D. A. Cairns et J. H. L. Hansen. ICASRUS : An Mwave Based Real-time Speech Recognition System in Noise and Lombard Effect. Dans *Proc. Int. Conf. on Spoken Language Processing*, pages 703–706, Banff, Alberta, Canada, 1992. ICSLP'92.

[CALLIOPE, 1989] CALLIOPE. Perception auditive et perception de parole. Dans *La parole et son traitement automatique*, chapitre 5, pages 147–214. Masson, Paris, 1989.

[Carlson et Clements, 1991] B. A. Carlson et M. A. Clements. Application of a weighted projection measurement for robust hidden Markov model based speech recognition. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pages 921–924, Toronto, Canada, 1991. ICASSP'91.

[Carlson et Clements, 1992] B. A. Carlson et M. A. Clements. Speech recognition in noise using a projection-based likelihood measure for mixture density HMM's. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pages 237–240, San Francisco, California, 1992. ICASSP'92.

[Chang et Zue, 1994] J. Chang et V. Zue. A Study of Speech Recognition System Robustness to Microphone Variations: Experiments in Phonetic Classification. Dans *Proc. Int. Conf. on Spoken Language Processing*, volume 3, pages 995–998, Yokohama, Japan, Septembre 1994.

[Chen, 1988] Y. Chen. Cepstral Domain Talker Stress Compensation for Robust Speech Recognition. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 36(4):433–439, Avril 1988.

[Cheng et O'Shaughnessy, 1991] Y. M. Cheng et D. O'Shaughnessy. Speech enhancement based conceptually on auditory evidence. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 39(9):1943–1954, 1991.

[Chollet *et al.*, 1990] G. Chollet, K. Choukri, et J.-P. Tubach. Experiments in Speech Analysis and Recognition: Tackling the variability of Speech. *Advances in Speech, Hearing and Language Processing*, 1:79–105, 1990.

[Chou *et al.*, 1992] W. Chou, B. H. Juang, et C. H. Lee. Segmental gpd training of hmm based speech recognizer. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, volume 1, pages 473–476, San Francisco, California, 1992.

[Class *et al.*, 1990] F. Class, A. Kaltenmeier, P. Regel, et K. Trottler. Fast speaker adaptation for speech recognition systems. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pages 133–136, Albuquerque, New Mexico, Avril 1990. ICASSP'90.

[Cohen, 1985] J. Cohen. Application of an adaptive auditory model to speech recognition. *Journal of the Acoustical Society of America*, 78 (supplement)(1):S50(A), 1985.

[Cox et Malah, 1981] R. V. Cox et D. Malah. A Technique for Perceptually Reducing Periodically Structured Noise in Speech. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, 1981.

[Cung et Normandin, 1992] H. M. Cung et Y. Normandin. Noise adaptation algorithms for robust speech recognition. Dans *Proc. ESCA Workshop, Speech Processing in Adverse Conditions*, pages 171–174, Cannes-Mandelieu, France, Novembre 1992. ESCA.

[Cung et Normandin, 1993] H. M. Cung et Y. Normandin. Noise adaptation algorithms for robust speech recognition. *Speech Communication*, 12:267–276, 1993.

[Dabis et Wrench, 1991] H. S. Dabis et A. A. Wrench. An evaluation of adaptive noise cancelling for speech recognition. Dans *Proceedings of European Conference on Speech Communication and Technology*, pages 1301–1304. EUROSPEECH'91, 1991.

[Dal Degan et Prati, 1988] N. Dal Degan et C. Prati. Acoustic noise analysis and speech enhancement techniques for mobile radio applications. *Signal Processing*, 15:43–56, 1988.

[Das *et al.*, 1993] S. Das, R. Bakis, A. Nadas, D. Nahamoo, et M. Picheny. Influence of background noise and microphone on the performance of the IBM TANGORA speech recognition system. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, volume 2, pages 71–74, Minneapolis, Minnesota, USA, Avril 1993. ICASSP'93.

[Das *et al.*, 1994] S. Das, A. Nadas, D. Nahamoo, et M. Picheny. Adaptation techniques for ambience and microphone compensation in the IBM Tangora speech recognition system. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, volume 1, pages 21–24, Adelaide, Australia, 1994. ICASSP'94.

[Dautrich *et al.*, 1983a] B. A. Dautrich, L. R. Rabiner, et T. B. Martin. On the effects of varying filter bank parameters on isolated word recognition. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 31:793–806, 1983.

[Dautrich *et al.*, 1983b] B. A. Dautrich, L. R. Rabiner, et T. B. Martin. On the Use of Filter Bank Features for Isolated Word Recognition. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pages 1061–1064, Boston, USA, 1983.

[Davis et Mermelstein, 1980] S. B. Davis et P. Mermelstein. Comparison of Parametric Representation for Monosyllabic Word Recognition in Continuously Spoken Sentences. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 28(4):357–366, Août 1980.

[DeGroot, 1970] M. H. DeGroot. *Optimal Statistical Decisions*, chapitre 9, pages 155–189. McGraw-Hill, 1970.

[Dempster *et al.*, 1977] A. P. Dempster, N. M. Laird, et D. B. Rubin. Maximum Likelihood from incomplete data via the EM Algorithm. *Journal of Royal Statistical Society Ser. B*, 39:1–39, 1977.

[Diday *et al.*, 1982] E. Diday, J. Lemaire, J. Pouget, et F. Testu. *Éléments d'analyse de données*. Dunod, Paris, 1982.

[Digalakis *et al.*, 1992] V. V. Digalakis, M. Ostendorf, et J. R. Rohliceck. Fast algorithms for phone classification and recognition using segment–based models. *IEEE Transactions on Signal Processing*, 40(12):2885–2896, Décembre 1992.

[Dobler *et al.*, 1993] S. Dobler, D. Geller, R Haeb-Umbach, P. Meyer, H. Ney, et H. W. Ruehl. Design and use of speech recognition algorithms for a mobile radio telephone. *Speech Communication*, 12:221–229, 1993.

[Doddington, 1989] G. R. Doddington. Phonetically sensitive discriminants for improved speech recognition. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pages 556–559, Glasgow, UK, Mai 1989. ICASSP'89.

[Drucker, 1968] H. Drucker. Speech processing in a high ambient noise environment. Dans *IEEE Trans. on Audio and Electroacoustics*, volume AU-16, pages 165–168, Juin 1968.

[Ephraim *et al.*, 1987] Y. Ephraim, J. G. Wilpon, et L. R. Rabiner. A linear predictive front-end processor for speech recognition in noisy environment. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pages 1324–1327, Dallas, Texas, USA, Avril 1987. ICASSP'87.

[Ephraim *et al.*, 1989] Y. Ephraim, D. Malah, et B. H. Juang. On the Application of Hidden Markov Models for Enhancing Noisy Speech. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 37(12):1846–1856, Décembre 1989.

[Ephraim et Malah, 1985] Y. Ephraim et D. Malah. Speech enhancement using a Minimum Mean-Square Error Log-Spectral amplitude estimator. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 33(2):443–445, 1985.

[Ephraim, 1992a] Y. Ephraim. A Bayesian Estimation Approach for Speech Enhancement Using Hidden Markov Models. *IEEE Transactions on Signal Processing*, 40(4):725–735, Avril 1992.

[Ephraim, 1992b] Y. Ephraim. Statistical-model-based speech enhancement systems. *Proc. of the IEEE*, 80(10):1526–1555, Octobre 1992.

[Erell et Weintraub, 1990] A. Erell et M. Weintraub. Estimation using log-spectral-distance criterion for noise-robust speech recognition. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, volume 2, pages 853–856, Albuquerque, New Mexico, Avril 1990. ICASSP'90.

[Erell et Weintraub, 1993a] A. Erell et M. Weintraub. Energy Condition Spectral Estimation for Recognition of Noisy Speech. *IEEE Transactions on Speech and Audio Processing*, 1(1):84–89, 1993.

[Erell et Weintraub, 1993b] A. Erell et M. Weintraub. Filterbank-energy estimation using mixture and Markov models recognition of noisy speech. *IEEE Transactions on Speech and Audio Processing*, 1(1), 1993.

[Erell et Weintraub, 1994] A. Erell et M. Weintraub. Estimation of Noise-Corrupted Speech DFT-Spectrum Using the Pitch Period. *IEEE Transactions on Speech and Audio Processing*, 2(1), 1994.

[Euler et Zinke, 1992] S. Euler et J. Zinke. Experiments on the use of the generalized probabilistic descent method in speech recognition. Dans *Proc. Int. Conf. on Spoken Language Processing*, volume 1, pages 157–160, Banff, Alberta, Canada, Octobre 1992.

[Feder *et al.*, 1987] M. Feder, A. V. Oppenheim, et E. Weinstein. Methods for noise cancellation based on the EM algorithm. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Dallas, Texas, USA, 1987. ICASSP'87.

[Feder *et al.*, 1988] M. Feder, E. Weinstein, et A. V. Oppenheim. A new class of sequential and adaptive algorithms with application to noise cancellation. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pages 557–560, 1988.

[Feder *et al.*, 1989] M. Feder, A. V. Oppenheim, et E. Weinstein. Maximum likelihood noise cancellation using the EM Algorithm. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 37(2):204–216, 1989.

[Fellbaum et Becker, 1991] K. Fellbaum et D. Becker. Isolated word recognition with integrated noise reduction. Dans *Proceedings of European Conference on Speech Communication and Technology*, pages 87–90. EUROSPEECH'91, 1991.

[Fissore *et al.*, 1992] L. Fissore, P. Laface, G. Micca, et G. Sperto. Channel Adaptation for a Continuous Speech Recognizer. Dans *Proc. Int. Conf. on Spoken Language Processing*, volume 2, pages 1495–1498, Banff, Alberta, Canada, Octobre 1992.

[Fletcher, 1940] H. Fletcher. Auditory patterns. *Rev. Mod. Phys.*, 12:45–65, 1940.

[Frangoulis et Gaganelis, 1992] E. Frangoulis et D. A. Gaganelis. Adaptation of the HMM distribution: Application to a VQ codebook and to a noisy environment. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pages 489–492, San Francisco, California, 1992. ICASSP'92.

[Frangoulis et Sgardoni, 1991] E. Frangoulis et V. Sgardoni. A novel Speaker Adaptation approach for Continuous Densities HMM's. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pages 861–864, Toronto, Canada, Mai 1991. ICASSP'91.

[Frazier *et al.*, 1976] R. H. Frazier, S. Samsam, L. D. Braida, et A. V. Oppenheim. Enhancement of speech by adaptive filtering. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pages 251–253, Philadelphia, USA, Avril 1976. ICASSP'76.

[Furui, 1981] S. Furui. Cepstral Analysis Technique for Automatic Speaker Verification. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 29(2):254–272, 1981.

[Furui, 1986a] S. Furui. On the role of spectral transition for speech perception. *Journal of the Acoustical Society of America*, 80(4):1016–1025, 1986.

[Furui, 1986b] S. Furui. Speaker-independent isolated word recognition using dynamic features of speech spectrum. *IEEE trans. on ASSP*, 34(1):52–59, Février 1986.

[Furui, 1989a] S. Furui. Unsupervised speaker adaptation based on Hierarchical Spectral Clustering. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 37(12):1923–1930, Décembre 1989.

[Furui, 1989b] S. Furui. Unsupervised speaker adaptation based on Hierarchical Spectral Clustering. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pages 286–289. ICASSP'89, 1989.

[Furui, 1990] S. Furui. On the use of hiererchical spectral dynamics in speech recognition. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pages 789–792, Albuquerque, New Mexico, Avril 1990. ICASSP'90.

[Furui, 1992a] S. Furui. Recent advances in speech recognition technology at NTT laboratories. *Speech Communication*, 11(2–3):195–204, Juin 1992.

[Furui, 1992b] S. Furui. Toward robust speech recognition under adverse conditions. Dans *Proc. ESCA workshop on Speech Processing in Adverse Conditions*, pages 31–42, Cannes-Mandelieu, France, Novembre 1992. ESCA.

[Gales et Young, 1992] M. J. F. Gales et S. Young. An improved approach to the hidden Markov model decomposition of speech and noise. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, volume 1, pages 233–236, San Francisco, California, Avril 1992.

[Gales et Young, 1993a] M. J. F. Gales et S. J. Young. Cepstral parameter compensation for HMM recognition in noise. *Speech Communication*, 12:231–239, 1993.

[Gales et Young, 1993b] M. J. F. Gales et S. J. Young. HMM recognition in noise using parallel model combination. Dans *Proceedings of European Conference on Speech Communication and Technology*, volume 2, pages 837–840, Berlin, Germany, 1993.

[Gales et Young, 1993c] M. J. F. Gales et S. J. Young. Parallel Model Combination for Speech Recognition in Noise. Rapport Technique F-INFENG/TR 135, CUED, Cambridge University Engineering Department, Juin 1993.

[Gales et Young, 1993d] M. J. F. Gales et S. J. Young. PMC for Speech Recognition in Additive and Convolutional Noise. Rapport Technique F-INFENG/TR 154, CUED, Cambridge University Engineering Department, Décembre 1993.

[Gales et Young, 1994a] M. F. J. Gales et S. J. Young. Parallel Model Combination on a Noise Corrupted Resource Management Task. Dans *Proc. Int. Conf. on Spoken Language Processing*, volume 1, pages 255–258, Yokohama, Japan, Septembre 1994. ICSLP'94.

[Gales et Young, 1994b] M. J. F. Gales et S. J. Young. Robust continuous speech recognition using parallel model combination. Rapport Technique F-INFENG/TR 172, CUED, Cambridge University Engineering Department, Mars 1994.

[Gao *et al.*, 1992a] Y. Gao, J.-P. Haton, et Y. Gong. Noisy speech recognition tested on continuous speech recognition system. Dans *ESCA Workshop: Speech Processing in Adverse Conditions*, Cannes-Mandelieu, France, Novembre 1992.

[Gao *et al.*, 1992b] Y. Gao, T. Huang, S. Chen, et J.-P. Haton. Auditory model based speech processing. Dans *Proc. Int. Conf. on Spoken Language Processing*, volume 1, pages 73–76, Banff, Alberta, Canada, 1992. ICSLP'92.

[Gao et Haton, 1993] Y. Gao et J.-P. Haton. Noise Reduction and Speech Recognition in Noise conditions tested on LPNN-Based Continuous Speech Recognition System. Dans *Proceedings of European Conference on Speech Communication and Technology*, volume 2, pages 1035–1038, Berlin, Germany, 1993. EUROSPEECH'93.

[Gao et Haton, 1994] Y. Gao et J.-P. Haton. A hierarchical LPNN network for noise reduction and noise degraded speech recognition. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, volume 2, pages 89–92, Adelaide, Australia, 1994. ICASSP'94.

[Gauvain et Lee, 1992] J.-L. Gauvain et C. H. Lee. Bayesian learning for hidden Markov model with Gaussian mixture state observation densities. *Speech Communication*, 11(2–3):205–213, Juin 1992.

[Gauvain et Lee, 1994] J.-L. Gauvain et C.-H. Lee. Maximum a Posteriori Extimation for Multivariate Gaussian Mixture Observations of Markov Chains. *IEEE Transactions on Speech and Audio Processing*, 2(2):291–298, Avril 1994.

[Ghitza, 1986] O. Ghitza. Auditory nerve representation as a front-end for speech recognition in a noisy environment. *Computer Speech and Language*, 1:109–130, 1986.

[Ghitza, 1987] O. Ghitza. Robustness against noise: the role of timing-synchrony measurement. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pages 2372–2375, Dallas, Texas, USA, 1987. ICASSP'87.

[Ghitza, 1988] O. Ghitza. Auditory neural feedback as a basis for speech recognition. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pages 91–94. ICASSP'88, 1988.

[Ghitza, 1992] O. Ghitza. Auditory Nerve Representation as a Basis for Speech Processing. Dans *Advances in Speech Signal Processing*, rédacteurs S. Furui et M. M. Sondhi, chapitre 15, pages 453–485. Marcel Dekker, Inc, New York, 1992.

[Gish *et al.*, 1990] H. Gish, Y.-L. Chow, et J. R. Rohlicek. Probabilistic Vector mapping of noisy speech parameters for HMM word spotting. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, volume 1, pages 117–120, Albuquerque, New Mexico, Avril 1990. ICASSP'90.

[Giuliani *et al.*, 1994] D. Giuliani, M. Omologo, et P. Svaizer. Talker localisation and speech recognition using a microphone array and a cross-powerspectrum phase analysis. Dans *Proc. Int. Conf. on Spoken Language Processing*, pages 1243–1246, Yokohama, Japan, Septembre 1994. ICSLP'94.

[Gomez-Mena *et al.*, 1991] J. Gomez-Mena, J. Santos-Suarez, et R. Garcia-Gomez. A robust feature extraction method for automatic speech recognition in noisy environments. Dans *Proceedings of European Conference on Speech Communication and Technology*, pages 1383–1386. EUROSPEECH'91, 1991.

[Gong *et al.*, 1991] Y. Gong, J.-P. Haton, et F. Mouria. Continuous speech recognition based on high plausibility regions. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pages 725–728, Toronto, Canada, Mai 1991. ICASSP'91.

[Gong *et al.*, 1992] Y. Gong, O. Siohan, et J.-P. Haton. Minimization of Speech Alignment Error by Iterative Transformation for Speaker Adaptation. Dans *Proc. Int. Conf. on Spoken Language Processing*, pages 377–380, Banff, Alberta, Canada, Octobre 1992. ICSLP'92.

[Gong *et al.*, 1994] Y. Gong, J.-P. Haton, et J.-F. Mari. Issues in acoustic modeling of speech for automatic speech recognition. Dans *Progress and Prospects of Speech Research and Technolog*, rédacteurs H. Niemann, R. De Mori, et G. Hanrieder. INFIX Sankt Augustin, 1994.

[Gong et Haton, 1991] Y. Gong et J.-P. Haton. VINICS: A continuous speech recognizer based on a new robust formulation. Dans *Proceedings of European Conference on Speech Communication and Technology*, volume 3, pages 1221–1224, Genova, Italy, Septembre 1991.

[Gong et Haton, 1993] Y. Gong et J.-P. Haton. Iterative Transformation And Alignment For Speech Labeling. Dans *Proceedings of European Conference on Speech Communication and Technology*, volume 3, pages 1759–1762, Berlin, Germany, Septembre 1993. EUROSPEECH'93.

[Gong et Haton, 1994] Y. Gong et J.-P. Haton. Stochastic Trajectory Modeling For Speech Recognition. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, volume 1, pages 57–60, Adelaide, Australia, Avril 1994. ICASSP'94.

[Gong, 1993] Y. Gong. Base transformation for environment adaptation in continuous speech recognition. Dans *Proceedings of European Conference on Speech Communication and Technology*, volume 3, pages 2227–2230, Berlin, Germany, Septembre 1993. EUROSPEECH'93.

[Gong, 1994] Y. Gong. Stochastic Trajectory Modeling and Sentence Searching for Continuous Speech Recognition. Rapport de recherche, CRIN – CNRS & INRIA Lorraine, 1994.

[Gong, 1995] Y. Gong. Speech Recognition in Noisy Environments: A Survey. *Speech Communication*, 16(3):261–291, Avril 1995.

[Graf et Hubing, 1993] J. T. Graf et N. Hubing. Dynamic Time-warping for the enhancement of speech degraded by white Gaussian noise. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, volume 2, pages 339–342, Minneapolis, Minnesota, USA, 1993. ICASSP'93.

[Gray *et al.*, 1980] R. M. Gray, A. Buzo, A. H. Gray, et Y. Matsuyama. Distorsion measures for speech processing. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 28:367–376, Août 1980.

[Griffiths et Jim, 1982] L. J. Griffiths et C. W. Jim. An alternative approach to linearly constrained adaptive beamforming. *IEEE Trans. Antennas Propag.*, 30:27–34, 1982.

[Gu et Mason, 1989] Y. Gu et J. S. Mason. Speaker normalization via a linear transformation on a perceptual feature space and its benefits in ASR adaptation. Dans *Proc. European Conf. on Speech Communication and Technology*, pages 258–261. Eurospeech-89, 1989.

[Guan *et al.*, 1993] C. Guan, Y. Chen, et B. Wu. Direct modification on LPC coefficients with application to speech enhancement and improving the performance of speech recognition in noise. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, volume 2, pages 107–110, Minneapolis, Minnesota, USA, Avril 1993. ICASSP'93.

[Haeb-Umbach *et al.*, 1993] R. Haeb-Umbach, D. Geller, et H. Ney. Improvements in Connected Digit Recognition Using Linear Discriminant Analysis and Mixture Densities. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, volume 2, pages 239–242, Minneapolis, Minnesota, USA, Avril 1993. ICASSP'93.

[Haeb-Umbach et Ney, 1992] R. Haeb-Umbach et H. Ney. Linear Discriminant Analysis for Improved Large Vocabulary Continuous Speech Recognition. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, volume 1, pages 13–16, San Francisco, California, Mars 1992. ICASSP'92.

[Hansen *et al.*, 1994] J. H. L. Hansen, B. D. Womack, et L. M. Arslan. A Source Generator Based Production Model for Environmental Robustness in Speech Recognition. Dans *Proc. Int. Conf. on Spoken Language Processing*, volume 3, pages 1003–1006, Yokohama, Japan, Septembre 1994.

[Hansen et Bria, 1990] J. H. L. Hansen et O. N. Bria. Lombard effect compensation for robust automatic speech recognition in noise. Dans *Proc. Int. Conf. on Spoken Language Processing*, pages 1125–1128, 1990.

[Hansen et Clements, 1985] J. H. L. Hansen et M. A. Clements. Enhancement of Speech Degraded by Non-White Additive Noise. Rapport Technique DSPL-85-6, Georgia Institute of Technology, Atlanta, Août 1985.

[Hansen et Clements, 1987] J. H. L. Hansen et M. A. Clements. Iterative speech enhancement with spectral constraints. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pages 189–192, Dallas, Texas, USA, Avril 1987. ICASSP'87.

[Hansen et Clements, 1988] J. H. L. Hansen et M. A. Clements. Constrained Iterative speech Enhancement with Application to Automatic Speech Recognition. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pages 561–564, New York, USA, 1988. ICASSP'88.

[Hansen et Clements, 1989] J. H. L. Hansen et M. A. Clements. Stress compensation and noise reduction algorithms for robust speech recognition. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pages 266–269. ICASSP'89, 1989.

[Hansen et Clements, 1991] J. H. L. Hansen et M. A. Clements. Constrained Iterative Speech Enhancement with Application to Speech Recognition. *IEEE Transactions on Signal Processing*, 39(4):795–805, Avril 1991.

[Hansen, 1991] J. H. L. Hansen. A New Speech Enhancement Algorithm Employing Acoustic Endpoint Detection and Morphological Based Spectral Constraints. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pages 901–904, Toronto, Canada, 1991. ICASSP'91.

[Hansen, 1993] J. H. L. Hansen. Adaptive source generator compensation and enhancement for speech recognition in noisy stressful environments. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, volume 2, pages 95–98, Minneapolis, Minnesota, USA, 1993. ICASSP'93.

[Hanson et Applebaum, 1990a] B. A. Hanson et T. H. Applebaum. Features for noise-robust speaker-independent word recognition. Dans *Proc. Int. Conf. on Spoken Language Processing*, pages 1117–1120, 1990.

[Hanson et Applebaum, 1990b] B. A. Hanson et T. H Applebaum. Robust speaker-independent word recognition using static, dynamic and acceleration features: experiments with Lombard and noisy speech. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pages 857–860, Albuquerque, New Mexico, Avril 1990. ICASSP'90.

[Hanson et Applebaum, 1993] B. A. Hanson et T. H. Applebaum. Suband or cepstral domain filtering for recognition of Lombard and channel-distorted speech. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, volume 2, pages 79–82, Minneapolis, Minnesota, USA, 1993. ICASSP'93.

[Hanson et Wakita, 1986a] B. A. Hanson et H. Wakita. Spectral Slope Based Distorsion Measures for All-pole Models of Speech. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pages 757–760, Tokyo, Japan, 1986. ICASSP'86.

[Hanson et Wakita, 1986b] B. A. Hanson et H. Wakita. Spectral slope distance measures with linear prediction analysis for word recognition in noise. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pages 741–744, Tokyo, Japan, Avril 1986.

[Hanson et Wakita, 1987] B. A. Hanson et H. Wakita. Spectral slope distance measures with linear prediction analysis for word recognition in noise. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 35(7):968–973, 1987.

[Haton, 1993] J.-P. Haton. Automatic recognition of noisy speech. Dans *Proc. NATO ASI on New Advances and Trends in Speech Recognition and Coding*, 1993.

[Hermansky *et al.*, 1985] H. Hermansky, B. A. Hanson, et H. Wakita. Perceptually Based Linear Predictive Analysis of Speech. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, volume 1, pages 509–512. ICASSP'85, 1985.

[Hermansky *et al.*, 1991] H. Hermansky, N. Morgan, A. Bayya, et P. Kohn. Compensation for the effect of the communication channel in auditory-like analysis of speech (RASTA-PLP). Dans *Proceedings of European Conference on Speech Communication and Technology*, pages 1367–1370. EUROSPEECH'91, 1991.

[Hermansky *et al.*, 1993] H. Hermansky, N. Morgan, et H.-G. Hirsch. Recognition of speech in additive and convolutional noise based on RASTA spectral processing. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, volume 2, pages 83–86, Minneapolis, Minnesota, USA, 1993. ICASSP'93.

[Hermansky et Morgan, 1992] H. Hermansky et N. Morgan. Towards Handling the Acoustic Environment in Spoken Language Processing. Dans *Proc. Int. Conf. on Spoken Language Processing*, volume 1, pages 85–88, Banff, Alberta, Canada, Octobre 1992. ICSLP'92.

[Hermansky, 1990] H. Hermansky. Perceptual linear predictive (PLP) analysis of speech. *Journal of the Acoustical Society of America*, 87(4):1738–1752, Avril 1990.

[Hernando *et al.*, 1994] J. Hernando, C. Nadeu, C. Villagrassa, et E. Monte. Speaker Identification in Noisy Conditions Using Linear Prediction of the One-Sided Autocorrelation Sequence. Dans *Proc. Int. Conf. on Spoken Language Processing*, volume 4, pages 1847–1850, Yokohama, Japan, Septembre 1994.

[Hernando et Nadeu, 1991] J. Hernando et C. Nadeu. A comparative study of parameters and distances for noisy speech recognition. Dans *Proceedings of European Conference on Speech Communication and Technology*, pages 91–94, Genova, Italy, 1991.

[Hernando et Nadeu, 1994] J. Hernando et C. Nadeu. Speech recognition in noisy car environment based on OSALPC representation and robust similarity measuring techniques. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, volume 2, pages 69–72, Adelaide, Australia, 1994. ICASSP'94.

[Hirsch *et al.*, 1991] H. G. Hirsch, P. Meyer, et H. W. Ruehl. Improved Speech Recognition using high-pass filtering of subband Enveloppes. Dans *Proceedings of European Conference on Speech Communication and Technology*, pages 413–416, Genova, Italy, Septembre 1991. EUROSPEECH'91.

[Hirsch et Ruhl, 1989] H. G. Hirsch et H. W. Ruhl. Automatic speech recognition in a noisy environment. Dans *Proc. European Conf. on Speech Communication and Technology*, pages 652–654, Paris, France, Septembre 1989. EuroSpeech'89.

[Holmes et Sedgwic, 1986] J. N. Holmes et N. C. Sedgwic. Noise compensation for speech recognition using probabilistic models. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pages 741–744, Tokyo, Japan, Avril 1986. ICASSP'86.

[Huang *et al.*, 1990] X. D. Huang, Y. Ariki, et M.A. Jack. *Hidden Markov Models for Speech Recognition.* Edinburgh University Press, 1990.

[Hunt *et al.*, 1991] M. J. Hunt, D. C. Bateman, S. M. Richardson, et P. Piau. An investigation of PLP and IMELDA acoustic representation and of their potential combination. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pages 881–884, Toronto, Canada, 1991. ICASSP'91.

[Hunt et Lefèbvre, 1986] M. J. Hunt et C. Lefèbvre. Speech recognition using a cochlear model. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, volume 3, pages 1979–1982, Tokyo, Japan, 1986.

[Hunt et Lefèbvre, 1987] M. J. Hunt et C. Lefèbvre. Speech recognition using an Auditory Model with Pitch-Synchronous Analysis. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pages 813–816, Dallas, Texas, USA, 1987.

[Hunt et Lefèbvre, 1988] M. J. Hunt et C. Lefèbvre. Speaker Dependent and Independent Speech Recognition Experiments with an Auditory Model. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pages 215–218, New York, USA, 1988.

[Hunt et Lefèbvre, 1989] M. J. Hunt et C. Lefèbvre. A comparison of several acoustic representations for speech recognition with degraded and undegraded speech. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pages 262–265. ICASSP'89, 1989.

[Imamura, 1991] A. Imamura. Speaker-adaptive HMM-based speech recognition with a stochastic speaker classifier. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pages 841–844, Toronto, Canada, 1991. ICASSP'91.

[Itakura et Saito, 1968] F. Itakura et S. Saito. An Analysis-Synthesis Telephony based on Maximum Likelihood Method. Dans *Proc. Int. Congr. Acoust.*, pages C–5–5, Tokyo, Japan, Août 1968.

[Itakura et Umezaki, 1987] F. Itakura et T. Umezaki. Distance Measure for Speech Recognition Based on the Smoothed Group Delay Spectrum. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pages 1257–1261, Dallas, Texas, USA, 1987. ICASSP'87.

[Itakura, 1975] F. Itakura. Minimum Prediction Residual Principle Applied to Speech Recognition. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 23, Février 1975.

[Jelinek et Mercer, 1980] F. Jelinek et R. L. Mercer. Interpolated estimation of Markov source parameters from sparse data. Dans *Patern Recognition in Practice*, rédacteurs E. S. Gelsema et L. N. Kanal, pages 381–397, 1980.

[Juang *et al.*, 1986] B. H. Juang, L. R. Rabiner, et J. G. Wilpon. On the use of bandpass lifting in speech recognition. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pages 765–768, Tokyo, Japan, 1986. ICASSP'86.

[Juang *et al.*, 1987] B. H. Juang, L. R. Rabiner, et J. G. Wilpon. On the use of bandpass lifting in speech recognition. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 35(7):947–954, 1987.

[Juang et Paliwal, 1992] B. H. Juang et K. K. Paliwal. Vector equalization in hidden Markov models for noisy speech recognition. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, volume 1, pages 301–304, San Francisco, California, 1992. ICASSP'92.

[Juang et Rabiner, 1985] B. H. Juang et L. R. Rabiner. Mixture Autoregressive Hidden Markov Models. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 33(6):1404–1413, Décembre 1985.

[Juang et Rabiner, 1987] B. H. Juang et L. R. Rabiner. Signal Restoration by Spectral Mapping. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pages 2368–2371, Dallas, Texas, USA, 1987. ICASSP'87.

[Juang, 1991] B. H. Juang. Speech recognition in adverse environments. *Computer Speech and Language*, 5(1):275–294, 1991.

[Junqua et Anglade, 1990] J.-C. Junqua et Y. Anglade. Acoustic and perceptual studies of Lombard speech: application to isolated-words automatic speech recognition. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pages 841–844, Albuquerque, New Mexico, 1990. ICASSP'90.

[Junqua et Wakita, 1989] J.-C. Junqua et H. Wakita. A comparative study of cepstral lifters and distance measures for all-pole models of speech in noise. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pages 476–479, Glasgow, UK, Mai 1989. ICASSP'89.

[Junqua, 1989] J.-C. Junqua. *Contribution à l'amélioration de la robustesse des systèmes de reconnaissance automatique de mots isolés*. Thèse de doctorat, Université de NANCY 1, 1989.

[Junqua, 1993] J.-C. Junqua. The Lombard reflex and its role on human listeners and automatic speech recognizers. *Journal of the Acoustical Society of America*, 93(1):510–524, Janvier 1993.

[Junqua, 1994] J.-C. Junqua. A Duration Study of Speech Vowels Produced in Noise. Dans *Proc. Int. Conf. on Spoken Language Processing*, volume 1, pages 419–422, Yokohama, Japan, Septembre 1994.

[Kadirkamanathan, 1992] M. Kadirkamanathan. Hidden Markov Model Decomposition recognition of speech in noise: A comprehensive experimental study. Dans *ESCA Workshop Proceedings of Speech Processing in Adverse Conditions*, pages 187–190, Cannes-Mandelieu, France, 1992.

[Kanal et Chandrasekaran, 1971] L. N. Kanal et B. Chandrasekaran. On dimensionality and sample size in statistical pattern classification. *Pattern Recognition*, 3:225–234, Octobre 1971.

[Kitamura *et al.*, 1992] T. Kitamura, S. Ando, et E. Hayahara. Speaker-independent spoken digit recognition in noisy environments using dynamic spectral features and neural networks. Dans *Proc. Int. Conf. on Spoken Language Processing*, volume 1, pages 699–702, Banff, Alberta, Canada, Octobre 1992.

[Klatt, 1976] D. H. Klatt. A Digital Filter-bank for Spectral Matching. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pages 573–576, Philadelphia, USA, 1976. ICASSP'76.

[Kobatake et Matsunoo, 1994] H. Kobatake et Y. Matsunoo. Degraded word recognition based on segmental signal-to-noise ratio weighting. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, volume 1, pages 425–428, Adelaide, Australia, 1994. ICASSP'94.

[Kobayashi *et al.*, 1994] T. Kobayashi, R. Mine, et K. Shirai. Markov model based noise modeling and its application to noisy speech recognition using dynamical features of speech. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, volume 2, pages 57–60, Adelaide, Australia, 1994. ICASSP'94.

[Kobayashi et Imai, 1984] T. Kobayashi et S. Imai. Spectral Analysis Using Generalised Cepstrum. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 32(5), 1984.

[Koehler *et al.*, 1994] J. Koehler, N. Morgan, H. Hermansky, H. G. Hirsch, et G. Tong. Integrating RASTA-PLP into speech recognition. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, volume 1, pages 421–424, Adelaide, Australia, 1994. ICASSP'94.

[Kushner *et al.*, 1989] W. S. Kushner, V. Goncharoff, C. Wu, V. Nguyen, et J. N. Damoulakis. The effect of Subtractive-type Speech Enhancement/Noise Reduction Algorithms on Parameter Estimation for Improved Recognition and Coding in hign Noise Environment. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, volume 1, pages 211–214, Glasgow, UK, 1989. ICASSP'89.

[Le Bouquin, 1991] R. Le Bouquin. *Traitement pour la réduction du bruit sur la parole, application aux communications radio-mobiles*. Thèse de doctorat, Université de Rennes I, Juillet 1991.

[Lecomte *et al.*, 1989] I. Lecomte, M. Lever, M. Boudy, et A. Tassy. Car noise processing for speech input. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pages 512–515, Glasgow, UK, Mai 1989. ICASSP'89.

[Lee *et al.*, 1990] C. H. Lee, C. H. Lin, et B. H. Juang. A study on speaker adaptation of continuous density HMM parameters. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pages 145–148, Albuquerque, New Mexico, Avril 1990. ICASSP'90.

[Lee *et al.*, 1991] C. H. Lee, C. H. Lin, et B. H. Juang. A study on speaker adaptation of the parameters of continuous density Hidden Markov Models. *IEEE Transactions on Signal Processing*, 39(4):806–814, Avril 1991.

[Lee et Gauvain, 1993] C. H. Lee et J.-L. Gauvain. Speaker Adaptation Based on MAP Estimation of HMM Parameters. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, volume 2, pages 558–561, Minneapolis, Minnesota, USA, Avril 1993. ICASSP'93.

[Lee et Lin, 1993] C. H. Lee et C. H. Lin. On the use of a family of signal limiters for recognition of noisy speech. *Speech Communication*, 12:383–392, 1993.

[Lee et Wang, 1994] L.-M. Lee et H.-C. Wang. A Study on Adaptations of Cepstral and Delta Cepstral Coefficients for Noisy Speech Recognition. Dans *Proc. Int. Conf. on Spoken Language Processing*, volume 3, pages 1011–1014, Yokohama, Japan, Septembre 1994.

[Lee, 1989] C. H. Lee. On the use of some robust modeling techniques for speech recognition. *Computer Speech and Language*, 3:35–52, 1989.

[Lefèbvre *et al.*, 1992] C. Lefèbvre, D. Zwierzynski, D. Starks, et G. Birch. Further optimization of a robust IMELDA speech recogniser for applications with severely degraded speech. Dans *Proc. Int. Conf. on Spoken Language Processing*, volume 1, pages 691–694, Banff, Alberta, Canada, 1992. ICSLP'92.

[Leggetter et Woodland, 1994a] C. J. Leggetter et P. C. Woodland. Speaker Adaptation of Continuous Density HMMs Using Multivariate Linear Regression. Dans *Proc. Int. Conf. on Spoken Language Processing*, volume 1, pages 451–454, Yokohama, Japan, Septembre 1994.

[Leggetter et Woodland, 1994b] C. J. Leggetter et P. C. Woodland. Speaker adaptation of HMMs using linear regression. Rapport Technique F-INFENG/TR.181, CUED, Cambridge University Engineering Department, UK, Juin 1994.

[Lim et Oppenheim, 1978] J. S. Lim et A. V. Oppenheim. All-pole modeling of degraded speech. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 26(3):197–210, Juin 1978.

[Lim et Oppenheim, 1979] J. S. Lim et A. V. Oppenheim. Enhancement and Bandwidth compression of noisy speech. *Proc. of the IEEE*, 67:1586–1604, Décembre 1979.

[Lim et Oppenheim, 1983] J. S. Lim et A. V. Oppenheim. All pole modeling of degraded Speech. Dans *Speech Enhancement*, rédacteur J. Lim, pages 101–114. Prentice-Hall, Englewood Cliffs, NJ, 1983.

[Lim, 1978] J. S. Lim. Evaluation of a correlation subtraction method for ennhancing speech degraded by additive white noise. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 26:471–472, 1978.

[Lim, 1979] J. S. Lim. Spectral Root Homomorphic Deconvolution System. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 27(3):223–232, Juin 1979.

[Lin *et al.*, 1994] Q. Lin, E.-E. Jan, C. W. Che, et B. de Vries. System of Microphone Arrays and Neural Networks for Robust Speech Recognition in Multimedia environments. Dans *Proc. Int. Conf. on Spoken Language Processing*, volume 3, pages 1247–1250, Yokohama, Japan, Septembre 1994.

[Linde *et al.*, 1980] Y. Linde, A. Buzo, et R. M. Gray. An algorithm for Vector Quantizer Design. *IEEE Transactions on Communications*, 28(1):84–95, Janvier 1980.

[Lippmann *et al.*, 1987] R. P. Lippmann, E. A. Martin, et D. B. Paul. Multi-style training for robust isolated word speech recognition. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pages 705–708, Dallas, Texas, USA, Avril 1987. ICASSP'87.

[Liu *et al.*, 1994] F.-H. Liu, R. M. Stern, A. Acero, et P. J. Moreno. Environment normalization for robust speech recognition using direct cepstral compensation. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, volume 2, pages 61–64, Adelaide, Australia, 1994. ICASSP'94.

[Lockwood *et al.*, 1991] P. Lockwood, C. Baillargeat, J.-M. Gillot, J. Boudy, et G. Faucon. Noise reduction for speech enhancement in cars: Non-Linear Spectral Subtraction/Kalman filtering. Dans *Proceedings of European Conference on Speech Communication and Technology*, pages 83–86. EUROSPEECH'91, 1991.

[Lockwood et Alexandre, 1994] P. Lockwood et P. Alexandre. Root adaptive homomorphic deconvolution schemes for speech recognition in noise. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, volume 1, pages 441–444, Adelaide, Australia, 1994. ICASSP'94.

[Lockwood et Boudy, 1991] P. Lockwood et J. Boudy. Experiments with a Non-Linear Spectral Subtractor (NSS), Hidden Markov Models and the projection, for robust speech recognition in cars. Dans *Proceedings of European Conference on Speech Communication and Technology*, pages 79–82. EUROSPEECH'91, 1991.

[Lockwood et Boudy, 1992] P. Lockwood et J. Boudy. Experiments with a Nonlinear Spectral Subtractor (NSS), Hidden Markov Models and the projection, for robust speech recognition in cars. *Speech Communication*, 11(2–3):215–228, Juin 1992.

[Lombard, 1911] E. Lombard. Le Signe de l'Élévation de la Voix. *Ann. Maladies Oreille, Larynx, Nez, Pharynx*, 37:101–119, 1911.

[Lyon et Dyer, 1986] R. F. Lyon et L. Dyer. Experiments with a computational model of the cochlea. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pages 1975–1978, Tokyo, Japan, 1986.

[Lyon, 1982] R. F. Lyon. A computational model of filtering, detection, and compression in rhe cochlea. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pages 1282–1285, 1982.

[Lyon, 1984] R. F. Lyon. A computational model of filtering, detection and compression in the cochlea. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pages 1975–1978, 1984.

[Malah et Cox, 1982] D. Malah et R. V. Cox. A generalized comb filtering technique for speech enhancement. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pages 160–163, 1982.

[Mansour et Juang, 1988] D. Mansour et B. H Juang. The Short-time Modified Coherence representation and its application for noisy speech recognition. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pages 525–528, New York, USA, Avril 1988. ICASSP'88.

[Mansour et Juang, 1989] D. Mansour et B. H. Juang. A family of distortion measures based upon projection operation for robust speech recognition. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 37(11):1659–1671, 1989.

[Martin *et al.*, 1993] F. Martin, K. Shikano, et Y. Minami. Recognition of Noisy Speech by Composition of Hidden Markov Models. Dans *Proceedings of European Conference on Speech Communication and Technology*, volume 2, pages 1031–1034, Berlin, Germany, Septembre 1993. EUROSPEECH'93.

[Matsukoto et Inoue, 1992] H. Matsukoto et H. Inoue. A piecewise linear spectral mapping for supervised speaker adaptation. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, volume 1, pages 449–452, San Francisco, California, Mars 1992. ICASSP'92.

[Matsumoto et Imai, 1986] H. Matsumoto et H. Imai. Comparative study of various spectrum matching measures on noise robustness. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Tokyo, Japan, Avril 1986. ICASSP'86.

[Matsuoka et Shikano, 1991] T. Matsuoka et K. Shikano. Robust HMM Phoneme Modeling for Different Speaking Styles. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pages 265–268, Toronto, Canada, Mai 1991. ICASSP'91.

[McAulay et Malpass, 1980] R. J. McAulay et M. L. Malpass. Speech enhancement using a soft-decision noise suppression filter. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 28:137–145, Avril 1980.

[Mellor et Varga, 1993] B. A. Mellor et A. P. Varga. Noise masking in a transform domain. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, volume 2, pages 87–90, Minneapolis, Minnesota, USA, Avril 1993. ICASSP'93.

[Mena *et al.*, 1990] J. G. Mena, L. S. Sandoval, et R. G. Gomez. A Comparative Study of Feature Extraction Methods for Noisy Speech Recognition. Dans *Signal Processing V: Theories and Applications*, rédacteurs L. Torres, E. Masgrau, et M. A. Lagunas, pages 1191–1194. Elsevier Science Publishers B. V., 1990.

[Meyer, 1992] Y. Meyer. *Ondelettes : algorithmes et applications.* Armand Colin, Paris, 1992.

[Mizuta et Nakajima, 1992] S. Mizuta et K. Nakajima. Optimal Discriminative Training for HMMs to Recognize Noisy Speech. Dans *Proc. Int. Conf. on Spoken Language Processing*, volume 2, pages 1519–1522, Banff, Alberta, Canada, Octobre 1992.

[Mokbel *et al.*, 1992a] C. Mokbel, L. Barbier, et G. Chollet. Adapting a HMM speech recognizer to noisy environments. Dans *Proc. ESCA Workshop, Speech Processing in Adverse Conditions*, pages 211–214, Cannes-Mandelieu, France, Novembre 1992. ESCA.

[Mokbel *et al.*, 1992b] C. Mokbel, L. Barbier, Y. Kerlou, et G. Chollet. Word Recognition in the Car: Adaptating Recognizers to New Environments. Dans *Proc. Int. Conf. on Spoken Language Processing*, pages 707–710, Banff, Alberta, Canada, 1992. ICSLP'92.

[Mokbel *et al.*, 1993] C. Mokbel, J. Monné, et D. Jouvet. On-line adaptation of a speech recognizer to variation in telephone line conditions. Dans *Proceedings of European Conference on Speech Communication and Technology*, pages 1247–1250, Berlin, Germany, Septembre 1993.

[Mokbel *et al.*, 1994] C. Mokbel, P. Pachès-Leal, D. Jouvet, et J. Monné. Compensation of Telephone Line Effects for Robust Speech Recognition. Dans *Proc. Int. Conf. on Spoken Language Processing*, volume 3, pages 987–990, Yokohama, Japan, Septembre 1994.

[Mokbel et Chollet, 1991a] C. Mokbel et G. Chollet. Speech recognition in adverse environments: speech enhancement and spectral transformations. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pages 925–928, Toronto, Canada, Mai 1991. ICASSP'91.

[Mokbel et Chollet, 1991b] C. Mokbel et G. Chollet. Word recognition in a car. Speech enhancement/spectral transformations. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Toronto, Canada, Mai 1991. ICASSP'91.

[Mokbel, 1992] C. Mokbel. *Reconnaissance de la parole dans le bruit: bruitage/débruitage.* Thèse de doctorat, ENST Paris, Juin 1992.

[Montacié et Chollet, 1988] C. Montacié et G. Chollet. Evaluating speech recognizers and databases. Dans *Recent advances in speech understanding and dialog systems*, rédacteur H. Niemann, volume F46. NATO ASI Series, 1988.

[Moreno et Stern, 1994] P. J. Moreno et R. M. Stern. Sources of degradation of speech recognition in the telephone network. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, volume 1, pages 109–112, Adelaide, Australia, 1994. ICASSP'94.

[Morgan *et al.*, 1990] N. Morgan, H. Hermansky, et C. Wooters. SPOONS'90: SPeech recOgnition frOnt eNd workShop. Rapport Technique TR-90-045, International Computer Science Institute, 1990.

[Morgan et Hermansky, 1992] N. Morgan et H. Hermansky. RASTA extensions: Robustness to Additive and Convolutional Noise. Dans *ESCA Workshop Proceedings of Speech Processing in Adverse Conditions*, pages 115–118, Cannes-Mandelieu, France, 1992. ESCA.

[Morii *et al.*, 1990] S. Morii, T. Morii, et M. Hoshimi. Noise Robustness in Speaker Independent Speech Recognition. Dans *Proc. Int. Conf. on Spoken Language Processing*, pages 1145–1148, Novembre 1990.

[Nadas *et al.*, 1989] A. Nadas, D. Nahamoo, et M. A Picheny. Speech recognition using noise-adaptive prototypes. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 37(10):1495–1502, Octobre 1989.

[Nakamura *et al.*, 1993] S. Nakamura, T. Akabane, et S. Hamaguchi. Robust word spotting in adverse car environments. Dans *Proceedings of European Conference on Speech Communication and Technology*, volume 2, pages 1045–1049, Berlin, Germany, 1993. EUROSPEECH'93.

[Nakamura et Shikano, 1989] S. Nakamura et K. Shikano. Speaker adaptation applied to HMM and Neural Networks. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pages 89–92. ICASSP'89, 1989.

[Nandkumar et Hansen, 1992] S. Nandkumar et J. H. L. Hansen. Dual-Channel speech enhancement with auditory spectrum based constraints. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pages 297–300, San Francisco, California, Mars 1992. ICASSP'92.

[Nandkumar et Hansen, 1994] S. Nandkumar et J. H. L. Hansen. Speech enhancement based on a new set of auditory constrained parameters. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, volume 1, pages 1–4, Adelaide, Australia, 1994. ICASSP'94.

[Neumeyer et Weintraub, 1994] L. Neumeyer et M. Weintraub. Probabilistic optimum filtering for robust speech recognition. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, volume 1, pages 417–420, Adelaide, Australia, 1994. ICASSP'94.

[Ng *et al.*, 1992] K. Ng, H. Gish, et J. R. Rohlicek. Robust Mapping of noisy speech parameters for HMM word spotting. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, volume 2, pages 109–112, San Francisco, California, Mars 1992. ICASSP'92.

[Nicol *et al.*, 1992] N. Nicol, S. Euler, M. Falkhausen, H. Reininger, D. Wolf, et J. Zinke. Improving the robustness of automatic speech recognizers using state duration information. Dans *ESCA Workshop Proceedings of Speech Processing in Adverse Conditions*, pages 183–186, Cannes-Mandelieu, France, 1992.

[Nocerino *et al.*, 1985] N. Nocerino, F. K. Soong, L. R. Rabiner, et D. H. Klatt. Comparative Study of Several Distorsion Measures for Speech Recognition. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pages 25–28. ICASSP'85, 1985.

[Nolazco Flores et Young, 1993] J. A. Nolazco Flores et S. J. Young. Adapting a HMM-based Recogniser for Noisy Speech Enhanced by Spectral Subtraction. Rapport Technique F-INFENG/TR 123, Cambridge University Engineering Department, Avril 1993.

[Nolazco Flores et Young, 1994] J. A. Nolazco Flores et S. J. Young. Continuous speech recognition in noise using spectral subtraction and HMM adaptation. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, volume 1, pages 409–412, Adelaide, Australia, 1994. ICASSP'94.

[Ohkura *et al.*, 1992] K. Ohkura, M. Sugihama, et S. Sagayama. Speaker Adaptation Based on Transfer Vector Field Smoothing with Continuous Mixture Density HMMs. Dans *Proc. Int. Conf. on Spoken Language Processing*, volume 1, pages 369–372, Banff, Alberta, Canada, Octobre 1992. ICSLP'92.

[Ohkura *et al.*, 1993] K. Ohkura, D. Rainton, et M. Sugiyama. Noise-Robust HMMs Based on Minimum Error Classification. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, volume 2, pages 75–78, Minneapolis, Minnesota, USA, Avril 1993. ICASSP'93.

[Ohkura et Sugiyama, 1991] K. Ohkura et M. Sugiyama. Speech recognition in a noisy environment using a noise reduction neural network and a codebook mapping technique. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pages 929–932, Toronto, Canada, Mai 1991. ICASSP'91.

[Ohshima et Stern, 1994] Y. Ohshima et R. M. Stern, Jr. Environmental Robustness in Automatic Speech Recognition Using Physiologically Motivated Signal Processing. Dans *Proc. Int. Conf. on Spoken Language Processing*, volume 3, pages 1347–1350, Yokohama, Japan, Septembre 1994.

[Openshaw et Mason, 1994] J. Openshaw et J. S. Mason. On the limitations of cepstral features in noise. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, volume 2, pages 49–52, Adelaide, Australia, 1994. ICASSP'94.

[O'Shaughnessy, 1988] D. O'Shaughnessy. Speech enhancement using vector quantization and a formant distance measure. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pages 549–552. ICASSP'88, 1988.

[O'Shaughnessy, 1989] D. O'Shaughnessy. Enhancing speech Degraded by Additive noise or interfering speakers. *IEEE Communications Magazine*, pages 46–52, 1989.

[Ostendorf et Roukos, 1989] M. Ostendorf et S. Roukos. A Stochastic Segment Model for Phoneme-Based Continuos Speech Recognition. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 37(12):1857–1869, Décembre 1989.

[Pai et Wang, 1992] H. Pai et H. Wang. A Study of the Two-Dimensional Cepstrum Approach for Speech Recognition. *Computer Speech and Language*, 6:361–375, 1992.

[Paliwal et Atal, 1994] K. K. Paliwal et B. S. Atal. A Comparative Study of Feature Representations for Robust Speech Recognition in Adverse Environments. Dans *Proc. Int. Conf. on Spoken Language Processing*, volume 3, pages 1015–1018, Yokohama, Japan, Septembre 1994.

[Paliwal et Basu, 1987] K. K. Paliwal et A. Basu. A speech enhancement based on Kalman filtering. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pages 177–180, Dallas, Avril 1987. ICASSP'87.

[Paliwal, 1982] K. K. Paliwal. On the performance of frequence-weighted cepstral coefficients in vowel recognition. *Speech Communication*, 1:151–154, Mai 1982.

[Paliwal, 1990] K. K. Paliwal. Neural net classifier for robust speech recognition under noisy environments. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pages 429–432, Albuquerque, New Mexico, Avril 1990. ICASSP'90.

[Paliwal, 1992] K. K. Paliwal. Dimensionality Reduction of the Enhanced Feature Set for the HMM-Based Speech Recognizer. *Digital Signal Processing*, pages 157–173, 1992.

[Paliwal, 1993] K. K. Paliwal. Use of Temporal Correlation Between Successive Frames in A Hidden Markov Model Based Speech Recognizer. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, volume 2, pages 215–218, Minneapolis, Minnesota, USA, 1993.

[Papoulis, 1991] A. Papoulis. *Probability, Random Variables, and Stochastic Processes*. McGraw-Hill International Editions, third édition, 1991.

[Perlmutter *et al.*, 1977] Y. M. Perlmutter, L. D. Braida, R. H. Frazer, et A. V. Oppenheim. Evaluation of speech Enhancement system. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pages 212–215. ICASSP'77, Mai 1977.

[Porter et Boll, 1984] J. E. Porter et S. F. Boll. Optimal estimators for spectral restoration of noisy speech. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*. ICASSP'84, Mars 1984.

[Press *et al.*, 1988] W. H. Press, B. P. Flannery, S. A. Teukolsky, et W. T. Vetterling. *Numerical recipes in C. The Art of Scientific Computing*. Cambridge University Press, 1988.

[Puel et André-Obrecht, 1994] J.-B. Puel et R. André-Obrecht. Robust Signal Preprocessing for HMM Speech Recognition in Adverse Conditions. Dans *Proc. Int. Conf. on Spoken Language Processing*, volume 1, pages 259–262, Yokohama, Japan, Septembre 1994.

[Rabiner, 1989] L.R. Rabiner. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *Proceedings of the IEEE*, 77(2):257, Février 1989.

[Ramalho et Mammone, 1994] M. A. Ramalho et R. J. Mammone. A New Speech enhancement technique with application to speaker identification. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, volume 1, pages 29–32, Adelaide, Australia, Avril 1994. ICASSP'94.

[Redner et Walker, 1984] R. A. Redner et H. F. Walker. Mixture densities, Maximum Likelihood and the EM Algorithm. *SIAM Review*, 26(2):195–239, Avril 1984.

[Roe, 1987] D. B. Roe. Speech recognition with a noise-adapting codebook. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pages 1139–1142, Dallas, Texas, USA, Avril 1987. ICASSP'87.

[Rose *et al.*, 1994] R. C. Rose, E. M. Hofstetter, et D. A. Reynolds. Integrated Models of Signal and Background With Application to Speaker Identification in Noise. *IEEE Transactions on Speech and Audio Processing*, 2(2):245–257, Avril 1994.

[Roth *et al.*, 1993] R. Roth, J. Baker, L. Gillic, M. Hunt, Y. Ito, S. Lowe, J. Orloff, B. Peskin, et F. Scattone. Large vocabulary continuous speech recognition of Wall street journal data. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, volume 2, pages 640–643, Minneapolis, Minnesota, USA, 1993. ICASSP'93.

[Roucos *et al.*, 1988] S. Roucos, M. Ostendorf, H. Gish, et A. Derr. Stochastic Segment Modelling Using the Estimate-Maximize Algorithm. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pages 127–130, New York, USA, Avril 1988. ICASSP'88.

[Roukos et Dunham, 1987] S. Roukos et M. O. Dunham. A Stochastic Segment Model for Phoneme-Based Continuous Speech Recognition. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, pages 73–76, Dallas, Texas, USA, Avril 1987. ICASSP'87.

[Sambur, 1978] M. R. Sambur. Adaptive noise canceling for speech signal. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 26:419–423, Octobre 1978.

[Schroeder et Hall, 1974] M. R. Schroeder et J. L. Hall. Model for Mechanical to neural transduction in the auditory receptor. *Journal of the Acoustical Society of America*, 55:1055–1060, 1974.

[Schroeder, 1981] M. R. Schroeder. Direct (nonrecursive) relations between cepstrum and predictor coefficients. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 29(2):297–301, 1981.

[Seide et Mertins, 1994] F. Seide et A. Mertins. Non-linear regression based feature extraction for connected-word recognition in noise. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, volume 2, pages 85–88, Adelaide, Australia, 1994. ICASSP'94.

[Seneff, 1988] S. Seneff. A joint synchrony/mean-rate model of auditory speech processing. *Journal of Phonetics*, 16(1):55–76, Janvier 1988.

[Shamma, 1985] S. A. Shamma. Speech processing in the auditory system II: Lateral inhibition and the central processing of speech evoked activity in the auditory nerve. *Journal of the Acoustical Society of America*, pages 1622–1632, 1985.

[Sheikhzadeh *et al.*, 1994] H. Sheikhzadeh, H. Sameti, L. Deng, et R. L. Brennan. Comparative performance of spectral subtraction and HMM–based speech enhancement strategies with application to hearing aid design. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, volume 1, pages 13–16, Adelaide, Australia, 1994. ICASSP'94.

[Shields, 1970] V. C. Shields. *Separation of Added Speech Signals by Digital Comb Filtering*. S. m. thesis, Departement of Electrical Engineering, MIT, 1970.

[Shikano *et al.*, 1986] K. Shikano, K. F. Lee, et R. Reddy. Speaker adaptation through Vector Quantization. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Tokyo, Japan, 1986. ICASSP'86.

[Shikano et Sugiyama, 1982] K. Shikano et M. Sugiyama. Evaluation of LPC Spectral Matching Measures for Spoken Word Recognition. *Trans. IECE*, J65-D(5):535–541, Mai 1982.

[Singer et Lippmann, 1992] E. Singer et R. P. Lippmann. A Speech Recognizer Using Radial Basis Function Neural Networks in an HMM Framework. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, San Francisco, California, Mars 1992.

[Siohan *et al.*, 1993] O. Siohan, Y. Gong, et J.-P. Haton. A Bayesian Approach to Phone Duration Adaptation for Lombard Speech Recognition. Dans *Proceedings of European Conference on Speech Communication and Technology*, volume 3, pages 1639–1642, Berlin, Germany, Septembre 1993. EUROSPEECH'93.

[Siohan *et al.*, 1994] O. Siohan, Y. Gong, et J.-P. Haton. A comparison of three noisy speech recognition approaches. Dans *Proc. Int. Conf. on Spoken Language Processing*, volume 3, pages 1031–1034, Yokohama, Japan, Septembre 1994. ICSLP'94.

[Siohan *et al.*, 1995] O. Siohan, Y. Gong, et J.-P. Haton. Noise Adaptation Using Linear Regression for Continuous Noisy Speech Recognition. Dans *Proceedings of European Conference on Speech Communication and Technology*, Madrid, Spain, Septembre 1995. EUROSPEECH'95.

[Siohan, 1995] O. Siohan. On the robustness of Linear Discriminant Analysis as a preprocessing step for noisy speech recognition. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, volume 1, pages 125–128, Detroit, Michigan, USA, Mai 1995. ICASSP'95.

[Smolders *et al.*, 1994] J. Smolders, T. Claes, G. Sablon, et D. Van Compernolle. On the importance of the microphone position for speech recognition in the car. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, volume 1, pages 429–432, Adelaide, Australia, 1994. ICASSP'94.

[Soong et Sondhi, 1987] F. K. Soong et M. M. Sondhi. A frequency-weighted Itakura spectral distorsion measure and its application to speech recognition in noise. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pages 625–628, Dallas, Texas, USA, 1987. ICASSP'87.

[Sorensen et Hartmann, 1993] H. B. D. Sorensen et U. Hartmann. Robust speaker-independent speech recognition using Non-Linear Spectral Subtraction based IMELDA. Dans *Proceedings of European Conference on Speech Communication and Technology*, volume 1, pages 235–238, Berlin, Germany, Septembre 1993. EUROSPEECH'93.

[Sorensen et Hartmann, 1994] H. B. D. Sorensen et U. Hartmann. Hybrid Model Decomposition of Speech and Noise in a Radial Basis Function Neural Model Framework. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, volume 2, pages 657–660, Adelaide, Australia, Avril 1994.

[Sorensen, 1991] H. Sorensen. A cepstral noise reduction multi-layer neural network. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pages 933–936, Toronto, Canada, 1991. ICASSP'91.

[Stern *et al.*, 1992] R. M. Stern, F.-H. Liu, Y. Ohshima, T. M. Sullivan, et A. Acero. Multiple approches to robust speech recognition. Dans *Proc. Int. Conf. on Spoken Language Processing*, pages 695–698, Banff, Alberta, Canada, 1992. ICSLP'92.

[Stern *et al.*, 1994] R. M. Stern, F.-H. Liu, P. J. Moreno, et A. Acero. Signal Processing for Robust Speech Recognition. Dans *Proc. Int. Conf. on Spoken Language Processing*, volume 3, pages 1027–1030, Yokohama, Japan, Septembre 1994.

[Stern et Lasry, 1987] R. M. Stern et M. J. Lasry. Dynamic speaker adaptation for feature-based isolated word recognition. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 35(6):751–762, Juin 1987.

[Sullivan et Stern, 1993] T. M. Sullivan et R. M. Stern. Multi-microphone correlation-based processing for robust speech recognition. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, volume 2, pages 91–94, Minneapolis, Minnesota, USA, 1993. ICASSP'93.

[Suzuki *et al.*, 1994] T. Suzuki, K. Nakajima, et Y. Abe. Isolated Word Recognition Using Models for Acoustics Phonetic Variability by Lombard Effect. Dans *Proc. Int. Conf. on Spoken Language Processing*, volume 3, pages 999–1002, Yokohama, Japan, Septembre 1994.

[Takahashi *et al.*, 1993] S. Takahashi, T. Matsuoka, Y. Minami, et K. Shikano. Phoneme HMMS Constrained by Frame Correlations. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, volume 2, pages 219–222, Minneapolis, Minnesota, USA, 1993.

[Takahashi et Sagayama, 1994] J.-I. Takahashi et S. Sagayama. Telephone Line Characteristic Adaptation Using Vector Field Smoothing Technique. Dans *Proc. Int. Conf. on Spoken Language Processing*, volume 3, pages 991–994, Yokohama, Japan, Septembre 1994.

[Takizawa et Hamada, 1990] Y. Takizawa et M. Hamada. Lombard speech recognition by formant-frequency-shifted LPC cepstrum. Dans *Proc. Int. Conf. on Spoken Language Processing*, pages 293–296, 1990.

[Tamura et Waibel, 1988] S. Tamura et A. Waibel. Noise reduction using connectionist models. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pages 553–556. ICASSP'88, Avril 1988.

[Tamura, 1989] S. Tamura. An analysis of a noise reduction neural network. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pages 2001–2003, Glasgow, UK, Mai 1989. ICASSP'89.

[Tohkura, 1987] Y. Tohkura. A weighted Cepstral Distance Measure for Speech Recognition. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 35(10):1414–1422, Octobre 1987.

[Treurniet et Gong, 1994] W. C. Treurniet et Y. Gong. Noise Independent Speech Recognition for a Variety of Noise Types. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, volume 1, pages 437–440, Adelaide, Australia, Avril 1994. ICASSP'94.

[Trompf *et al.*, 1993] M. Trompf, R. Richter, H. Eckhardt, et H. Hackbarth. Combination of distortion-robust feature extraction and neural noise reduction for ARS. Dans *Proceedings of European Conference on Speech Communication and Technology*, volume 2, pages 1039–1042, Berlin, Germany, 1993. EUROSPEECH'93.

[Trompf, 1992] M. Trompf. Experiments with Noise Reduction Neural Networks for Robust Speech Recognition. Rapport Technique TR-92-035, International Computer Science Institute, Berkeley, CA 94704, Mai 1992.

[Tseng *et al.*, 1987] H. P. Tseng, M. J. Sabin, et E. A. Lee. Fuzzy vector quantization applied to Hidden Markov modeling. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pages 641–644, Dallas, Texas, USA, Avril 1987. ICASSP'87.

[Tubach *et al.*, 1991] 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 multicouches. *Traitement du signal*, 7(4):285–292, 1991.

[Usagawa *et al.*, 1994] T. Usagawa, M. Iwata, et M. Ebata. Speech parameter extraction in noisy environment using a masking model. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, volume 2, pages 81–84, Adelaide, Australia, 1994. ICASSP'94.

[Van Compernolle *et al.*, 1990] D. Van Compernolle, W. Ma, F. Xie, et M. Van Diest. Speech recognition in noisy environments with the aid of microphone arrays. *Speech Communication*, 9(5–6):433–442, Décembre 1990.

[Van Compernolle, 1987] D. Van Compernolle. Improved noise immunity in large vocabulary speech recognition with the aid of spectral subtraction. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pages 1143–1146, Dallas, Texas, USA, Avril 1987. ICASSP'87.

[Van Compernolle, 1989a] D. Van Compernolle. Noise adaptation in a hidden Markov model speech recognition system. *Computer Speech and Language*, 3(2):151–168, 1989.

[Van Compernolle, 1989b] D. Van Compernolle. Spectral Estimation Using a Log-Distance Error Criterion Applied to Speech Recognition. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pages 258–261, Glasgow, UK, 1989. ICASSP'89.

[Van Compernolle, 1992] D. Van Compernolle. DSP techniques for speech enhancement. Dans *ESCA workshop: Speech processing in adverse conditions*, pages 21–30, Cannes-Mandelieu, France, Novembre 1992. ESCA.

[Van Trees, 1968] H. L. Van Trees. *Detection, Estimation, and Modulation Theory. Part I.* John Wiley and Sons, 1968.

[Varga *et al.*, 1988] A. Varga, R. Moore, J. Bridle, K. Ponting, et M. Russel. Noise compensation algorithms for use with Hidden Markov Model based speech recognition. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, volume 1, pages 481–484, New York, USA, Avril 1988. ICASSP'88.

[Varga *et al.*, 1992] A. P. Varga, H. J. M. Steeneken, M. Tomlinson, et D. Jones. The NOISEX-92 study on the effect of additive noise on automatic speech recognition. Rapport technique, DRA Speech Research Unit, 1992.

[Varga et Moore, 1990] A. P. Varga et R. K. Moore. Hidden Markov Model Decomposition of Speech and Noise. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pages 845–848, Albuquerque, New Mexico, Avril 1990. ICASSP'90.

[Varga et Moore, 1991] A. P. Varga et R. K. Moore. Simultaneous recognition of concurrent speech signals using Hidden Markov Model Decomposition. Dans *Proceedings of European Conference on Speech Communication and Technology*, pages 1175–1178. EUROSPEECH'91, 1991.

[Varga et Ponting, 1989] A. P. Varga et K. M. Ponting. Control Experiments on Noise Compensation in Hidden Markov Model Based Continuous word Recognizers. Dans *Proceedings of European Conference on Speech Communication and Technology*, Paris, France, 1989. EUROSPEECH'89.

[Vaseghi *et al.*, 1994] S. V. Vaseghi, B. P. Milner, et J. J. Humphries. Noisy speech recognition using Cepstral-Time features and spectral-time filters. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, volume 2, pages 65–68, Adelaide, Australia, 1994. ICASSP;94.

[Vaseghi et Milner, 1992] S. V. Vaseghi et B. P. Milner. Speech Recognition in Noisy Environments. Dans *Proc. Int. Conf. on Spoken Language Processing*, volume 2, pages 1487–1490, Banff, Alberta, Canada, Octobre 1992.

[Vaseghi et Milner, 1993a] S. V. Vaseghi et B. P. Milner. Noise-adaptive Hidden Markov models based on Wiener Filters. Dans *Proceedings of European Conference on Speech Communication and Technology*, volume 2, pages 1023–1026, Berlin, Germany, 1993.

[Vaseghi et Milner, 1993b] S. V. Vaseghi et B. P. Milner. Noisy Speech Recognition Based on HMM, Wiener Filters and Re-evaluation of Most Likely Candidates. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, volume 2, pages 103–106, Minneapolis, Minnesota, USA, 1993.

[Wang *et al.*, 1993] K. Wang, S. A. Shamma, et W. J. Byrne. Noise robustness in the auditory representation of speech signals. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, volume 2, pages 335–338, Minneapolis, Minnesota, USA, 1993. ICASSP'93.

[Wang et Lim, 1982] D. L. Wang et J. S. Lim. The Unimportance of Phase in Speech Enhancement. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 30(4):679–681, Août 1982.

[Weintraub et Neumeyer, 1994] M. Weintraub et L. Neumeyer. Constructiong telephone acoustic models from a high-quality speech corpus. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, volume 1, pages 85–88, Adelaide, Australia, 1994. ICASSP'94.

[Weiss *et al.*, 1974] M. R. Weiss, E. Aschkenasy, et T. W. Parsons. Study and development of the INTEL technique for improving speech intelligibility. Rapport Technique NSN-FR/4023, Nicolet Scientific Corp., 1974.

[Wellekens, 1987] C. Wellekens. Explicite time correlation in hidden Markov Models for speech Recognition. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pages 384–386, Dallas, Texas, USA, 1987.

[Whipple, 1994] G. Whipple. Low residual noise speech enhancement utilizing time-frequency filtering. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, volume 1, pages 5–8, Adelaide, Australia, Avril 1994. ICASSP'94.

[Widrow *et al.*, 1975] B. Widrow, J. R. Glover, J. M. McCool, J. Kaunitz, C. S. Williams, R. H. Hearn, J. R. Zeidler, E. Dong, et R. C. Goodlin. Adaptive noise cancelling: principles and applications. *Proc. of the IEEE*, 63(12):1692–1716, Décembre 1975.

[Witbrock et Haffner, 1992] M. Witbrock et P. Haffner. Rapid connectionist speaker adaption. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pages 453–456, San Francisco, California, 1992. ICASSP'92.

[Wood *et al.*, 1991] L. Wood, D. Pearce, et F. Novello. Improved vocabulary-independent sub-word HMM modeling. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pages 181–184, Toronto, Canada, Mai 1991. ICASSP'91.

[Xie et Van Compernolle, 1993] F. Xie et D. Van Compernolle. Speech enhancement by nonlinear spectral estimation - A unifying approach. Dans *Proceedings of European Conference on Speech Communication and Technology*, volume 1, pages 617–620, Berlin, Germany, Septembre 1993. EUROSPEECH'93.

[Xie et Van Compernolle, 1994] F. Xie et D. Van Compernolle. A family of MLP based nonlinear spectral estimators for noise reduction. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, volume 2, pages 53–56, Adelaide, Australia, 1994. ICASSP'94.

[Xu *et al.*, 1989] L. Xu, J. Oglesby, et J. Mason. The optimization of Perceptually-based features for speaker identification. Dans *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pages 520–523, Glasgow, UK, 1989.

[Xu et Mason, 1989] L. Xu et J. S. Mason. Instantaneous ans transitional perceptually-based features in speaker identification. Dans *Proceedings of European Conference on Speech Communication and Technology*, pages 271–274, Paris, France, Septembre 1989. EuroSpeech'89.

[Young, 1992a] S. J. Young. Cepstral Mean Compensation for HMM recognition in noise. Dans *ESCA Workshop Proceedings of Speech Processing in Adverse Conditions*, pages 123–126, Cannes-Mandelieu, France, 1992. ESCA.

[Young, 1992b] S. J. Young. HTK: Hidden Markov Model Toolkit V1.4 reference manual. Rapport technique, Speech group, Cambridge University Engineering Department, Septembre 1992.

[Zwicker et Feldtkeller, 1981] E. Zwicker et R. Feldtkeller. *Psychoacoustique: l'oreille récepteur d'information.* Masson, 1981.

Index

A
Adaptation
des modèles acoustiques, 31–32
des modèles de durée, 32–33
Analyse
homomorphique en racine, 40
IMELDA, 42
MFCC, 13
OSALPC, 39
PLP, 41
RASTA, 45
SMC, 39
Apprentissage
discriminant, 33
multiréférences, 33–35

C
CDCN, 25

D
Décomposition de modèles, 27–29
Distance
(distorsion) Itakura-Saito, 37
de projection cepstrale, 40
RPS, 38
SWL, 38
WLR, 37
DTW, 13

E
Effet Lombard, 2
Compensation de, 25–26
Estimation
à base de modèles, 22–25

F
Filtrage
adaptatif, 14–16
en peigne, 20
par état, 29–30

L
LDA, 42

M
Masquage
de bruit, 21–22
proactif, 22
simultané, 22
Mel, 44
Modèles
de Markov cachés, 4
stochastiques de trajectoires, 59–69

O
OLA, 11

P
Paramétrage
à base de modèles auditifs, 43–45
PMC, 28

S
Soustraction
cepstrale, 46
spectrale, 11–14

T
Transformation
de la parole, 11–26
des systèmes de reconnaissance, 27–35

V
Vinics, 53–78

Annexes

Annexe A

Algorithme EM

L'algorithme EM [Dempster *et al.*, 1977] est une méthode générale permettant de résoudre des problèmes d'estimation du maximum de la vraisemblance, en présence de données incomplètes. Cette méthode joue un role considérable en estimation, puisqu'elle permet de maximiser par une approche itérative, des fonctions de vraisemblance dont la maximisation directe n'est pas possible. Nous présentons tout d'abord l'inégalité de Jensen, qui est à la base de EM, avant de donner brièvement la formulation générale de EM. Le lecteur pourra se reporter à [Dempster *et al.*, 1977; Huang *et al.*, 1990; Feder *et al.*, 1989; Redner et Walker, 1984] pour des détails et applications de EM.

1 Inégalité de Jensen

Soit $f_X(x)$ et $g_X(x)$ deux *pdfs* d'une variable aléatoire X. On a :

$$\int f_X(x)dx = \int g_X(x)dx = 1 \tag{A.1}$$

L'équation (A.1) peut se réécrire, pour tous les x dont $f_X(x) \neq 0$:

$$\int f_X(x)(\frac{g_X(x)}{f_X(x)} - 1)dx = 0 \tag{A.2}$$

Comme $f_X(x)$ et $g_X(x)$ sont des *pdfs*, on a également :

$$y = \frac{g_X(x)}{f_X(x)} \geq 0 \tag{A.3}$$

Or, on a :

$$y - 1 \geq \log y, \quad \forall y > 0 \tag{A.4}$$

En remplaçant (A.3) dans (A.4), on obtient :

$$\frac{g_X(x)}{f_X(x)} - 1 \geq \log \frac{g_X(x)}{f_X(x)} \tag{A.5}$$

Comme $f_X(x) \geq 0$, multiplier les 2 cotés de (A.5) conserve le signe de l'inégalité :

$$f_X(x)(\frac{g_X(x)}{f_X(x)} - 1) \geq f_X(x) \log \frac{g_X(x)}{f_X(x)} \tag{A.6}$$

En intégrant (A.6) par rapport à x, et en utilisant (A.2), on obtient :

$$0 \geq \int f_X(x) \log \frac{g_X(x)}{f_X(x)} dx \tag{A.7}$$

Soit en définitive, la formulation de l'inégalité de Jensen utilisée dans EM :

$$\int f_X(x) \log f_X(x) \geq \int f_X(x) \log g_X(x) \tag{A.8}$$

2 Algorithme EM

Soit x une donnée observée, qualifiée de donnée « incomplète », réalisation d'une variable aléatoire X. Soit y une donnée «non observée» réalisation d'une variable aléatoire Y. La connaissance simultanée de x et y, notée (x, y), constitue une donnée qualifiée de «complète». Soit $\overline{\Phi}$, un ensemble de paramètres caractérisant la fonction de densité de probabilité de X.

D'après la règle de Bayes, on a :

$$p_{X,Y}(x, y|\overline{\Phi}) = p_{Y|X}(y|x, \overline{\Phi}) p_X(x|\overline{\Phi}) \tag{A.9}$$

où $p_{X,Y}(x, y|\overline{\Phi})$ désigne la *pdf* conjointe de la variable aléatoire (X, Y), $p_{Y|X}(y|x, \overline{\Phi})$ la *pdf* de la donnée non observée, connaissant la donnée observée, et $p_X(x|\overline{\Phi})$ la *pdf* de la donnée incomplète X.

Écrivons le logarithme de l'équation (A.9) :

$$\log p_X(x|\overline{\Phi}) = \log p_{X,Y}(x, y|\overline{\Phi}) - \log p_{Y|X}(y|x, \overline{\Phi}) \tag{A.10}$$

Calculons maintenant l'espérance mathématique sur la donnée complète, de l'équation (A.9) étant donné la donnée observée x et un ensemble Φ des paramètres des *pdfs* :

$$\begin{aligned} &E\{\log p_X(x|\overline{\Phi})|X = x, \Phi\} \\ &\quad = E\{\log p_{X,Y}(x, y|\overline{\Phi})|X = x, \Phi\} - E\{\log p_{Y|X}(y|x, \overline{\Phi})|X = x, \Phi\} \end{aligned} \tag{A.11}$$

On a :

$$\begin{aligned} E\{\log p_X(x|\overline{\Phi})|X = x, \Phi\} &= \int \log p_X(x|\overline{\Phi}) p_{Y|X}(y|x, \Phi) dy \\ &= \log p_X(x|\overline{\Phi}) \\ &= L(x, \overline{\Phi}) \end{aligned} \tag{A.12}$$

Posons $Q(\Phi, \overline{\Phi}) = E\{\log p_{X,Y}(x, y|\overline{\Phi})|X = x, \Phi\}$ et $H(\Phi, \overline{\Phi}) = E\{\log p_{Y|X}(y|x, \overline{\Phi})|X = x, \Phi\}$. En utilisant (A.12), l'équation (A.11) s'écrit :

$$L(x, \overline{\Phi}) = Q(\Phi, \overline{\Phi}) - H(\Phi, \overline{\Phi}) \tag{A.13}$$

Écrivons $H(\Phi, \overline{\Phi})$ et $H(\Phi, \Phi)$. On a :

$$H(\Phi, \overline{\Phi}) = \int \log p_{Y|X}(y|x, \overline{\Phi}) p_{Y|X}(y|x, \Phi) dy \tag{A.14}$$

$$H(\Phi, \Phi) = \int \log p_{Y|X}(y|x, \Phi) p_{Y|X}(y|x, \Phi) dy \tag{A.15}$$

D'après l'inégalité de Jensen (A.8), on obtient :

$$H(\Phi, \Phi) \geq H(\Phi, \overline{\Phi}) \tag{A.16}$$

Cela signifie que si l'on est capable de trouver un ensemble de paramètres $\overline{\Phi}$, tel que $Q(\Phi, \overline{\Phi}) \geq Q(\Phi, \Phi)$, alors on a $L(x, \overline{\Phi}) \geq L(x, \Phi)$. Par conséquent, la log-vraisemblance de la donnée incomplète x augmente de façon monotone à chaque itération lors de la mise à jour de Φ avec $\overline{\Phi}$, obtenu par maximisation de la fonction auxiliaire Q.

La forme générale de EM est la suivante. Étant donné un ensemble de paramètres Φ maximisant la log-vraisemblance de la donnée incompléte, $L(x, \Phi)$, la nouvelle estimation $\overline{\Phi}$ est obtenue par :

1. Choisir une estimation initiale de Φ.
2. Étape E. Calculer la fonction auxiliaire $Q(\Phi, \overline{\Phi})$, en utilisant Φ.
3. Étape M. Calculer $\overline{\Phi}$ tel que $\overline{\Phi} = \text{argmax}_{\overline{\Phi}} Q(\Phi, \overline{\Phi})$.
4. Remplacer Φ par $\overline{\Phi}$, et retour en 2 jusqu'à convergence.

L'algoritme EM est donc utilisé dans des applications où la maximisation de la fonction $Q(\Phi, \overline{\Phi})$ est plus simple que la maximisation directe de la log-vraisemblance de la donnée incomplète $L(x, \Phi)$.

Dans cette présentation, nous n'avons considéré que le cas où une seule donnée incomplète x est disponible. La généralisation à un ensemble de données $x_1, \ldots x_N$ est immédiate [Huang *et al.*, 1990]. Appelons $Q_k(\Phi, \overline{\Phi})$ la fonction auxiliaire associée à la donnée x_k. On peut montrer que la maximisation de la log-vraisemblance des N données x_k, $L(x_1, \ldots x_N, \Phi)$ peut s'effectuer itérativement en maximisant la somme des fonctions auxiliaires $Q(\Phi, \overline{\Phi}) = \sum_{k=1}^{N} Q_k(\Phi, \overline{\Phi})$.

Annexe B

Transformation de base

1 Introduction

Dans cette annexe, nous présentons la méthode de transformation d'espace spectral, appelée transformation de base, définie dans [Gong, 1993] et utilisée dans des applications de RAP en environnement bruité [Treurniet et Gong, 1994; Siohan *et al.*, 1994].

L'objectif de cette méthode est de définir une transformation permettant de projeter l'espace des paramètres de parole bruitée, sur l'espace des paramètres de parole propre. La transformation est déterminée à partir d'un corpus d'adaptation disponible dans les deux environnement, propre et bruité. La transformation étant définie, il est possible de transformer la parole bruitée, afin de la reconnaître en utilisant des modèles acoustiques construits à partir de parole propre.

Les différentes hypothèses utilisées pour la transformation de base sont présentés paragraphe 2. La transformation est définie paragraphe 3, et le paragraphe 4 conclut cette annexe.

2 Notation et principe

Soit respectivement χ et ψ un environnement de test (p.ex. bruité) et un environnement de référence (p.ex. propre), définis dans un espace de dimension D. Soit ϕ un ensemble de J étiquettes, $\phi = \{q_1, \ldots q_J\}$. Une observation correspond à la réalisation acoustique d'une étiquette dans un espace de dimension D. Notons $\mathbf{b}_i^\chi$ une observation de q_i dans l'environnement χ, et $\mathbf{b}_i^\psi$ l'observation associée dans ψ. Soit f une fonction permettant de transformer une observation de l'environnement χ en l'observation correspondante dans ψ, $f : \chi \rightarrow \psi$. On a :

$$\mathbf{b}_i^\psi = f(\mathbf{b}_i^\chi) \tag{B.1}$$

Soit $\mathrm{B}^\psi = \{\mathbf{b}_i^\psi\}$ et $\mathrm{B}^\chi = \{\mathbf{b}_i^\chi\}$, deux ensembles d'observations de l'ensemble d'étiquettes ϕ dans les environnements ψ et χ. B^ψ et B^χ sont appelés *bases* de ψ et χ. Soit $\mathbf{v}^\chi$ une observation dans χ, et $\mathbf{v}^\psi = f(\mathbf{v}^\chi)$, l'observation associée dans ψ. Notons $\mu(\mathbf{a}, \mathbf{b})$, une mesure de similarité entre les vecteurs $\mathbf{a}$ et $\mathbf{b}$.

L'hypothèse principale de la transformation de base est qu'il existe un sous ensemble $\Omega_{\mathbf{v}}^{\chi}$ de $\mathbf{B}^{\chi}$, et un sous ensemble associé $\Omega_{\mathbf{v}}^{\psi}$ de $\mathbf{B}^{\psi}$, tels que les similarités relatives entre $\mathbf{v}^{\chi}$ et les éléments de $\Omega_{\mathbf{v}}^{\chi}$ sont les mêmes que les similarités relatives entre $\mathbf{v}^{\psi}$ et les éléments de $\Omega_{\mathbf{v}}^{\psi}$.

Cela se note :

$$\forall \mathbf{v}^{\chi} \in \chi, \quad \exists \Omega_{\mathbf{v}}^{\chi} \subseteq \mathbf{B}^{\chi}, \quad \text{tel que } \forall \mathbf{b}_i \in \Omega_{\mathbf{v}}^{\chi}, \quad \mu(\mathbf{v}^{\chi}, \mathbf{b}_i^{\chi}) = \gamma \mu(\mathbf{v}^{\psi}, \mathbf{b}_i^{\psi}) \tag{B.2}$$

Le scalaire γ est une constante, ce qui signifie que les similarités relatives entre une observation et un ensemble d'éléments d'une base ne sont pas affectées par la transformation. Par conséquent, ces similarités sont indépendantes de la base dans laquelle on se place.

Généralement, $\Omega_{\mathbf{v}}^{\chi}$ est constitué des éléments de $\mathbf{B}^{\chi}$ les plus proches de $\mathbf{v}$. $\Omega_{\mathbf{v}}^{\psi}$ est directement obtenu par les éléments de $\mathbf{B}^{\psi}$ donc les étiquettes sont celles des éléments de $\Omega_{\mathbf{v}}^{\chi}$.

3 Adaptation par transformation de base

Le vecteur transformé $\mathbf{v}^{\psi}$ peut s'exprimer comme une fonction de $\mathbf{v}^{\chi}$ et des éléments de $\Omega_{\mathbf{v}}^{\chi}$ et $\Omega_{\mathbf{v}}^{\psi}$. Le calcul des similarités μ entre un vecteur $\mathbf{x}$ et les éléments d'une base $\mathbf{B}$ est appelé décomposition de $\mathbf{x}$ dans $\mathbf{B}$.

Supposons que l'apprentissage d'un système de RAP soit effectué dans un environnement ψ, et qu'une procédure existe pour transformer un environnement χ en ψ. Lors de la reconnaissance dans l'environnement χ, les phrases sont décomposées trames par trames dans l'environnement χ. Une transformation inverse est ensuite appliquée dans l'environnement ψ, qui conduit à exprimer les vecteurs de parole dans ψ. Ces phrases peuvent alors être reconnues avec les modèles entraînés dans ψ.

Deux problèmes doivent être résolu pour mettre en œuvre la transformation de base. Le premier consiste à définir une base d'un environnement, le second concerne la définition de la transformation. Ces deux aspects sont abordés respectivement paragraphes 3.1 et 3.2.

3.1 Définition d'une base

Plusieurs contraintes doivent être respectées pour définir une base d'un environnement.

1° Les observations des étiquettes doivent former un hypervolume convexe dans l'espace des paramètres de parole, afin de fournir le plus possible d'information sur l'environnement.

2° Les observations des étiquettes doivent être uniformément distribuées dans cet hypervolume.

3° L'obtention de la correspondance entre les éléments des différentes bases doit être simple.

4° La base doit caractériser le contexte phonétique.

En conséquence, les étiquettes choisies sont les différents phones d'un corpus d'adaptation. Le corpus d'adaptation est étiqueté, et une version bruitée du corpus est obtenue par ajout de bruit. La contrainte 3 est donc directement respectée car la correspondance est immédiate entre les éléments de la base propre et ceux de la base bruitée. Si la transformation de base est utilisée dans un but d'adaptation au locuteur, il est nécessaire de procéder à un alignement temporel dynamique des phrases prononcées par les différents locuteurs. En pratique, chaque élément de la base est constitué de la moyenne de 3 vecteurs de paramètres consécutifs situés au centre de la réalisation acoustique de l'étiquette associée. La satisfaction des contraintes 1 et 2 est réalisée en utilisant un corpus d'adaptation riche et équilibré au niveau phonétique. Les bases des différents environnements proviennent des réalisations d'un même texte, les informations contextuelles sont donc les mêmes dans les différentes bases, et la contrainte 4 est respectée.

3.2 Définition de la transformation de base

Soit $\mathbf{x}_n^\chi$ le vecteur associé à la trame n de la phrase à reconnaître. Soit $s(\mathbf{x}, \mathbf{y})$ la similarité entre les vecteurs $\mathbf{x}$ et $\mathbf{y}$ de dimension D, définie comme l'inverse d'une distance Euclidienne pondérée :

$$s(\mathbf{x}, \mathbf{y}) = \frac{1}{\sqrt{\sum_{k=1}^{D} \lambda_k (x_k - y_k)^2}} \quad \in [0, \infty[\tag{B.3}$$

La mesure de similarité entre le vecteur $\mathbf{x}_n^\chi$ et chaque vecteur de $\Omega_\mathbf{v}^\chi$ est définie par :

$$\forall \mathbf{b}_i^\chi \in \Omega_{\mathbf{x}_n}^\chi, \quad \mu_i(\mathbf{x}_n^\chi, \mathbf{b}_i^\chi) = \frac{s(\mathbf{x}_n^\chi, \mathbf{b}_i^\chi)^\alpha}{\sum\limits_j s(\mathbf{x}_n^\chi, \mathbf{b}_j^\chi)^\alpha} \tag{B.4}$$

avec $\alpha = \frac{1}{(m-1)}$ et $m \in [1, \infty[$.

Il est possible d'exprimer $\mathbf{x}_n^\chi$ comme une somme pondérée des éléments de $\Omega_\mathbf{v}^\chi$:

$$\mathbf{x}_n^\chi = \frac{\sum_j (\mu_j)^m \cdot \mathbf{b}_j^\chi}{\sum_k (\mu_k)^m} \tag{B.5}$$

La transformation de base consiste à remplacer les éléments $\Omega_\mathbf{v}^\chi$ par les éléments correspondant de $\Omega_\mathbf{v}^\psi$, les coefficients de pondération μ étant conservés :

$$\mathbf{x}_n^\psi = \frac{\sum_j (\mu_j)^m \cdot \mathbf{b}_j^\psi}{\sum_k (\mu_k)^m} \tag{B.6}$$

Dans les équations (B.3), (B.3) et (B.6), les sommes sur j (et k) sont effectuées sur tous les j (et k) tels que $\mathbf{b}_j^\chi \in \Omega_{\mathbf{x}_n}^\chi$. En pratique, l'ensemble $\Omega_{\mathbf{x}_n}^\chi$ est constitué des K éléments de $\mathbf{B}^\chi$ les plus proches de $\mathbf{x}_n^\chi$.

Il faut remarquer que cette transformation se distingue des approches à base de quantification vectorielle floue, dans la mesure ou chaque élément d'une base est ici associé explicitement à un symbole phonétique.

4 Conclusion

Dans cette annexe, nous avons rapidement présenté la transformation de base de [Gong, 1993], pour son application à la reconnaissance de la parole dans le bruit.

L'hypothèse de la transformation de base est que le bruit ne modifie pas les distances relatives entre un vecteur et les éléments d'une base d'un espace. Un vecteur étant décomposé comme une somme pondéré des éléments d'une base dans un espace, exprimer ce vecteur dans un autre espace consiste simplement à changer de base.

Les constituants d'une base sont définis comme l'ensemble des phones d'un corpus d'adaptation de taille réduite. Le corpus d'adaptation étant étiqueté, la correspondance entre les éléments de la base de l'espace propre et ceux de la base de l'espace bruitée est immédiate.

Le problème de cette approche est que la transformation est définie sans utiliser un critère objectif, et en dehors de tout cadre d'optimalité, contrairement aux différentes approches proposées partie III.

Annexe C

Tableaux des résultats de reconnaissance

SNR	Trans. de base		Comb. de modèles		Régression linéaire		Fitrage par états	
0	60.88	[59.63,62.11]	65.20	[63.98,66.40]	63.06	[61.82,64.28]	53.09	[51.82,54.36]
6	83.70	[82.74,84.62]	81.75	[80.74,82.71]	85.43	[84.50,86.30]	76.08	[74.97,77.15]
12	90.75	[89.99,91.47]	90.01	[89.22,90.75]	94.28	[93.66,94.84]	87.09	[86.22,87.92]
18	93.78	[93.13,94.36]	93.23	[92.57,93.85]	96.73	[96.24,97.15]	91.90	[91.18,92.57]
24	95.56	[95.01,96.06]	95.38	[94.81,95.88]	97.50	[97.07,97.87]	94.94	[94.35,95.47]
30	95.93	[95.40,96.41]	96.25	[95.74,96.71]	97.79	[97.39,98.14]	95.70	[95.15,96.19]
36	96.17	[95.65,96.63]	97.07	[96.61,97.47]	97.84	[97.44,98.18]	96.59	[96.10,97.03]

SNR	Appr. dans le bruit		Modèles propres	
0	72.82	[71.68,73.94]	3.25	[2.83, 3.74]
6	85.71	[84.80,86.58]	12.23	[11.42,13.09]
12	91.91	[91.18,92.57]	35.90	[34.69,37.13]
18	95.18	[94.60,95.69]	75.19	[74.07,76.27]
24	96.73	[96.24,97.15]	93.37	[92.71,93.98]
30	97.67	[97.25,98.02]	96.69	[96.21,97.12]
36	98.00	[97.61,98.32]	97.71	[97.29,98.06]

TAB. C.1 - *Comparaison des performances des différentes approches. Bruit blanc Gaussien. Moyenne sur tous les locuteurs. (cf. fig. 9.10 et 9.13)*

SNR	Trans. de base		Comb. de modèles		Régression linéaire		Fitrage par états	
0	71.41	[70.24,72.54]	67.35	[66.14,68.53]	59.13	[57.87,60.37]	46.63	[45.36,47.90]
6	85.88	[84.97,86.74]	82.23	[81.24,83.19]	85.16	[84.23,86.04]	71.78	[70.62,72.91]
12	91.26	[90.52,91.95]	88.41	[87.57,89.20]	94.59	[93.98,95.13]	87.30	[86.43,88.12]
18	93.44	[92.78,94.04]	91.98	[91.27,92.65]	96.59	[96.10,97.03]	92.56	[91.86,93.20]
24	95.19	[94.62,95.71]	94.77	[94.18,95.31]	97.27	[96.82,97.65]	95.98	[95.36,96.51]
30	95.95	[95.42,96.42]	96.24	[95.72,96.69]	97.62	[97.20,97.98]	96.51	[96.01,96.95]
36	96.24	[95.72,96.69]	96.61	[96.12,97.04]	97.91	[97.51,98.24]	96.93	[96.46,97.34]

SNR	Appr. dans le bruit		Modèles propres	
0	53.86	[52.59,55.13]	4.37	[3.88, 4.92]
6	84.30	[83.35,85.20]	18.83	[17.85,19.84]
12	89.37	[88.56,90.13]	44.80	[43.54,46.07]
18	95.01	[94.42,95.53]	80.43	[79.40,81.42]
24	96.86	[96.38,97.27]	95.17	[94.60,95.69]
30	97.59	[97.17,97.95]	97.16	[96.71,97.56]
36	98.08	[97.70,98.40]	97.69	[97.28,98.04]

TAB. C.2 - *Comparaison des performances des différentes approches. Bruit Sèche-cheveux. Moyenne sur tous les locuteurs. (cf. fig. 9.10 et 9.13)*

SNR	Trans. de base		Comb. de modèles		Régression linéaire		Fitrage par états	
0	84.68	[83.74,85.58]	78.49	[77.43,79.52]	86.54	[85.64,87.38]	71.12	[69.95,72.26]
6	89.83	[89.03,90.57]	87.99	[87.14,88.79]	94.31	[93.69,94.87]	89.04	[88.22,89.81]
12	93.88	[93.24,94.46]	92.73	[92.04,93.36]	96.81	[96.34,97.23]	93.66	[93.01,94.25]
18	95.50	[94.94,96.00]	95.08	[94.49,95.60]	97.35	[96.91,97.73]	95.81	[95.27,96.30]
24	96.10	[95.58,96.57]	96.20	[95.69,96.66]	98.01	[97.62,98.34]	96.61	[96.12,97.04]
30	96.36	[95.85,96.81]	97.08	[96.62,97.48]	97.94	[97.55,98.27]	97.33	[96.89,97.72]
36	96.54	[96.05,96.98]	97.69	[97.28,98.04]	97.94	[97.55,98.27]	97.47	[97.04,97.84]

SNR	Appr. dans le bruit		Modèles propres	
0	89.97	[89.05,90.82]	34.47	[33.27,35.68]
6	94.45	[93.84,95.00]	67.65	[66.44,68.82]
12	96.63	[96.14,97.06]	87.01	[86.07,87.89]
18	97.57	[97.15,97.93]	95.11	[94.53,95.63]
24	97.88	[97.48,98.21]	97.08	[96.62,97.48]
30	97.78	[97.37,98.12]	97.75	[97.35,98.10]
36	98.07	[97.66,98.41]	97.96	[97.57,98.29]

TAB. C.3 - *Comparaison des performances des différentes approches. Bruit Lynx. Moyenne sur tous les locuteurs. (cf. fig. 9.11 et 9.13)*

SNR	Trans. de base		Comb. de modèles		Régression linéaire		Fitrage par états	
0	73.25	[72.10,74.36]	73.43	[72.29,74.54]	71.44	[70.28,72.58]	63.93	[62.70,65.15]
6	88.61	[87.78,89.40]	87.67	[86.81,88.48]	93.59	[92.94,94.19]	85.48	[84.56,86.35]
12	92.81	[92.13,93.44]	93.35	[92.69,93.96]	96.29	[95.78,96.74]	92.97	[92.29,93.59]
18	95.75	[95.21,96.23]	95.63	[95.08,96.12]	97.60	[97.18,97.96]	95.85	[95.31,96.33]
24	95.60	[95.04,96.09]	96.00	[95.47,96.47]	97.88	[97.48,98.21]	96.66	[96.17,97.09]
30	96.96	[96.42,97.43]	96.73	[96.24,97.15]	98.13	[97.75,98.44]	97.05	[96.59,97.45]
36	96.85	[96.30,97.33]	97.27	[96.82,97.65]	97.79	[97.39,98.14]	96.95	[96.48,97.36]

SNR	Appr. dans le bruit		Modèles propres	
0	79.50	[78.46,80.51]	13.50	[12.65,14.39]
6	92.28	[91.48,93.01]	37.31	[36.09,38.55]
12	94.01	[93.38,94.59]	71.95	[70.79,73.08]
18	96.41	[95.90,96.85]	93.56	[92.90,94.15]
24	97.60	[97.18,97.96]	96.24	[95.72,96.69]
30	97.77	[97.36,98.12]	97.35	[96.91,97.73]
36	97.67	[97.26,98.03]	97.84	[97.44,98.18]

TAB. C.4 - *Comparaison des performances des différentes approches. Bruit avion F16. Moyenne sur tous les locuteurs. (cf. fig. 9.11 et 9.13)*

SNR	Trans. de base		Comb. de modèles		Régression linéaire		Fitrage par états	
0	75.88	[74.77,76.95]	81.75	[80.74,82.71]	87.83	[86.98,88.64]	70.07	[68.89,71.22]
6	90.05	[89.26,90.79]	91.30	[90.55,91.99]	95.98	[95.45,96.46]	89.20	[88.39,89.97]
12	93.96	[93.32,94.54]	95.23	[94.65,95.74]	97.23	[96.79,97.62]	94.95	[94.37,95.48]
18	95.46	[94.90,95.96]	96.42	[95.92,96.87]	97.91	[97.51,98.24]	96.22	[95.70,96.68]
24	96.59	[96.10,97.03]	97.23	[96.78,97.62]	98.23	[97.86,98.54]	97.37	[96.93,97.75]
30	96.49	[95.99,96.93]	97.67	[97.26,98.03]	97.92	[97.53,98.26]	97.94	[97.55,98.27]
36	96.19	[95.67,96.65]	97.76	[97.35,98.10]	97.86	[97.46,98.20]	97.39	[96.95,97.76]

SNR	Appr. dans le bruit		Modèles propres	
0	90.92	[90.17,91.63]	23.24	[22.11,24.41]
6	95.22	[94.62,95.75]	44.78	[43.32,46.25]
12	96.88	[96.41,97.29]	81.69	[80.53,82.80]
18	97.61	[97.18,97.96]	94.24	[93.52,94.89]
24	98.09	[97.71,98.41]	97.02	[96.55,97.42]
30	97.99	[97.61,98.32]	97.57	[97.15,97.93]
36	98.08	[97.70,98.40]	97.72	[97.31,98.07]

TAB. C.5 - *Comparaison des performances des différentes approches. Bruit autobus. Moyenne sur tous les locuteurs. (cf. fig. 9.12 et 9.13)*

Nom : **SIOHAN**

Prénom : **Olivier**

DOCTORAT de l'UNIVERSITE HENRI POINCARE, NANCY-I

en INFORMATIQUE

VU, APPROUVÉ ET PERMIS D'IMPRIMER

Nancy, le 14 SEP. 1995 n° 337

Le Président de l'Université

J.P. FINANCE

3C-F9

UNIVERSITE HENRI POINCARE - NANCY 1
24-30, rue Lionnois B.P.3069 54013 Nancy cedex
Tél : 83 32 81 81 Fax : 83 32 95 90
Adresse électronique : _____ @presn1.u-nancy.fr

Résumé

Les systèmes actuels de reconnaissance automatique de la parole (RAP) sont généralement peu robustes aux variations du signal intervenant entre les conditions de test et d'apprentissage. Dans cette thèse, nous proposons et évaluons différentes approches pour améliorer la robustesse au bruit du système de reconnaissance de parole continue VINICS du CRIN-INRIA Lorraine, système fondé sur des modèles stochastiques de trajectoires de parole (STM), alternative efficace aux traditionnels modèles de Markov cachés (HMM).

Dans une première partie, nous dressons un bilan des principales approches développées ces dix dernières années dans le domaine de la RAP dans le bruit.

La seconde partie est constituée d'une étude et comparaison de trois approches. Nous développons d'une part une approche permettant d'estimer un STM hybride de parole bruitée, à partir d'un HMM de bruit et d'un STM de parole propre. D'autre part, nous proposons d'appliquer un filtrage du signal, spécifique à chaque état de chaque STM et optimisé selon un critère significatif au niveau perceptif. Ensuite, nous appliquons une méthode d'adaptation des STM de parole propre aux variations des conditions d'environnement, calculée par régression linéaire. Ces trois approches sont comparées expérimentalement sur une tâche de reconnaissance de la parole continue, en mode dépendant du locuteur, pour un vocabulaire d'un millier de mots, en présence de différents bruits additifs réels. L'adaptation par transformation linéaire s'avère beaucoup plus efficace que les autres approches.

Enfin, dans une dernière partie, nous développons d'une part une étude expérimentale sur l'utilisation de l'analyse discriminante linéaire pour mettre en œuvre un paramétrage du signal de parole robuste au bruit. Nous mettons en évidence que l'analyse discriminante permet d'obtenir un paramétrage efficace pour la reconnaissance de la parole dans le bruit. Cependant, nos expériences montrent qu'un tel paramétrage est peu robuste aux variations du rapport signal-à-bruit, mais cette conclusion reste très dépendante de la nature du bruit. D'autre part, nous prenons en compte les variations du rythme d'élocution provoquées par l'effet Lombard, en utilisant une méthode d'adaptation des modèles de durée des phonèmes, sous le cadre général de l'apprentissage Bayésien. Cette méthode, évaluée sur une tâche de reconnaissance de mots isolés permet d'améliorer de façon significative les taux de reconnaissance.

Mots-clés: reconnaissance automatique de parole bruité, filtrage de bruit, combinaison de modèles stochastiques, modèles stochastiques de trajectoires, regression linéaire, analyse linéaire discriminante, estimation Bayésienne