Contributions à la reconnaissance automatique de la parole avec données manquantes

Sébastien Demange

Nancy-Université

Département de formation doctorale en informatique **École doctorale IAEM Lorraine**
UFR STMIA

Contributions à la reconnaissance automatique de la parole avec données manquantes

THÈSE

présentée et soutenue publiquement le 8 novembre 2007

pour l'obtention du

Doctorat de l'université Henri Poincaré – Nancy 1

(spécialité informatique)

par

Sébastien Demange

Composition du jury

Rapporteurs : **Laurent Miclet**
Professeur - ENSSAT - Lanion - France
Dirk Van Compernolle
Professeur - Université Catholique de Louvain - Leuven - Belgique

Examinateurs : **Jean-Paul Haton**
Professeur - Université Henri Poincaré - Nancy - France
Noureddine Ellouze
Professeur - ENIT - Tunis - Tunisie
Salvatore Tabbone
MCF (HDR) - Université de Nancy 2 - Nancy - France
Christophe Cerisara
Chargé de recherche CNRS - Nancy - France

Laboratoire Lorrain de Recherche en Informatique et ses Applications — UMR 7503

Loria INRIA LORRAINE

Mis en page avec la classe thloria.

Remerciements

Une thèse est une étape inportante dans la vie de toute personne attirée par le monde de la recherche. Cette histoire est riche de nouvelles expériences, d'échanges et de rencontres. A ce titre, j'aimerais remercier de nombreuses personnes pour avoir contribué, chacune à leur manière, à la rédaction de ce mémoire. J'adresse donc mes remerciements les plus chaleureux et les plus sincères :

- à Violaine, qui partage ma vie depuis plus de 10 ans et qui m'a toujours encouragé et soutenu ces trois années durant.
- à Jean-Paul Haton et Christophe Cerisara pour m'avoir mis le pied à l'étrier, suivi et conseillé ainsi que pour leur grande disponibilité et gentillesse.
- aux autres membres de mon jury de thèse, Salvatore Tabbone, Nourredine Ellouze, Laurent Miclet et Dirk Van Compernolle.
- à tous les membres de l'équipe PAROLE.
- à toute ma famille et plus particulièrement mes parents, mon frère ainsi que Michel, Gene et Amandine.
- à tous mes amis avec une pensée plus particulière pour Caroline, Emmanuel, Slim, Joseph, Romain, Gilles, Loïc, et bien d'autres

A Violaine.

Des paroles de douceur
Des moments de bonheur
Des regards échangés
Un zeste de complicité
Des étoiles dans les yeux
La joie d'être deux
Main dans la main
Suivre le même chemin
Juste la magie de l'amour
Une belle vie pour toujours

Table des matières

Table des figures

Liste des tableaux

Introduction générale

« Ouvre la porte, HAL ! »

- 2001 : l'odyssée de l'espace -

Parler avec les machines est une des visions récurrentes de notre imagination collective de l'informatique du futur. Dès 1968 Stanley Kubrick avait imaginé et mis en scène un ordinateur intelligent, appelé HAL, capable de raisonner, de réfléchir mais aussi de communiquer verbalement. Pourtant, à cette époque, la technologie ne permettait de reconnaître tout au plus quelques centaines de mots par des systèmes simulés sur de gros ordinateurs. Quarante ans plus tard, grâce à l'avènement de l'informatique et aux efforts déployés, les systèmes de reconnaissance vocale sont devenus des produits de consommation destinés à un très large publique. La reconnaissance vocale est devenue une des technologies prépondérantes dans le développement d'interfaces Homme-Machine avancées. Toutefois, malgré les avancées très importantes de ces dernières années dans ce domaine, les systèmes actuels sont encore en deçà des performances de notre système d'audition. Un des principaux obstacles au déploiement des systèmes de reconnaissance vocale est la robustesse au bruit. Les différences entre les conditions d'utilisation (généralement bruitées) et d'apprentissage (absence de bruit) des modèles acoustiques provoquent une dégradation significative des taux de reconnaissance, même si ces dégradations semblent minimes à l'oreille.

La reconnaissance de la parole avec données manquantes est une approche qui fut développée dans le domaine de la vision [Ahmad 93] et transposée à la reconnaissance automatique de la parole il y a près de 10 ans [Cooke 96, Cooke 97, Cooke 01b]. A la différence des méthodes de débruitage ou d'adaptation, la reconnaissance avec données manquantes utilise un masque. Ce masque correspond à l'identification dans le signal observé (plus précisément dans le domaine spectral [1]) des paramètres trop corrompus par le bruit pour fournir une information exploitable et pertinente au moteur de reconnaissance. Il est montré qu'il est plus judicieux d'ignorer de tels paramètres durant le processus de décodage. Ces paramètres sont communément appelés données manquantes ou masquées. Une fois ces paramètres identifiés, des algorithmes de reconnaissance en présence de données manquantes sont mis en œuvre. Ces algorithmes peuvent être divisés en

[1] domaine de paramétrisation résultant d'une analyse fréquentielle du signal.

deux familles. Les techniques dites d'*imputation* [Raj 00] estiment la contribution énergétique du signal de la parole pour les paramètres masqués. Les données manquantes sont reconstruites afin de fournir au décodeur un ensemble complet de paramètres décrivant le signal à reconnaître. Les techniques dites de *marginalisation* [Vizinho 99, Barker 01b, Morris 01a] reposent sur une adaptation du moteur de reconnaissance pour que ce dernier puisse reconnaître un signal de parole à partir d'une représentation incomplète de celui-ci. Plus précisément, la vraisemblance des paramètres masqués et substituée par son espérance calculée sur l'ensemble des valeurs possibles de ces paramètres. De nombreux travaux ont montré que de telles stratégies permettent d'améliorer considérablement les taux de reconnaissance. Lorsque les masques sont connus *a priori* (masques oracles), c'est-à-dire lorsque les paramètres masqués sont clairement identifiés à partir des signaux de parole seule et du bruit, les taux de reconnaissance sont proches de ceux obtenus en absence de bruit. Cependant ces performances représentent seulement les performances potentiellement atteignables. En pratique les masques de données manquantes sont estimés à partir des seules observations bruitées et constituent par conséquent une approximation erronée des masques oracles. Les différentes erreurs d'identification commises par les estimateurs de masques se traduisent par une chute des performances illustrant le rôle de premier plan des masques. L'estimation des masques constitue donc un problème central en reconnaissance automatique de la parole avec données manquantes faisant l'objet de nombreuses publications dans ce domaine.

Nos travaux se placent dans le contexte de l'estimation de masques à partir de modèles stochastiques. Cette approche fut initiée à l'université de Carnegie Mellon par l'équipe de Richard Stern. Les publications [Seltzer 00, Raj 00, Kim 05, Kim 06] dont elle fait l'objet montrent qu'il est possible d'apprendre des modèles de masques et les résultats reportés sont très encourageants. Notre première contribution concerne la définition de ces modèles. L'estimateur bayésien proposé par Seltzer et Raj Ramakrishnan classifie chaque coefficient spectral du signal observé comme fiable ou manquant de manière indépendante. En d'autres termes, le masque d'un coefficient spectral particulier ne dépend pas des masques des coefficients de son voisinage. Pourtant nous montrons qu'une similitude existe entre la structure des masques et l'enveloppe énergétique du signal de parole dans le domaine spectral. Les coefficients de faible énergie sont plus sensibles au bruit que les coefficients de forte énergie et sont donc plus souvent masqués. L'enveloppe énergétique du signal de parole étant très structurée, nous supposons que les masques de données manquantes le sont de la même manière. A cet égard nous proposons de nouvelles architectures d'estimateurs bayésiens dans le but de restituer cette structure. Des erreurs de masque locales peuvent ainsi être évitées en considérant un masque dans sa globalité (le masque d'un phrase par exemple) et non plus comme une composition d'entités (masques à l'échelle du coefficient) indépendantes.

La mise en œuvre d'un algorithme de décodage de la parole sur des observations partielles repose sur la définition même du masque de données manquantes. La fiabilité d'un paramètre

acoustique est le plus souvent déterminée à partir du SNR [2]. Tout coefficient spectral dont le SNR est inférieur à un seuil prédéterminé est considéré comme manquant. Dans le cadre de la marginalisation de données, la prise en compte de cette définition de masque a permis d'affiner l'algorithme de décodage, notament en proposant des intervalles de marginalisation spécifiques aux données manquantes et fiables. Ces intervalles sont plus fins que ceux initialement proposés permettant un gain significatif en terme de taux de reconnaissance. Nous proposons dans cette optique une nouvelle définition de masque et montrons comment l'exploiter dans le but de minimiser les intervalles de marginalisation.

Le premier chapitre constitue une rapide introduction à la reconnaissance robuste de la parole. Nous présentons dans un premier temps les principes généraux de la reconnaissance automatique de la parole et relatons différentes stratégies de décodage usuellement utilisées. Nous décrivons plus particulièrement le modèle de Markov caché ainsi que sa mise en œuvre puisque ce modèle s'est imposé comme modèle de référence dans la communauté du traitement des langues notament pour sa capacité à modéliser un signal à évolution temporelle tel le signal de parole. Les systèmes de reconnaissance de la parole actuels exploitent pour la plupart ce modèle. Dans un second temps nous adressons le problème de la robustesse au bruit. Nous mettons en évidence les principales techniques permettant d'améliorer la robustesse des systèmes. Celles-ci interviennent à des étapes distinctes du processus de reconnaissance allant de la paramétrisation du signal à l'algorithme de décodage.

La reconnaissance de la parole avec données manquantes est présentée au chapitre 2. Nous relatons des travaux montrant que notre système auditif se comporte de manière sélective vis-à-vis des différents stimuli qu'il traite. L'oreille humaine est capable de distinguer les différents acteurs d'une scène auditive et peut par un processus de masquage se focaliser sur une source sonore particulière. Ces études ne montrent pas comment nous sélectionnons les portions d'intérêt du signal mais montrent que nous sommes capables de reconnaître de la parole à partir d'une représentation parcellaire du signal acoustique. Nous définissons ensuite les notions de données manquantes et de masque de données manquantes dans le cadre de la reconnaissance automatique de la parole. Les différents algorithmes d'imputations et de marginalisation sont décrits. Nous concluons ce chapitre par une évaluation comparative de 3 techniques de marginalisation mettant en évidence le fort potentiel de la reconnaissance de la parole avec données manquantes mais aussi le rôle de premier plan que jouent les masques.

L'estimation de masques de données manquantes constitue aujourd'hui un enjeu important et motive de nombreux travaux. Nous proposons au troisième chapitre un état de l'art de cet axe de recherche. Les principales approches proposées dans la littérature sont présentées avec comme seule limitation l'usage d'un unique microphone pour l'acquisition du signal. Ce travail prospectif

[2]Rapport signal sur bruit. Cette mesure permet de quantifier le degré de corruption du signal.

n'est pas limité au seul cadre applicatif que constitue la reconnaissance de la parole mais couvre également des domaines connexes comme la séparation aveugle de sources, l'analyse computationnelle de scène auditive ou encore la détection de parole utile. Nous avons choisi de classer ces travaux en deux catégories : d'une part les méthodes s'inspirant du fonctionnement de notre appareil auditif, et d'autre part, les méthodes orientées traitement du signal. L'objectif n'est pas d'opposer ces deux approches. Au contraire, de récents travaux, le décodeur multi-sources de Barker [Barker 06] par exemple, montrent le bénéfice de combiner des concepts issus de ces deux approches.

Nous proposons au chapitre 4 deux nouvelles modélisations des masques. La première a pour objectif la modélisation des dépendances existantes entre les valeurs de masque des coefficients spectraux. Nous motivons cette approche en mettant en évidence les similitudes entre l'enveloppe énergétique du signal de parole et la structure des masques dans le domaine spectral. Nous définissons deux types de dépendance : les dépendances temporelles et les dépendances fréquentielles. Nous décrivons comment ces dépendances peuvent être prise en compte pendant le processus d'estimation de masque, et nous proposons de nouveaux modèles stochastiques de masques intégrant individuellement ou conjointement ces dépendances. Nous proposons ensuite une nouvelle définition de masque permettant, dans le cadre de la marginalisation de données, d'affiner l'algorithme de décodage. Cette nouvelle définition de masque permet de réduire les intervalles de marginalisation comparativement aux intervalles dérivés des masques fondés sur le seuillage du SNR classiquement utilisés.

Ces propositions sont évaluées au chapitre 5. Une comparaison des masques générés par nos estimateurs avec les masques oracles est présentée afin de rendre compte de leur qualité en terme d'identification des données masquées. Nous présentons également une étude qualitative des masques résultant de nos propositions en les comparant aux masques obtenus à partir d'estimateur de référence que nous définirons. L'objectif affiché de nos travaux est d'améliorer la qualité des masques ainsi que leur prise en compte par le moteur de reconnaissance. Nous présentons dans cette optique une évaluation comparative des résultats de reconnaissance obtenus à partir de nos propositions sur différentes bases de données par rapport aux taux de reconnaissance obtenus avec les systèmes de référence.

Chapitre 1

La reconnaissance automatique de la parole robuste au bruit

« J'entends ta voix dans tous les bruits du monde. »

- Paul Eluard -

Sommaire

Ce chapitre présente le problème de la reconnaissance automatique de la parole (RAP). Nous caractérisons dans un premier temps le signal acoustique de la parole. Nous évoquons ensuite le principe général de la RAP et en particulier l'approche bayésienne qui est la plus répandue. Nous détaillons un modèle bayésien particulier :le modèle de Markov caché (HMM : Hidden Markov Model). Ce modèle fournit de très bons taux de reconnaissance en condition d'utilisation maîtrisée. Cependant ces performances sont loin d'être aussi bonnes lorsque les conditions d'utilisation se dégradent. Cette différence de performance due aux conditions d'expérimentation relève de la robustesse au bruit du système de reconnaissance. Nous exposons les grandes approches de reconnaissance robuste de la parole. Plusieurs ouvrages traitent de ce problème et plus généralement de la reconnaissance de la parole [Boite 00, Mariani 02, Haton 06].

1.1 Reconnaissance automatique de la parole

1.1.1 Le signal de la parole

Le signal de parole est une onde acoustique modulée par l'appareil phonatoire en fréquence et en amplitude. Cette onde est généralement présentée sous la forme d'une courbe (Fig. 1.1) représentant les variations d'amplitude du signal au cours du temps.

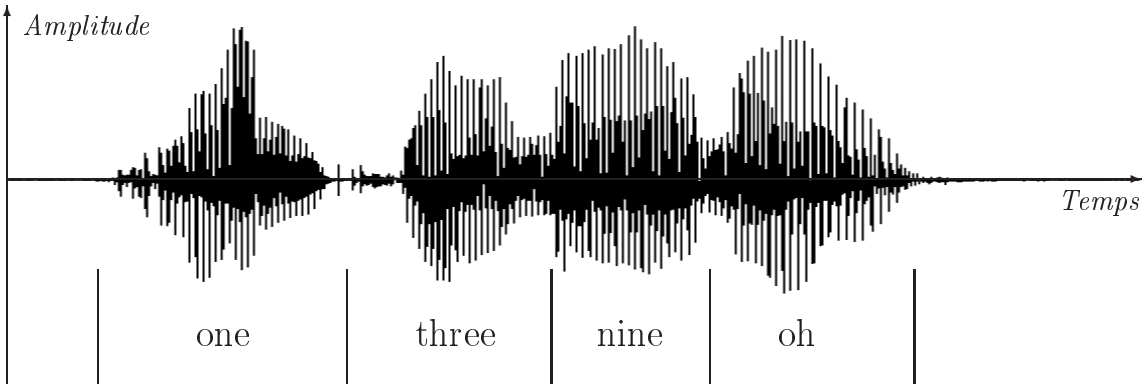


FIG. 1.1 – *Représentation temporelle d'un signal de parole correspondant à la séquence de mots "one three nine oh".*

Le signal de parole est une concaténation de réalisations acoustiques élémentaires. Ces réalisations sont plus connues sous le nom de *phonèmes*. Un phonème est une entité abstraite définie comme la plus petite unité acoustique. Chaque langue peut être alors caractérisée par un ensemble de phonèmes qui constituent en quelque sorte les briques acoustiques élémentaires à partir desquelles les syllabes, les mots et les phrases sont construits. Tout signal de la parole peut alors être exprimé comme une succession de phonèmes. Ce signal véhicule un ensemble d'informations très diverses : le message que veut faire passer le locuteur, son humeur, son identité, etc. Le signal à reconnaître fait, dans un premier, l'objet d'un prétraitement, appelé paramétrisation, consistant à extraire de ce signal des paramètres pertinents permettant d'identifier la séquence des phonèmes prononcés.

1.1.2 Paramétrisation du signal

1.1.2.1 Représentations paramétriques du signal de la parole

Le rôle d'un module de paramétrisation du signal est de fournir et d'extraire des informations caractéristiques et pertinentes du signal. Ces informations sont restituées sous la forme d'une suite discrète de vecteurs, appelés *vecteurs acoustiques* ou *vecteurs d'observations*. Chaque vecteur contient un nombre fini de paramètres représentant les caractéristiques d'un segment du signal. La concaténation de ces vecteurs fournit une représentation discrète et paramétrique du signal à traiter [d'allessandro 92]. La conversion du signal en séquence de vecteurs d'observations est régie par un modèle paramétrique caractérisant le point de vue sous lequel le signal est observé. La paramétrisation consiste à estimer les paramètres de ce modèle. Ces modèles peuvent être classés en quatre familles :

Les modèles articulatoires

Ils permettent d'extraire les informations régissant le mécanisme de phonation. Cette paramétrisation s'appuie sur un formalisme issu de la mécanique des fluides puisque l'onde acoustique que nous produisons en parlant résulte de la circulation d'un flux d'air au travers du conduit vocal constitué d'articulateurs. Les paramètres extraits codent la position des différents articulateurs (position des lèvres, ouverture de la bouche, protusion, position de la langue, etc).

Les modèles de production

Ils permettent de réaliser une simulation de l'équivalent électrique de l'appareil phonatoire. Ces modèles sont une simplification (ou approximation) des modèles articulatoires. On trouve dans cette catégorie, les codages LPC (*Linear Prediction Coding*) et AR (*AutoRegressive coding*).

Les modèles phénoménologiques

Ces modèles tentent de modéliser le signal indépendamment de la façon dont il a été produit. Les modèles basés sur l'analyse de Fourier en sont un exemple. Ils proposent des représentations du signal basées sur une analyse fréquentielle de celui-ci. Parmi les paramétrisations dérivées de ces modèles, nous détaillerons dans le paragraphe suivant la paramétrisation spectrale. Cette paramétrisation présente l'avantage de fournir une représentation temps-fréquence (spectrogramme) du signal pour laquelle les énergies des différents signaux constituant une scène auditive peuvent être considérées comme additives.

Les modèles d'audition

Ces modèles tentent de mettre à profit les connaissances acquises sur la perception des sons et sur le fonctionnement de notre système auditif afin d'améliorer la robustesse des modèles précédents. Par exemple l'introduction de connaissances issues de la psychoacoustique dans l'estimation

des modèles AR ou spectraux a conduit respectivement aux analyses PLP (Perceptual Linear Prediction) et MFCC (Mel Frequency Cepstral Coefficient).

1.1.2.2 Le spectrogramme

Le signal de la parole étant variable au cours du temps, l'extraction des vecteurs d'observation est généralement faite sur des fenêtres d'analyse temporelles de faible durée (de l'ordre de quelques dizaines de millisecondes), de telle sorte que le signal puisse être considéré comme stationnaire sur chacune d'elles. De nombreuses fenêtres ont été étudiées en traitement du signal (Hamming, Hanning, Kaiser,etc). La fenêtre la plus utilisée en reconnaissance de la parole est la fenêtre de Hamming, illustrée par la figure 1.2 et définie par l'équation :

$$h(n) = \begin{cases} 0.54 \ - \ 0.46 \ \cos(2\pi \frac{n}{N-1}) \text{ si } 0 \leq n \leq N-1 \\ 0 \text{ sinon} \end{cases}$$

où N est la taille de la fenêtre en nombre d'échantillons du signal. Par ailleurs, un filtre de préaccentuation très simple est souvent appliqué au signal pour renforcer les sons aigus, toujours plus faibles en énergie que les sons graves.

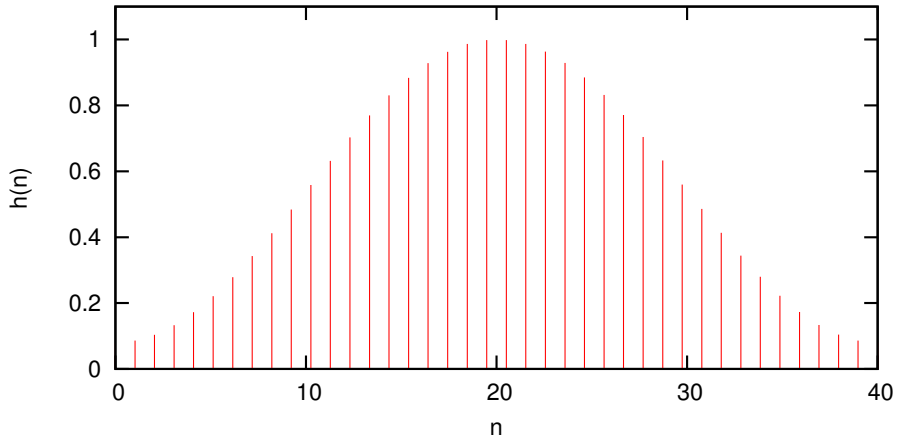


FIG. 1.2 – *Fenêtre de Hamming h(n).*

Une représentation spectrale d'un signal acoustique est une concaténation de spectres à court terme. Un spectre à court terme, appelé également spectre instantané, est le résultat de la transformée de Fourier sur une fenêtre d'analyse telle la fenêtre de Hamming. Dans le cas d'un signal discret, comme celui de la parole une fois échantillonnée, le spectre à court terme $S_N(f)$ peut s'écrire pour une fenêtre $h(.)$ centrée sur m :

$$S_N(f) \ = \ \sum_{n=0}^{n=N} s_m(n) \ h(n-m) \ e^{-i2\pi fn}$$

La concaténation des spectres à court terme successifs obtenus par glissement de la fenêtre d'analyse forme un spectrogramme qui représente l'évolution dans le plan temps-fréquence de l'énergie du signal, comme l'illustre la figure 1.3. Dans le but de limiter les effets de bord et

de réduire les discontinuités, les fenêtres d'analyse successives se recouvrent en partie (le plus souvent de moitié) et sont aplaties à leurs extrémités.

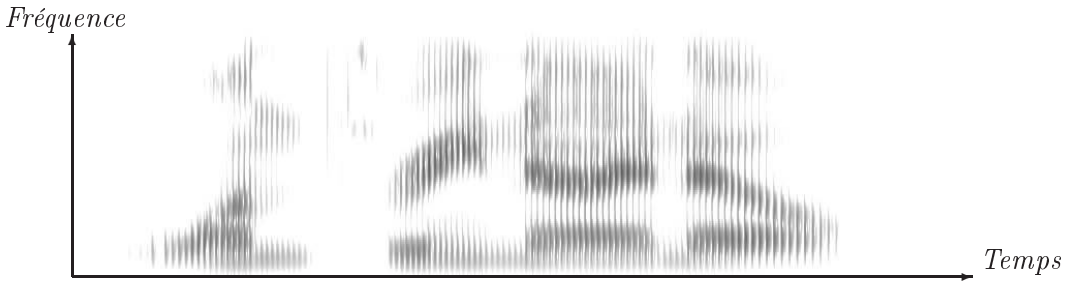


FIG. 1.3 – *Représentation spectrale d'un signal de la parole correspondant à la phrase : "one three nine oh".*

Des études perceptives ont montré que notre oreille possède une meilleure résolution pour les sons de basses fréquences que pour les sons de hautes fréquence. A cet égard, le spectrogramme obtenu par la transformation de Fourier à court terme est, en reconnaissance de la parole, souvent perçu comme un ensemble de signaux temporels contenant une partie de l'information sur le signal vocal dans chacune des bandes de fréquences d'un banc de filtres. Les fréquences centrales des filtres sont déterminées de manière à restituer la résolution fréquentielle de notre oreille. Les deux principales échelles perceptives sont les échelles *Bark* et *Mel*.

Un Bark correspond à la largeur d'une bande critique, qui croît proportionnellement à sa fréquence centrale. Cette échelle correspond au fait que l'oreille possède une bonne résolution spectrale en basses fréquences et médiocre en hautes fréquences.

$$B_{ark} \;=\; 13\; Arctg\left(\frac{0.76\; F_{Hz}}{100}\right) \;+\; 3.5\; Arctg\left(\frac{F_{Hz}}{7500}\right)^2$$

L'échelle Mel est linéaire jusqu'à 100 Hz et logarithmique au-delà. Une expression analytique possible est la suivante [O'Shaugnessy 00] :

$$M_{Mel} \;=\; 2595\; log_{10}\left(1 \;+\; \frac{F_{Hz}}{700}\right)$$

La plupart des systèmes actuels de reconnaissance de la parole fondent leur analyse sur cette échelle. La figure 1.4 représente un banc de filtres à échelle Mel.
Le groupement des énergies des spectres à court terme basée sur l'échelle Mel fournit une représentation Mel spectrale du signal, illustrée par la figure 1.5.

1.1.3 Principe de la reconnaissance

Considérons une séquence de vecteurs d'observations O correspondant à la prononciation d'une séquence de mots W. Le principe même de la RAP est de parvenir à déterminer W à

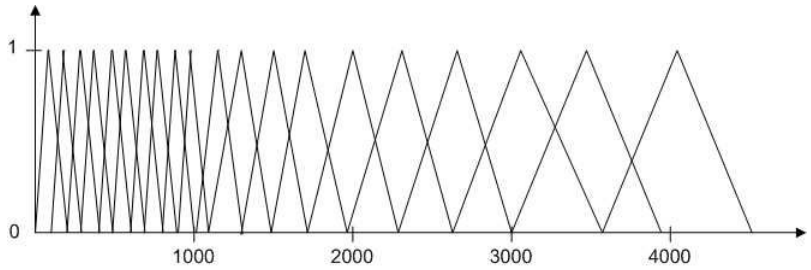


FIG. 1.4 – *Banc de filtres à échelle Mel.*

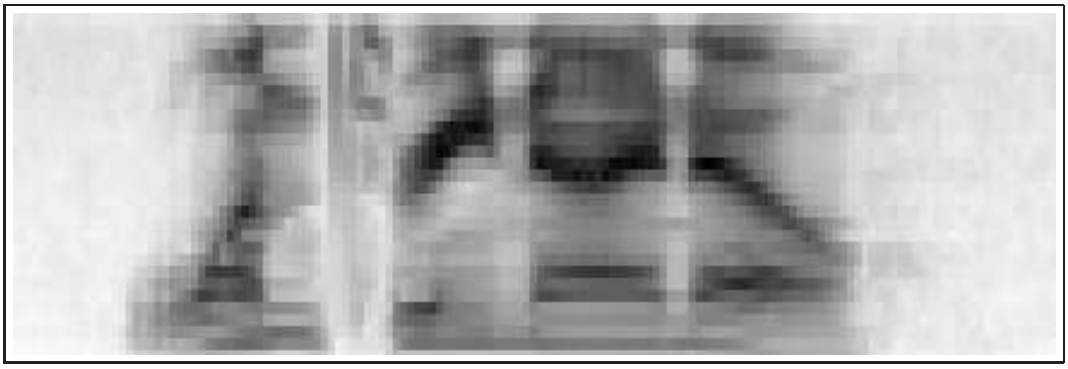


FIG. 1.5 – *Représentation Mel spectrale d'un signal parole. Chaque trame est exprimée ici par 32 coefficients Mel spectraux.*

partir des observations O. Trois stratégies de reconnaissance peuvent être envisagées plus une quatrième résultante de la combinaison des premières.

1. La reconnaissance à base d'exemples
2. La reconnaissance probabiliste
3. La reconnaissance par surfaces de décision et fonctions discriminantes
4. L'hybridation de modèles

Nous proposons dans le paragraphe suivant un bref survol de ces approches.

1.1.3.1 Reconnaissance à base d'exemples

Cette approche est la plus intuitive. Elle repose sur l'idée qu'une même classe regroupe des objets de formes similaires. De cette façon, il est possible à partir d'une métrique judicieusement choisie de mesurer la similitude entre deux formes. La reconnaissance d'une forme inconnue revient à comparer cette forme avec les formes représentatives des différentes classes. La comparaison de deux formes s'appuie sur des méthodes de comparaison élastique, fondées sur la programmation dynamique et fournissant une solution optimale au recalage temporel nécessaire à ce type de classification. Cette approche fut une des premières proposées dans le cadre de la RAP et était initialement dédiée à la reconnaissance de mots isolés [Vintsyuk 68, Sakoe 71].

Elle fut étendue par la suite afin de reconnaître des séquences continues de mots [Sakoe 79] et plus récemment à de la parole continue [Wachter 07]. Les résultats publiés dans [Wachter 07] montrent que cette approche peut rivaliser avec les meilleurs systèmes actuels. Cependant la reconnaissance à base d'exemples reste très coûteuse en temps et en mémoire car la forme à reconnaître doit être comparée à tous les exemples contenus dans la base des références.

1.1.3.2 Classification probabiliste

Cette catégorie de classification se base sur la connaissance des distributions des différents paramètres de chaque classe. Considérons un objet X devant être classé dans une des K classes C_k pour lesquelles un modèle paramétrique est disponible. La meilleure classe $C^\star$ est celle qui minimise le risque bayésien $R(C_i|X)$:

$$\begin{aligned} C^\star &= \arg\min_{C_i} R(C_i|X) \\ R(C_i|X) &= \sum_{k=1}^{K} L(C_i, C_k).P(C_k|X) \end{aligned}$$

$L(C_i, C_k)$ est le risque de mauvaise classification, ou encore le coût occasionné par la fait d'assigner la classe C_k à X sachant que X appartient à la classe C_i. $P(C_k|X)$ est la probabilité *a posteriori* de l'appartenance de X à C_k.

Pour des distributions de probabilités $P(X|C_i)$ connues (modèles paramétriques des classes C_i), la règle de Bayes est optimale dans le sens où aucune autre règle ne donnera un risque plus faible. Considérant le cas d'un risque binaire, c'est à dire assignant un coût de 1 à toute mauvaise classification et un coût de 0 à toute classification juste, minimiser le risque bayésien est équivalent à assigner à X la classe pour laquelle X a la plus forte probabilité d'appartenir (critère du Maximum *a posteriori* : MAP) :

Affecter à X la classe C_i si :

$$\begin{aligned} P(C_i|X) &> P(C_j|X) \quad \forall j \neq i \\ P(X|C_i).P(C_i) &> P(X|C_j).P(C_j) \quad \forall j \neq i \end{aligned} \tag{1.1}$$

Une description plus détaillée de ce formalisme est présentée en annexe A.2. En général, les probabilités *a priori* $P(C_i)$ de chaque classe ne sont pas connues et doivent donc être estimées à partir d'une base d'apprentissage. Les modèles de Markov cachés, décrits au paragraphe 1.1.4, s'appuient sur ce formalisme.

1.1.3.3 Surfaces de décision et fonctions discriminantes.

Les paramètres o d'objets appartenant à une même classe sont, s'ils sont bien choisis, localisés dans une région homogène de l'espace des paramètres. La classification par surfaces de décision

ou fonctions discriminantes repose sur un partitionnement de l'espace des paramètres en régions mutuellement exclusives ; chaque région de l'espace correspondant aux réalisations possibles des paramètres d'une même classe. Une région associée à la classe C_k est donc caractérisée par une fonction discriminante $g_k(o)$ tel que :

$$g_k(o) > g_j(o) \quad \forall o \in C_k \quad \text{et} \quad \forall j \neq k$$

Si les fonctions discriminantes sont des fonctions linéaires des paramètres, alors les régions sont séparées linéairement par des hyperplans. En pratique ce cas est rare et les surfaces de décision sont approchées en considérant que celles-ci sont linéaires par morceaux. Cette famille de méthodes de classification regroupe entre autres :

Le perceptron

Le perceptron est un réseau de neurones formels entrant dans la catégorie des modèles neuromimétiques. Il produit une classification par fonction linéaire dans le cas où il est constitué d'une seule couche de neurones [Rosenblatt 62]. Cependant la fonction discriminante peut être complexifiée par adjonction de couches.
Un neurone formel est une représentation mathématique d'un neurone biologique. Les actions excitatrices et inhibitrices des synapses sont représentées, la plupart du temps, par des coefficients numériques associées aux entrées. Les valeurs numériques sont ajustées automatique dans une phase d'apprentissage. Dans sa version la plus simple, un neurone formel calcule la somme pondérée de ses entrées, puis applique à cette valeur une fonction d'activation, généralement non linéaire. La valeur finale obtenue est la sortie du neurone. Individuellement, les neurones formels calculent des fonctions linéaires mais leur mise en réseau permet de simuler des fonctions très complexes.

La machine à vecteur support (SVM)

Une machine à vecteur support (SVM en anglais pour Support Vector Machine) consiste à séparer deux ensembles de points par un hyperplan. L'idée originale des MVS a été publiée par Vladimir Vapnik [Vapnik 82, Vapnik 98]. Elle est basée sur l'utilisation de fonctions dites *noyaux* qui permettent une séparation optimale (sans problème d'optimum local) des points de l'espace en différentes classes. Le principe est de projeter l'espace des paramètres sur un espace de plus grande dimension à l'aide de la fonction noyau de manière à pouvoir séparer linéairement les points exprimés dans ce nouvel espace. Les SVM ont été développés initialement dans le cadre d'une classification bi-classes, mais des extensions multi-classes ont été proposées, comme la M-SVM [Guermeur 05]. Les SVM ont été introduites récemment pour la reconnaissance de la parole et ont donné des résultats prometteurs, notament pour l'identification du locuteur [Wan 05b, Wan 07], la reconnaissance de formes acoustiques [Wan 05a, Bernal-Chaves 05, Scharenborg 06], la détection de mots-clés [Aye 02, Keshet 07] ainsi que pour la conception de modèles hybrides [Ganapathiraju 00].

Les arbres de décision

Les arbres de décision [Breiman 84] également appelés arbres de classification sont également des classifieurs non linéaires par surfaces séparatrices. Le principe est de déterminer la classe d'appartenance d'une forme par une suite de tests sur ses paramètres. Un arbre de décision est ainsi formé d'un ensemble de nœuds internes contenant des tests et par des feuilles représentant chaque classe. L'identification de la classe d'une forme est donnée par un chemin depuis la racine de l'arbre jusqu'à une feuille. Plusieurs progiciels d'arbres de décision tels CART ou C4.5 sont disponibles et ont été utilisés en reconnaissance de la parole.

1.1.3.4 Modèles hybrides

Les modèles de classification probabiliste et en particulier les HMM compte parmi les différentes approches de classification les plus utilisés en reconnaissance de la parole. L'intérêt qu'ils suscitent provient non seulement du fait qu'ils donnent de bonnes performances mais aussi du fait qu'ils sont particulièrement bien adaptés au traitement de données à évolution temporelle. Cependant, ils sont peu discriminants en raison d'un apprentissage dit au *maximum de vraisemblance* (voir paragraphe 1.1.4.2). Certes il existe des méthodes d'apprentissage discriminant mais l'apprentissage au maximum de vraisemblance reste le standard. Par conséquent des systèmes hybrides ont été proposés, combinant des HMM avec des modèles discriminants.

Parmi ces hybridations, nous pouvons citer des modèles combinant HMM et réseaux de neurones. De tels modèles utilisent un réseau de neurones comme préprocesseur [Lazli 02] ou post-processeur [Guo 93] d'un HMM. Dans le premier cas un perceptron est entraîné pour apprendre les probabilités *a posteriori* des classes phonétiques $P(S_i|O)$, S_i étant un état d'un HMM et O un vecteur d'observations. La formule de Bayes permet à partir de ces probabilités de calculer la vraisemblance des observations. Ces vraisemblances sont alors utilisées en lieu et place de celles initialement calculées par les modèles à mélange de gaussiennes utilisés par un HMM classique. Dans le deuxième cas, toutes les hypothèses de reconnaissance (ou seulement les N meilleures) calculées par le HMM sont mises en entrée du réseau. Le réseau distinguera alors, parmi ces hypothèses, la meilleure d'entre elles.

Une autre hybridation consiste à combiner un HMM avec une SVM [Ganapathiraju 00]. Une telle hybridation présente également l'avantage de combiner la capacité des HMM à modéliser des séries temporelles et le pouvoir discriminant des SVM. Ce système hybride possède la même architecture que le système de Lazli et Sellami [Lazli 02] mais le réseau de neurones calculant $P(S_i|O)$ est remplacé par une SVM.

1.1.4 Le modèle de Markov caché

Les modèles de Markov cachés (HMM : Hidden Markov Model) ont été décrits pour la première fois dans une série de publications de statistique par Leonard E. Baum [Baum 70,

Baum 72]. Ce n'est qu'en 1975 qu'ils ont été proposés dans le cadre de la reconnaissance automatique de la parole [Baker 75a, Baker 75b] et se sont imposés depuis comme modèles de référence dans ce domaine. Nous proposons dans les paragraphes suivants de définir ce qu'est un HMM et de décrire sa mis en œuvre dans le cadre de la reconnaissance automatique de la parole.

1.1.4.1 Définition d'un HMM

Un HMM est un cas particulier des modèles stochastiques graphiques, et peut être vu comme un automate probabiliste. Il est généralement caractérisé par un quadruplet (S, Π, A, B) :

- $S = \{S_0, \ldots, S_i, \ldots, S_k\}$ est l'ensemble des états de l'automate.
- $\Pi = \{\pi_0, \ldots, \pi_i, \ldots, \pi_k\}$, avec π_i étant la probabilité que S_i soit l'état initial.
- A est l'ensemble des probabilités de transition d'un état vers un autre. A est caractérisé par une matrice $k*k$ d'éléments a_{ij} avec $i\ et\ j \in [0, k]$ et k le nombre d'états. Tout élément a_{ij} de cette matrice est la probabilité d'atteindre l'état S_j au temps t sachant que nous étions dans l'état S_i au temps $t-1$.
- B est un ensemble de lois de probabilité $b_i(o)$ donnant la probabilité $P(o|S_i)$ que l'état S_i ait généré l'observation o. Cette probabilité est la vraisemblance de l'observation au regard de S_i.

Un HMM étant un automate probabiliste, les contraintes suivantes doivent être respectées :

1. La somme des probabilités des états initiaux doit être égale à 1 :

$$\sum_i \pi_i = 1$$

2. La somme des probabilités des transitions sortant d'un état doit être égale à 1 :

$$\forall i \quad \sum_j a_{ij} = 1$$

3. La somme des probabilités des émissions d'un état doit être égale à 1 :

$$\forall i \quad \sum_o b_i(o) = 1 \qquad \text{dans le cas d'observations discrètes.}$$

$$\forall i \quad \int_o b_i(o)\ \mathrm{d}o = 1 \qquad \text{dans le cas d'observations continues.}$$

Un HMM représente un objet par deux suites de variables aléatoires : l'une dite *cachée* et l'autre *observable*. La suite observable correspond à la suite d'observations $o_1, o_2, \ldots, o_T$ où les o_i sont des vecteurs d'observations du signal à reconnaître. La suite cachée correspond à une suite d'états $q_1, q_2, \ldots, q_T$, où les q_i puisent leurs valeurs parmi l'ensemble des N états du modèle $\{S_1, S_2, \ldots, S_N\}$. La suite observable est définie comme une réalisation particulière de la suite cachée. L'objectif est de déterminer la meilleure séquence d'états $Q^\star = (q_1^\star, q_2^\star, \ldots, q_T^\star)$ à partir

de la séquence d'observations $O = (o_1, o_2, \ldots, o_T)$. Le meilleur chemin $Q^\star$ est celui qui maximise la probabilité *a posteriori* $P(Q|O)$ (critère du maximum *a posteriori* : Eq. 1.1). En effet, en dérivant cette probabilité *a posteriori* par la règle de Bayes, il vient :

$$\begin{aligned} Q^\star &= \arg\max_Q P(Q|O) \\ &= \arg\max_Q \frac{P(O|Q)\ P(Q)}{P(O)} \end{aligned}$$

$P(O)$ étant constant pour tout Q :

$$Q^\star = \arg\max_Q P(O|Q)\ P(Q) \tag{1.2}$$

Un HMM présente plusieurs avantages : il s'inscrit dans un formalisme mathématique bien établi, il bénéficie de méthodes d'apprentissage automatique des ses paramètres et il est particulièrement bien adapté à la modélisation de processus à évolution temporelle.

1.1.4.2 Mise en œuvre

La mise en œuvre d'un système de reconnaissance de la parole à partir de HMM nécessite de formuler quelques hypothèses simplificatrices dans le but d'adapter le cadre théorique des HMM à la RAP mais aussi d'en simplifier le formalisme mathématique et ainsi proposer des algorithmes d'apprentissage et de classification optimaux sous ces hypothèses. Une fois ces hypothèses posées, trois points importants sont à considérer pour la reconnaissance de la parole à partir de HMM :

1. **La topologie du modèle :**
   Comment définir le nombre d'états du modèle ? Quelles transitions entre les états sont permises ? quelles lois de probabilité utiliser pour modéliser la distribution des paramètres de chaque état ?
2. **L'apprentissage des paramètres :**
   Étant donné un ensemble de J séquences d'observations O_j représentant chacune la même entité acoustique et donc associées au même HMM M_j, comment choisir les paramètres Λ_j de M_j afin de maximiser la probabilité que M_j engendre la suite d'observations O_j ?
3. **Le décodage :**
   Étant donnée une séquence d'observations O, et un ensemble de HMM, quelle est la séquence de modèles qui maximise la probabilité de généré O ?

Nous décrivons dans les paragraphes suivant la manière dont ces points sont traités dans le cadre de la reconnaissance automatique de la parole.

Hypothèses simplificatrices

Soit $O = (o_1, o_2, \ldots, o_T)$ une suite de T observations. Soit $Q = (q_1, q_2, \ldots, q_T)$ une séquence d'états alignée avec la suite d'observations ; au temps t le HMM est dans l'état q_t engendrant

l'observation o_t.

Hypothèse n° 1
La probabilité qu'une observation o_t soit émise au temps t ne dépend pas des observations antérieures.

$$P(o_t|q_t, q_{t-1} \ldots, q_1, o_{t-1}, o_{t-2}, \ldots, o_1) = P(o_t|q_t, q_{t-1}, \ldots, q_1) \tag{1.3}$$

Hypothèse n° 2
La probabilité qu'une observation soit émise au temps t ne dépend pas des états précédemment visités, mais seulement de l'état courant.

$$P(o_t|q_t, q_{t-1}, \ldots, q_1) = P(o_t|q_t) \tag{1.4}$$

Hypothèse n° 3
La probabilité que le HMM soit dans l'état q_t à l'instant t ne dépend que de l'état dans lequel il se trouvait à l'instant $t-1$.

$$P(q_t|q_{t-1}, q_{t-2}, \ldots, q_1) = P(q_t|q_{t-1}) \tag{1.5}$$

Un modèle respectant cette dernière hypothèse est appelé modèle de Markov du premier ordre par opposition aux modèles d'ordre N. Un modèle d'ordre N est un modèle pour lequel la probabilité de se trouver dans un état q_t est conditionnée par la suite d'états $q_{t-1}, q_{t-2}, \ldots, q_{t\_N}$. Un HMM du second ordre a été mis au point [Gong 94]. Les résultats obtenus en reconnaissance de chiffres ont montré une légère amélioration par rapport aux modèles du premier ordre. Cependant cette amélioration se fait au détriment d'une complexification accrue du modèle ce qui limite son intérêt.

Topologie du modèle

Le nombre d'états d'un HMM dépend de l'entité acoustique qu'il modélise. L'entité la plus répandue est le phonème, mais il est possible de considérer des entités plus grandes (supra-phonétique), comme la syllabe ou le mot. Cependant construire un système possédant un modèle pour chaque mot d'une langue n'est pas envisageable pour des raisons de temps et d'espace de calcul mais aussi pour des raisons de taille de la base d'apprentissage devant contenir suffisamment d'exemples de chaque mot pour obtenir des modèles fiables. Une telle modélisation est alors inconcevable pour des systèmes grand vocabulaire permettant de reconnaître plusieurs dizaines de milliers de mots différents. Néanmoins sous certaines contraintes comme l'utilisation d'un vocabulaire restreint cette modélisation peut s'avérer avantageuse notamment pour la modélisation des phénomènes de co-articulation.

Un phonème est généralement décomposé en 3 parties : un début, une partie stable et une fin. Une topologie à 3 états est par conséquent utilisée. Le second état correspondant à la partie stable est l'état caractérisant le mieux le phonème alors que le premier et dernier état modélisent

les effets de la co-articulation, c'est à dire les transitions entre phonèmes. Ceux-ci correspondent donc aux parties instables du phonème car elles sont influencées par le contexte gauche et droit. Dans le but de restituer l'évolution temporelle du signal de la parole une topologie gauche-droite est adoptée dans la grande majorité des cas. Ceci veut dire qu'aucun retour en arrière n'est possible.

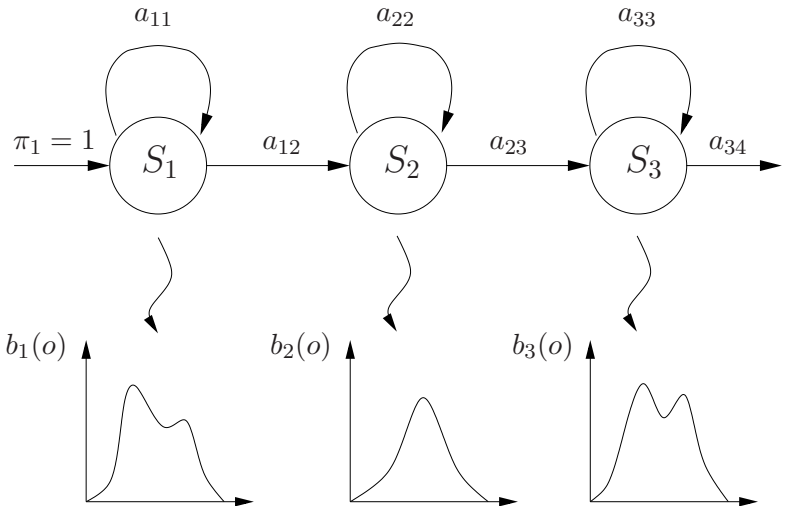


FIG. 1.6 – *HMM gauche-droite à 3 états usuellement utilisé pour la modélisation de phonèmes. Les lois de probabilité $b_i(o)$ fournissant les probabiltés qu'une observation o ait été générée par un état S_i sont modélisées par des modèles à mélange de gaussiennes (GMM).*

Chaque état S_i d'un HMM renvoie pour une observation o la probabilité que o ait été générée par S_i. Le calcul de cette probabilité appelée également vraisemblance de l'observation s'appuie sur une fonction de densité de probabilités $b_i(0)$. Cette fonction $b_i(0)$ est un modèle paramétrique de l'ensemble des observations pouvant être générées par l'état S_i. La plupart des systèmes s'appuient des densités de probabilités continues modélisée par un mélange de lois normales (distribution gaussienne des observations). La vraisemblance d'une observation o est donc donnée par :

$$b_i(o) = \sum_{j=1}^{N_\lambda} \lambda_j \, \mathcal{N}(o; \mu_j, \Sigma_j) \tag{1.6}$$

avec

$$\mathcal{N}(o; \mu_j, \Sigma_j) = \frac{1}{\sqrt{(2\pi)^M \ |\Sigma_j|}} \ \exp\Big(-\frac{1}{2}(o-\mu_j)'\Sigma_j^{-1}(o-\mu_j)\Big) \tag{1.7}$$

N_λ est le nombre de gaussiennes, λ_j est le poids le la $j^{\text{ième}}$ gaussienne, μ_j et Σ_j sont respectivement le vecteur moyen et la matrice de covariance de la $j^{\text{ième}}$ gaussienne et M la dimension du vecteur d'observations. La figure 1.6 présente un HMM gauche-droite à 3 états utilisé pour la modélisation de phonèmes.

Apprentissage

Considérons un ensemble de HMM M_j et un ensemble de T observations O_j. Apprendre les paramètres des HMM revient à chercher le meilleur ensemble de paramètres $\Lambda_j^\star = (\mu_j^\star, \Sigma_j^\star)$ tel que la probabilité que O_j ait été générée par M_j soit maximale (critère du maximum de vraisemblance).

$$\Lambda_j^\star = \arg\max_{\Lambda_j} \prod_{t=1}^{T} P(O_j(t)|M_j, \Lambda_j) \tag{1.8}$$

Idéalement, c'est $P(M_j|O_j, \Lambda_j)$ qui devrait être maximisée. L'apprentissage serait alors plus discriminant : lorsque la vraisemblance du modèle j augmente pour les exemples correspondant au modèle j, les vraisemblances des autres modèles devraient diminuer pour ces mêmes exemples. Les HMM devraient donc être entraînés, non seulement pour maximiser la probabilité de générer les exemples de sa propre classe, mais aussi pour les discriminer par rapport aux autres classes (critère du maximum *a posteriori*). Parce qu'il n'existe pas de méthode permettant de maximiser directement $P(O_j|M_j, \Lambda_j)$, les paramètres des modèles sont obtenus en maximisant l'équation 1.8 par la méthode itérative de Baum et Welch [Baum 72], qui est un cas particulier de l'algorithme EM (Expectation Maximisation) [Dempster 77].

Décodage

Le décodage de la parole par des modèles HMM revient à déterminer la meilleure séquence d'états $Q^\star = (q_1^\star, q_2^\star, \ldots, q_T^\star)$ pouvant engendrer la séquence d'observations $O = (o_1, o_2, \ldots, o_T)$:

$$\begin{aligned} Q^\star &= \arg\max_{Q} P(O|Q) \\ &= \arg\max_{Q} \ \pi_0 \prod_{t=1}^{T} a_{q_{t-1}q_t}.b_{q_t}(o_t) \end{aligned} \tag{1.9}$$

Une solution naïve est de calculer la probabilité $P(O|Q)$ de toutes les séquences d'états Q possibles et de ne retenir que la meilleure. Ceci peut se faire en construisant un arbre. A chaque temps t une couche de nœuds internes est ajoutée à l'arbre. Chaque nœud interne représente un état particulier des modèles et contient la probabilité de se trouver dans cet état à l'instant t. Les probabilités des différentes hypothèses de reconnaissance sont contenues dans les feuilles de cet arbre. Cependant une telle solution est en pratique inapplicable car le nombre d'hypothèses est très grand.

L'algorithme de Viterbi, variante stochastique de la programmation dynamique, propose de simplifier l'arbre au fur et à mesure de sa construction. En effet, lors de son déroulement on se trouve rapidement avec des branches proposant les mêmes substitutions, mais avec des probabilités différentes. Plusieurs hypothèses peuvent se retrouver dans le même état au même instant. L'algorithme de Viterbi stipule qu'il n'est pas nécessaire de dérouler les hypothèses de plus faible probabilité car elles ne peuvent plus être candidates pour décrire le message de plus probable.

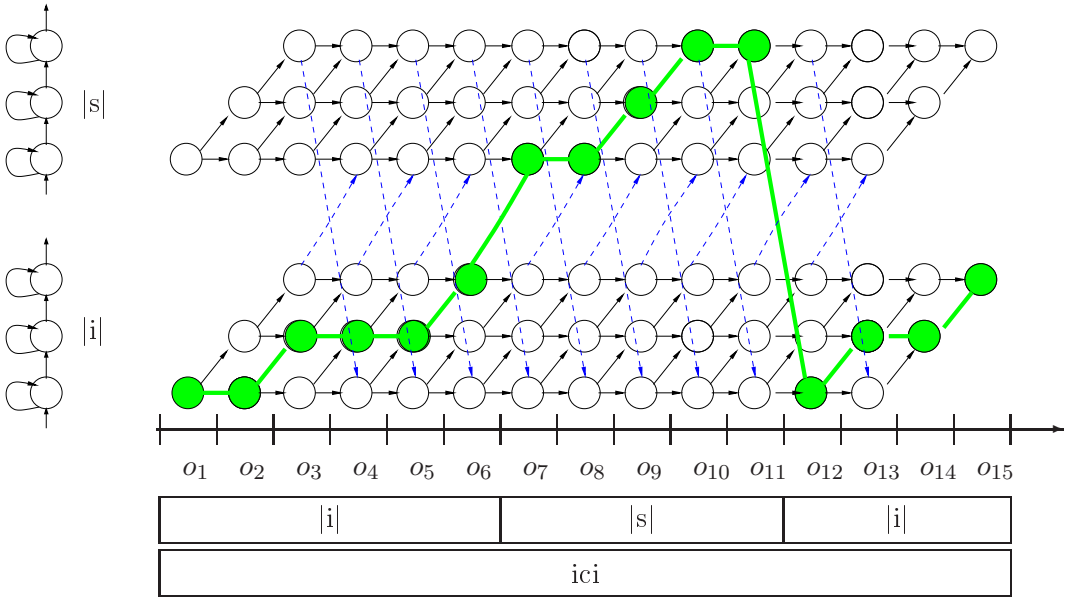


FIG. 1.7 – *Illustration de la reconnaissance de la parole par l'algorithme de Viterbi. La phrase reconnue correspond à l'hypothèse de reconnaissance (ou chemin) ayant la plus forte probabilité dans le treillis des hypothèses. Pour cet exemple la meilleure hypothèse correspond à la succession de phonèmes /i/ /s/ /i/ qui est la transcription phonétique du mot "ici".*

La mise en œuvre de cet algorithme consiste à construire de façon itérative la meilleure séquence d'états à partir d'un tableau $T * N$ (T : nombre d'observations, N : nombre d'états total des modèles) appelé *treillis des hypothèses* où chacun des nœuds (t, i) contient la vraisemblance $\delta_i(o_t)$ du meilleur chemin passant par l'état i à l'instant t. La vraisemblance $\delta_i(o_T)$ du meilleur chemin qui finit à l'état i au temps T est alors calculée par récurrence :

1. **Initialisation** : $\delta_i(o_1) = \pi_i$
2. **Récursion** : pour se trouver dans l'état i à l'instant t, le processus markovien se trouvait forcément dans un état j à l'instant $t-1$ pour lequel une transition vers l'état i est possible : $a_{ji} > 0$. D'après le principe d'optimalité de Bellman, $\delta_i(o_t) = \max_j \left(\delta_j(o_{t-1}) \,.\, a_{ji} \right) . b_i(o_t)$.
3. **Terminaison** : La vraisemblance des observations correspondant à la meilleure hypothèse est obtenue en recherchant l'état i qui maximise la valeur $\delta_i(o_T)$ à la dernière observation o_T :

$$P(O|Q^\star) = \max_i \left(\delta_i(o_T) \right)$$

Une représentation visuelle fictive de l'algorithme de Viterbi est présenté par la figure 1.7. Cette figure montre un treillis des hypothèses construit pour une séquence de 15 observations. Seulement deux modèles sont représentés ici ; deux HMM gauche-droite à 3 états modélisant les phonèmes |i| et |s|. Le meilleur chemin (en vert) correspond à la séquence de phonèmes : |i| |s| |i|. Cette séquence est la transcription phonétique du mot « ici ».

1.1.4.3 Limitation des HMM

L'utilisation des HMM en reconnaissance automatique de la parole repose sur plusieurs hypothèses simplificatrices. Celles-ci sont, certes, nécessaires, mais elles constituent également des points faibles des HMM.
La modélisation de la durée des phonèmes n'est qu'implicitement contenue au travers des probabilités de transitions entre les états. Une modélisation explicite de celle-ci a cependant été proposée avec succès [Russel 85, Levinson 86].
L'hypothèse d'indépendance conditionnelle des observations (équation 1.3) est irréaliste. Une solution efficace et largement répandue consiste à prendre en compte les dérivées premières Δ et secondes $\Delta\Delta$ des paramètres. Une deuxième solution est de modéliser explicitement la corrélation entre les vecteurs d'observations successifs [Russell 93, Gales 93b].

1.2 Robustesse au bruit

Malgré de nombreux efforts de recherche entrepris depuis plusieurs années, la robustesse des systèmes de reconnaissance de la parole au bruit reste problématique, ce qui explique probablement en grande partie leur diffusion et utilisation très limitée. Ce paragraphe définit les différents types de bruit et résume brièvement les grandes familles d'approche qui ont été proposées jusqu'alors pour résoudre ce problème de robustesse.

1.2.1 Le bruit

L'objectif d'un système de reconnaissance est de retranscrire ce qu'a prononcé un locuteur particulier. Nous considérons comme bruit toute distorsion du signal ou tout signal provenant d'une autre source sonore que le locuteur principal. On distingue deux types de bruits. Le bruit convolutif, conséquence de la distorsion du signal inhérent à l'acquisition par un microphone de mauvaise qualité, ou induite par les caractéristiques du canal de transmission comme les lignes téléphoniques et le bruit additif correspondant à une pollution sonore issue d'autres sources.

Le bruit est très pénalisant pour la reconnaissance. En effet les modèles acoustiques sont appris sur des corpus enregistrés en conditions maîtrisées, c'est-à-dire exempts de bruit. Ils ne représentent donc que les caractéristiques du signal de la parole. Ces modèles ne sont alors plus du tout adaptés pour reconnaître un signal de parole noyé dans le bruit ou subissant des distorsions.

Nous ne considérons par la suite que les bruits additifs.

Un bruit peut être caractérisé par différentes propriétés (TAB. 1.1). La connaissance de ces propriétés du bruit permet d'adopter une stratégie robuste adaptée.

Propriétés	Attributs de la propriété
structure temporelle	continu / impulsif / périodique
stationnarité	stationnaire / non-stationnaire
structure spectrale	large-bande / confiné en bande
dépendance avec la parole	corrélé / décorrélé
spatialisation	cohérent / incohérent avec la source de la parole
harmonicité	harmonique / inharmonique

TAB. 1.1 – *Propriétés caractérisantes du bruit (adapté de [Glotin 01]).*

Une des situations les moins pénalisantes en reconnaissance est de traiter un signal pollué par un bruit continu, stationnaire, décorrélé du signal de la parole et inharmonique ; un bruit blanc gaussien par exemple. Une situation beaucoup plus pénalisante est de reconnaître un signal de parole parmi d'autres signaux de parole. Une telle interférence est connue sous le nom de "cocktail party". La figure 1.8 illustre l'altération d'un spectrogramme de parole par du bruit.

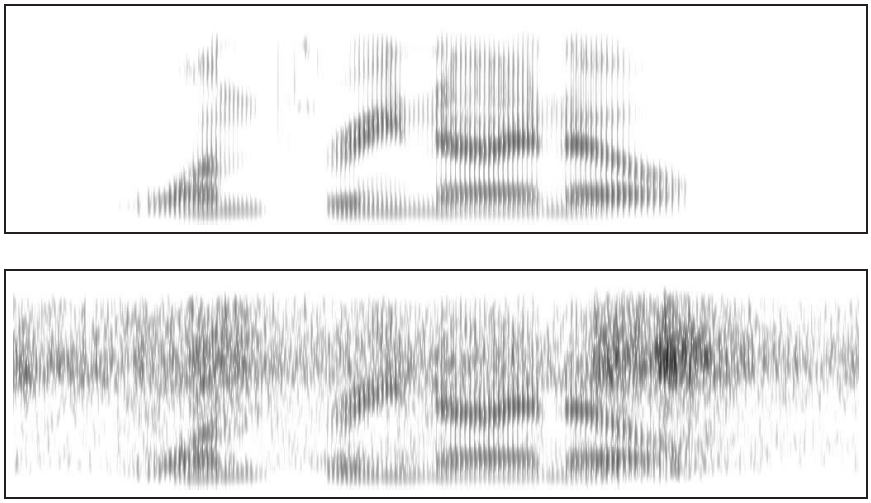


FIG. 1.8 – **Haut** : *Représentation spectrale d'un signal de la parole correspondant à la phrase : "one three nine oh".* **Bas** : *Le même signal, mais corrompu par le bruit du métro à 5 dB.*

Il est possible de quantifier le niveau de bruit dans une phrase. Le bruit est mesuré par le rapport signal sur bruit (SNR : Signal-to-Noise Ratio en anglais). Le SNR s'exprime en décibels (dB) :

$$SNR \;=\; 10 \;\log_{10} \frac{S}{N}$$

où S et N sont respectivement les énergies du signal de la parole et du bruit. Il est possible de calculer le SNR à différents niveaux de granularité. On peut donc distinguer :

1. **SNR global** : Le rapport est calculé en fonction des énergies totales sur la phrase de la parole et du bruit. La quantité de bruit est alors quantifiée par une seule valeur.
2. **SNR segmental** : Le rapport est calculé sur des segments temporels bien définis du signal. Le bruit est alors quantifié pour chacun des segments.
3. **SNR segmental temporel et fréquentiel** : Le calcul de ce rapport est identique que pour le précédent, cependant le calcul est effectué pour chaque bande de fréquence.
4. **SNR local** : Pour chaque coefficient du plan temps-fréquence (spectrogramme) un SNR est calculé. C'est la granularité la plus fine, mais aussi la plus délicate à estimer.

1.2.2 Stratégies pour la reconnaissance robuste de la parole

Plusieurs stratégies robustes de reconnaissance de la parole ont été proposées. Les principes sous-jacents sont souvent semblables, mais le point de vue différent adopté lors de la conception de chaque méthode aboutit à des hypothèses simplificatrices différentes et donc des implémentations différentes. Ces techniques peuvent être classées en 4 catégories (Fig. 1.9) :

Paramétrisation robuste du signal : extraire du signal des paramètres représentatifs de la parole possédant une sensibilité au bruit réduite.

Débruitage du signal : éliminer ou réduire l'influence du bruit sur le signal à reconnaître

Adaptation des modèles acoustiques : adapter les modèles acoustiques de manière à minimiser l'influence du bruit.

Modification de l'algorithme de décodage : modifier l'algorithme de décodage pour prendre en compte les différences entre le conditions d'apprentissage (parole seule) et de test (parole + bruit).

Ces différentes familles de stratégies sont décrites dans les paragraphes suivants. Pour chacune d'entre elles nous proposons une liste non exhaustive de méthodes.

1.2.2.1 Paramétrisation robuste du signal

Une première approche en reconnaissance robuste de la parole consiste à extraire du signal uniquement des paramètres pertinents pour le décodage phonétique et à réduire au maximum l'influence des autres sources. Parmi les paramétrisations robustes nous avons déjà cité (paragraphe 1.1.2) les méthodes MFCC et PLP.
Les campagnes d'évaluation Aurora [Pearce 00] de la reconnaissance de la parole robuste ont permis la conception d'un algorithme de paramétrisation standardisé par l'organisme ETSI [ETSI ES 202 050, 03] . Cet algorithme est connu sous le nom de *paramétrisation WI008*, ou encore *ETSI AFE* (ETSI Advanced Front End). Les résultats obtenus par cette méthode de paramétrisation robuste comptent parmi les meilleurs actuellement.

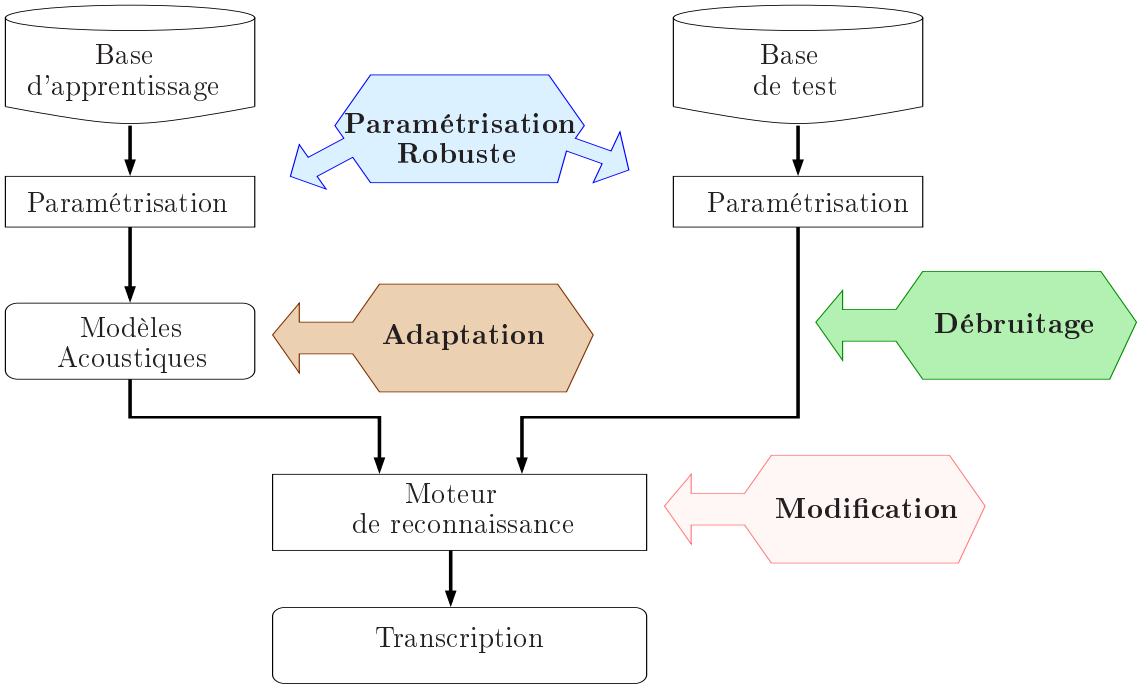


FIG. 1.9 – *Stratégies pour la reconnaissance robuste de la parole.*

1.2.2.2 Débruitage du signal

Le principe des méthodes de débruitage est de réduire les différences entre les conditions d'apprentissage et de test. Ces méthodes s'appuient sur des techniques de filtrage plus ou moins complexes qui tentent de supprimer ou de réduire l'influence du bruit sur le signal de la parole.

Soustraction spectrale

La combinaison des signaux de parole et de bruit est linéaire dans le domaine temporel :

$$y(t) \;=\; x(t) \;+\; n(t)$$

où $y(t)$, $x(t)$ et $n(t)$ représentent respectivement l'amplitude de la parole bruitée, de la parole seule et du bruit seul à l'instant t. Cette relation d'additivité est toujours valide dans le domaine spectral et préservée lors du passage dans le spectre de puissance à un terme de déphasage $cos(\Phi)$ près.

$$|Y(\tau)| \;=\; |X(\tau)| \;+\; |N(\tau)| \;.\; cos(\Phi)$$

Cependant il est montré, par diverses considérations, que ce terme peut être supposé proche de 1 et donc :

$$|Y(\tau)| \;=\; |X(\tau)| \;+\; |N(\tau)|$$

La soustraction spectrale propose de calculer une estimée du bruit sur des portions du signal ne contenant pas de parole. Sous l'hypothèse que le bruit soit stationnaire, l'estimée du bruit est soustraite du spectre de puissance du signal bruité.

Filtre de Wiener

Un filtre de Wiener est un estimateur $\hat{X}(t,f)$ du signal de parole seule, optimal au sens des moindres carrés, c'est à dire qui minimise :

$$E\left[\left(\hat{x}(t) \; - \; x(t)\right)^2\right]$$

Le filtre de Wiener fonctionne sous l'hypothèse que les trames de parole et de bruit sont issues de processus gaussiens et stationnaires de densités spectrales de puissances $\sigma_X^2(f)$ et $\sigma_N^2(f)$. Cet estimateur est exprimé par :

$$\hat{X}(t,f) \; = \; \underbrace{\frac{\sigma_Y^2(t,f) \; - \; \sigma_N^2(t,f)}{\sigma_Y^2(f)}}_{\alpha} \; . \; Y(t,f)$$

Le filtre de Wiener est défini par le terme multiplicatif α. Comme pour la soustraction spectrale, la principale difficulté réside dans l'estimation de $\sigma_N^2(t,f)$ qui s'apparente au calcul du SNR et qui est donc difficile à estimer lorsque le bruit est non stationnaire. Benoroya [Benaroya 03] proposa une adaptation permettant de prendre en compte au moins partiellement la non-stationnarité du bruit.

Débruitage paramétrique

Le débruitage paramétrique permet de transformer les trames de parole bruitée en trames de parole seule. Il s'agit de transformations non homogènes dans l'espace des paramètres acoustiques. Chacune des classes acoustiques qui peuvent être construites dans l'espace acoustique non bruité est mise en correspondance avec une classe acoustique dans l'espace acoustique bruité. Les transformations peuvent être apprises pendant la phase de construction du système en utilisant une base de données *stéréo*, c'est à dire possédant le même signal bruité et non bruité.

1.2.2.3 Adaptation des modèles acoustiques

Composition de modèles

Le principe de la composition de modèles est de combiner différents modèles (modèles acoustiques et modèles de bruit) pour ne former qu'un seul même modèle. La technique la plus utilisée est la combinaison parallèle de modèles (PMC : Parallel Model Combination) [Varga 90, Gales 93a]. Cette technique revient à construire un HMM équivalent aux deux modèles initiaux, supposant l'additivité des différentes sources sonores dans le spectre de puissance. Une telle combinaison est illustrée par la figure 1.10.
Il existe plusieurs problèmes inhérents à la combinaison parallèle de modèles :

- Un modèle de bruit doit être connu. De plus il est nécessaire d'estimer le SNR en condition de test afin de pouvoir combiner de manière *adéquate* les deux modèles.

- Il n'est pas possible de retirer du bruit, mais seulement d'en ajouter. Ceci implique que toutes les trames du signal de test soient plus bruitées que celles utilisées pour entraîner les modèles combinés.
- Les modèles combinés ont une complexité supérieure aux modèles initiaux.

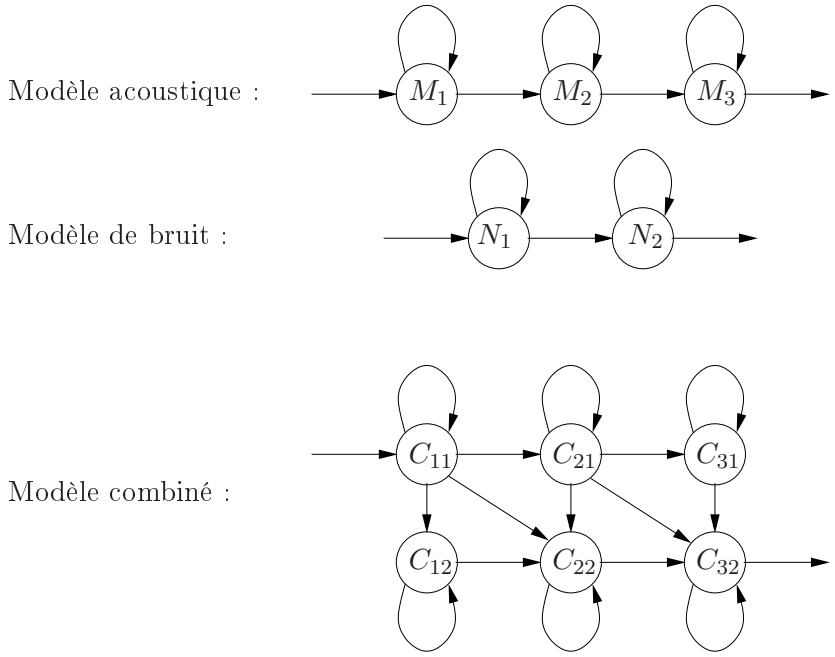


FIG. 1.10 – *Combinaison parallèle de deux modèles de Markov.*

Adaptation statistique

Le principe de l'adaptation statistique est d'adapter les paramètres des modèles acoustiques à la voix et/ou à l'environnement correspondant aux conditions de test. Cette adaptation nécessite un corpus appelé *corpus d'adaptation*. Ce dernier doit être constitué de phrases prononcées par le locuteur de test dans le cadre de l'adaptation au locuteur ou par des phrases enregistrées dans les mêmes conditions acoustiques que durant la phase de test pour l'adaptation à l'environnement. On distingue deux catégories d'adaptation statistique pour lesquelles différentes variantes sont proposées.

L'adaptation MAP (*maximum a posteriori*) [Lee 91, Gauvain 94] adapte les paramètres des modèles acoustiques (en général les moyennes) en fonction des données du corpus d'adaptation. L'équation régissant l'adaptation de la moyenne d'une gaussienne μ en $\hat{\mu}$ est la suivante :

$$\hat{\mu} = \frac{\tau\mu + \sum_{t=1}^{T} \gamma_t o_t}{\tau + \sum_{t=1}^{T} \gamma_t} \tag{1.10}$$

γ_t est la probabilité *a posteriori* que la gaussienne considérée soit alignée avec o_t et τ l'hyper-

paramètre qui règle l'équilibre entre la distribution *a priori* (initiale) et les nouvelles données. Si aucun exemple n'existe dans le corpus d'adaptation la moyenne reste inchangée. Inversement plus le nombre d'exemples dans le corpus d'adaptation est grand plus les nouvelles observations vont contribuer à adapter le modèle. Notons que seules les distributions observées sont adaptées et que la convergence des nouveaux paramètres est lente.
Les variantes SMAP (Structural Maximum A Posteriori) [Shinoda 97, Shinoda 01] et EMAP (Extended Maximum A Posteriori) [Lazri 84, Cox 95] sont des raffinements de cette adaptation. L'adaptation SMAP augmente la vitesse de convergence en regroupant les gaussiennes hiérarchiquement dans un arbre de régression. L'adaptation EMAP s'appuie sur des corrélations existantes entre plusieurs unités de parole pour adapter les modèles qui ne sont pas observés dans le corpus d'adaptation.

L'adaptation MLLR (*Maximum Likelihood Linear Regression*), contrairement à l'adaptation MAP, utilise un regroupement des gaussiennes "à plat". Il s'agit d'adaptation MLLR structurelle [Leggetteur 95] ou encore MAPLR structurelle [Siohan 00]. L'avantage de cette approche est de pouvoir adapter dynamiquement le nombre de classes de régression en fonction de la quantité de données présentes dans la base d'adaptation. La profondeur maximale des arbres de régression pouvant être atteinte est déterminée en fixant un nombre minimum d'exemples d'adaptation pour chaque classe. Chaque nouvelle observation est alors comptabilisée dans toutes les classes formant un chemin entre la racine de l'arbre et la feuille représentant la classe avec laquelle l'observation est alignée.

1.2.2.4 Modification de l'algorithme de décodage

La dernière technique de reconnaissance robuste de la parole que nous présentons concerne des approches basées sur la modification de l'algorithme de reconnaissance. Le principe est de considérer que dans un milieu bruité toutes les observations ne sont pas pertinentes. L'identification de ces observations et leur non prise en compte par le moteur de reconnaissance s'avère bénéfique.

Décodage incertain

Dans le cadre du décodage incertain, les observations sont considérées comme incertaines [Arrowood 03, Deng 05]. Sous cette hypothèse, le module de paramétrisation ne calcule plus un ensemble de paramètres mais considère ceux-ci comme des variables aléatoires et les substitue par leur densité de probabilité.
Lors du décodage, les variables aléatoires sont marginalisées, sur l'ensemble de leurs réalisations possibles. Pour chaque valeur, le résultat est pondéré par la probabilité de la variable aléatoire en ce point. En général, les densités de probabilité sont approchées par des lois normales dont les moyennes et variances sont calculées en fonction du SNR dans chaque bande de fréquences.

Reconnaissance partielle

La corruption d'un signal de la parole produit deux effets néfastes pour la reconnaissance. Le premier est que le bruit modifie les paramètres acoustiques des trames où la parole est présente. Ces paramètres ne sont donc plus pertinents lors de la comparaison avec les paramètres des modèles acoustiques. Le deuxième effet est que les silences généralement présents en début et fin de phrase ou encore entre deux mots ne sont plus détectables. En effet les silences correspondent à des intervalles de temps pendant lesquels l'énergie du signal est très faible. Le bruit pendant ces temps de silence peut alors être interprété, par le moteur de reconnaissance, comme la réalisation de phonèmes.
La reconnaissance partielle consiste à éliminer du signal à reconnaître toutes les trames exemptes de parole ou celles trop corrompues par le bruit pour fournir de l'information pertinente. La reconnaissance partielle, connue aussi sous le nom de *usable speech* [Yantorno 03, Chandra 02], est surtout utilisée en reconnaissance du locuteur [Khanwalkar 05, Ofoegbu 05]. Pour cette tâche il est possible de retirer un grand nombre de trames du signal d'origine. Cependant pour une application à la reconnaissance de la parole le nombre de trames éliminées du signal ne peut être aussi important.

Reconnaissance avec données manquantes

Le principe de la reconnaissance avec données manquantes est similaire à celui de la reconnaissance partielle mais de granularité plus fine. Il s'appuie sur l'hypothèse selon laquelle certains coefficients spectraux sont dominées par le bruit et d'autres par la parole. Alors que la reconnaissance partielle se base sur la détection de portions trop corrompues par le bruit au niveau de la trame, la reconnaissance de la parole avec données manquantes agit au niveau des coefficients spectraux. Il est montré qu'il vaut mieux ne pas utiliser les paramètres masqués par le bruit lors du décodage [Renevey 01a]. De plus, écarter certains paramètres ne pénalise pas ou peu la reconnaissance car le signal de la parole est très redondant dans le domaine spectral. Ainsi, il est démontré qu'il est possible de masquer jusqu'à 80 % des coefficients spectraux, dans le cadre d'un signal exempt de bruit, sans pour autant que cela affecte la reconnaissance.
L'identification des coefficients masqués est plus connue sous la dénomination d'estimation de masque. Un masque est une représentation temps-fréquence, comme le spectrogramme, où chaque coefficient $m(t, f)$ représente la confiance que l'on peut porter au coefficient spectral bruité $y(t, f)$. La reconnaissance de la parole avec données manquantes fait l'objet d'une étude plus approfondie exposée dans le chapitre suivant puisqu'elle est à la base de nos travaux.

1.3 Conclusion

L'élaboration de systèmes robustes de reconnaissance de la parole constitue aujourd'hui un des principaux enjeux du traitement automatique des langues. De nombreux systèmes ont été conçus s'appuyant sur des concepts relativement proches. Cependant, malgré les progrès réalisés,

les performances des systèmes actuels restent bien inférieures aux capacités du système auditif humain. A cet égard, les perspectives d'amélioration sont fortes. Les enjeux économiques liés à la reconnaissance de la parole font de ce domaine un secteur en constante évolution et pousse chacun de nous à imaginer sans cesse de nouvelles approches dans le but de pouvoir dépasser ou au moins égaler les performances de notre appareil auditif.
Des progrès importants sont nécessaires dans les différents niveaux de traitement du signal de la parole que nous venons de présenter. Alors que la compensation de bruits stationnaires est aujourd'hui relativement bien maîtrisée, l'un des principaux enjeux à l'heure actuelle concerne la robustesse aux bruits non stationnaires. Dans cette perspective, la reconnaissance de la parole avec données manquantes constitue une alternative très intéressante et prometteuse.

Chapitre 2

Reconnaissance automatique de la parole avec données manquantes

« On ne connaît pas complètement une science tant qu'on n'en connaît pas l'histoire. »

- Auguste Comte -

Sommaire

La reconnaissance avec données manquantes est une approche issue du domaine de la vision [Ahmad 93]. Elle fut transposée à la reconnaissance de la parole à la fin des années 90 par Martin Cooke et ses collègues de l'université de Sheffield [Cooke 96, Cooke 97, Cooke 01b]. Le principe général de cette approche repose sur un partitionnement des paramètres acoustiques en deux classes ; les paramètres fiables, peu ou pas affectés par le bruit fournissant une description assez fidèle du signal de la parole, et les paramètres manquants, trop affectés par le bruit décrivant de manière erronée le signal de parole à reconnaître. Ce partitionnement est plus connu sous la dénomination de masque de données manquantes.
Nous présentons dans ce chapitre les fondements sur lesquels s'appuie la théorie des données manquantes appliquée à la reconnaissance de la parole. Nous relatons en premier lieu des travaux montrant que notre système auditif traite tout stimuli de manière sélective. Nous sommes capable de reconnaître des mots prononcés à partir d'une représentation parcellaire du signal exploitant la redondance d'information du signal de parole : un processus de masquage des données manquantes est mis en œuvre. Le formalisme théorique de cette approche de reconnaissance sera ensuite détaillé mettant en avant l'importance des masques. Les différents processus de reconnaissance sur des observations partielles seront ensuite présentés avant de conclure par une évaluation du potentiel de cette approche.

2.1 Masquage en reconnaissance de la parole

La théorie des données manquantes en reconnaissance automatique de la parole est fortement inspirée du fonctionnement du système auditif humain. En effet, il a été établi que l'oreille humaine a un comportement sélectif vis-à-vis des différents stimuli qu'elle traite. Nous décrivons dans ce paragraphe le phénomène de masquage intervenant au niveau du système d'audition de l'Homme.

2.1.1 Théorie gestaltiste : perception et structuration du monde

Le mot allemand *Gestalt* est traduit par « forme » (ainsi *Gestalttheorie* signifie « théorie de la forme »), mais il s'agit en réalité de quelque chose de beaucoup plus complexe, qu'aucun mot ne traduit exactement dans aucune autre langue. Le verbe *gestalten* signifie « mettre en forme, donner une structure signifiante ». Le résultat, le « gestalt », est donc une forme structurée, complète et prenant sens pour nous. Par exemple, lorsqu'on regarde les étoiles, chacune d'elles est un stimulus visuel, pourtant on peut facilement les organiser en constellations, en ensembles structurellement cohérents formés de stimuli. Ainsi l'image mentale que nous avons en tête est une forme, et peut être évaluée par notre esprit en tant que telle, par exemple en la nommant « la Grande Ours ».
Pour comprendre un comportement ou une situation, il importe donc, non seulement de les analyser, mais surtout, d'en avoir une vue synthétique, de les percevoir dans l'ensemble plus vaste du contexte global, d'avoir un regard non plus pointu mais plus large : le contexte est

souvent plus signifiant que le texte (« com-prendre » c'est prendre ensemble). La structuration des formes ne se fait pas au hasard, mais selon certaines lois dites « naturelles » et qui s'imposent au sujet lorsqu'il perçoit le monde. Les principales lois de la Gestalt sont :

La loi de la bonne forme : loi principale dont les autres découlent : un ensemble de parties informes (comme des groupements aléatoires de points) tend à être perçu d'abord automatiquement comme une forme.

La loi de continuité : des points rapprochés tendent à représenter des formes lorsqu'ils sont perçus, nous les percevons d'abord dans une continuité, comme des prolongements les uns par rapport aux autres.

La loi de similitude : si la distance ne permet pas de regrouper les points, nous nous attacherons ensuite à repérer les plus similaires d'entre eux pour percevoir une forme.

La loi du destin commun : des parties en mouvement ayant la même trajectoire sont perçues comme faisant partie de la même forme.

La loi de clôture : une forme fermée est plus facilement identifiée comme une figure qu'une forme ouverte.

Ces lois déterminent la manière dont les éléments d'une scène perceptive sont organisés entre eux. Elles ont été proposées initialement pour l'explication des processus de perception visuelle [Köhler 29] mais Bregman [Bregman 90] montra qu'elles peuvent également se retrouver dans l'organisation des scènes auditives. Une étude plus détaillée de l'analyse de scène auditive est présentée dans la suite du manuscrit (chapitre 3 paragraphe 3.2).

2.1.2 Le masquage dans la perception humaine de la parole

La reconnaissance automatique de la parole avec données manquantes est un concept qui fut initié par des études menées dans le domaine de la perception humaine de la parole. Les années 50 ont vu émerger des techniques de codage de la parole pour les besoins du secteur des télécommunication.Ces techniques constituent aujourd'hui les bases de la théorie de la communication [Sha 49].

Ces travaux ont montré l'étonnante capacité de l'oreille humaine à reconnaître et comprendre un signal de parole même en présence de bruit à des niveaux élevés. Cette faculté est due au fait que le signal de la parole est extrêmement redondant dans le domaine spectral. Notre système auditif périphérique traite tout stimulus sonore par une analyse fréquentielle et peut donc exploiter pleinement cette redondance d'informations. Il est montré que nous pouvons reconnaître et comprendre la parole prononcée seulement à partir de petites portions du spectre sous la condition que ces portions ne soient pas ou très peu affectées par le bruit. Ces expériences ne montrent pas comment notre système auditif masque le signal mais elles montrent que reconnaître la parole sur des observations partielles est possible. Plus tard, des schémas de masquage ont été identifiés [Moore 82], dont les suivants :

Le masquage central

Ce type de masquage provient du fait que notre système auditif et composé de deux oreilles placées de part et d'autre de la tête et vers des directions opposées. Par conséquent chaque oreille ne perçoit pas la même information. Cette différence de perception est due aux différences de localisation et d'intensité des différentes sources sonores vis-à-vis de nos deux oreilles. Un processus de masquage est donc mis œuvre, basé sur ces différences de perception du monde sonore. Ce processus est à le base des principes d'identification et de localisation de sources sonores exploitant un banc de microphones.

L'effet de capture

Ce type de masquage ce produit lorsqu'un son domine localement la réponse neuronale. Plus précisément, le seuil d'audibilité d'un son est accru en présence d'un autre son de fréquence adjacente lorsque ces deux sons sont perçus simultanément ou avec un léger décalage temporel [Fletcher 37]. Si la différence des fréquences des sons devient supérieure à un seuil, ce phénomène disparaît. Ces travaux sont à l'origine de la notion de bandes critiques ayant conduit à l'utilisation des échelles perceptives Bark et Mel (chapitre 1 paragraphe 1.1.2.2).

Le masquage temporel

Le masquage temporel intervient dans le système d'audition lorsque qu'un son est précédé d'un autre son généralement d'intensité supérieur [Harris 79]. Ce phénomène est probablement lié à la capacité d'adaptation du système auditif à un stimulus de durée croissante.

2.2 Masquage en reconnaissance automatique de la parole

La reconnaissance avec données manquantes [Green 95, Lippmann 97, Morris 98] ainsi que les modèles multibandes [Boulard 96, Hermansky 96, Morgan 98, Cerisara 99] ont été développés dans le but de transposer ce phénomène de masquage à la reconnaissance automatique de la parole. L'approche multibandes, comme la reconnaissance avec données manquantes, ne transmet au moteur de reconnaissance qu'une partie des observations. Elle repose sur une sélection dynamique des bandes de fréquences qui serviront à la reconnaissance. Seules les bandes de fréquences les moins pénalisées par le bruit seront transmises. Bien que ces approches (modèles multibandes et reconnaissance avec données manquantes) soient quelque peu différentes elles sont toutes deux fondées sur l'identification des observations exploitables par le système RAP, c'est à dire, les observations les moins bruitées. Ce processus de classification des observations est appelé estimation de masque de données manquantes.

Nous définissons dans les paragraphes suivants ce qu'est un masque de données manquantes dans le cadre applicatif de la reconnaissance automatique de la parole.

2.2.1 Masque de données manquantes

2.2.1.1 Définition d'un masque

Un masque de données manquantes est une représentation d'un signal qui associe à chaque paramètre (exprimé dans un espace Λ) décrivant ce signal une mesure (exprimée dans un domaine D) de la fiabilité de l'information qu'il véhicule . Un masque est donc un étiquetage des observations. Dans le contexte de la RAP avec données manquantes cet étiquetage constitue une source d'informations additionnelles permettant au moteur de reconnaissance d'adapter son fonctionnement en fonction de la fiabilité des observations. De manière générale, un masque peut être défini comme une une fonction $\mathcal{M}$, qui associe à chaque paramètre du signal bruité exprimé dans un espace Λ, une valeur de masque à valeur dans D, en fonction d'un ensemble d'indices acoustiques exprimés dans un espace Ω :

$$\mathcal{M} \ : \ (\Lambda, \Omega) \ \rightarrow \ D \tag{2.1}$$

La fonction de masque $\mathcal{M}$ peut donc être vue comme l'application d'un critère de décision sur l'ensemble des indices acoustiques associés à chaque paramètre du signal Afin de mieux comprendre cette définition de masque, il convient d'en préciser les différents composants.

2.2.1.2 Critères de fiabilité et espace des masques

Considérons D qui est le domaine dans lequel un masque est exprimé, et qui contient donc l'ensemble des valeurs pouvant être attribuées aux masques. La nature des valeurs composant ce domaine peut être très diverse. Cependant ces valeurs doivent caractériser des propriétés ou indices pouvant être exploités par le moteur de reconnaissance. Deux formulations des masques ont été proposées :

Les masques discrets : $D \ = \ \{0, 1\}$: Les masques sont exprimés de manière discrète et binaire. Chaque paramètre acoustique étiqueté par une valeur valant 0 ou 1 signifiant respectivement que ce paramètre est fiable ou manquant.

Les masques continus : $D \ = \ [0, 1]$: Les valeurs de ce type de masque est un réel compris entre zéro et un. Ce réel traduit la probabilité qu'un paramètre soit manquant, c'est à dire trop corrompu par le bruit pour être signifiant.

Un paramètre est considéré comme masqué si la contribution du signal de parole du locuteur d'intérêt est plus faible que la contribution du bruit. Un tel critère peut être interprété, dans le spectrogramme, comme l'application d'un seuil sur la valeur du SNR_{local}. Cette formulation de

$\mathcal{M}$ fondée sur le seuillage du SNR_{local} est la plus intuitive et la plus largement adoptée par la communauté scientifique du traitement de la parole. Plusieurs études [Barker 00, Morris 01b] ont montré que les masques continus fournissent de meilleurs résultats que les masques discrets si l'on considère le masquage comme le seuillage du SNR. Ceci s'explique par le fait qu'une observation n'est jamais complètement manquante ou fiable. De plus, les erreurs de masques sont en quelque sorte lissées si ce dernier est exprimé de manière continue.

Bien que cette définition soit très largement répandue, elle n'est pas optimale. Cette définition est, certes, appropriée si l'objectif à atteindre est de débruiter le signal puisque le débruitage se traduit par une maximisation du SNR. Par conséquent un masque basé sur le SNR optimise directement l'objectif de débruitage. Cependant, la plupart des applications ont pour objectif d'améliorer la reconnaissance. Dans ce cas une amélioration du SNR ne se traduit pas systématique en une amélioration de l'intelligibilité. Dans le contexte de la reconnaissance automatique de la parole, il semblerait préférable de fonder le critère de fiabilité sur le taux de reconnaissance plutôt que sur le SNR, c'est à dire déterminer un critère tel que le masque résultant fournirait un taux de reconnaissance optimal, ce qui n'est pas le cas des masques basés sur le SNR. Cependant, il n'existe, à l'exception de Safayani qui propose dans [Safayani 07] un algorithme de soustraction spectrale optimisant la reconnaissance plutôt que le SNR, quasiment aucun travail publié exploitant un tel critère. Nous y voyons plusieurs raisons :

- L'unicité d'un masque optimisant le taux de reconnaissance n'est pas garantie. Au contraire il est fortement probable que plusieurs masques différents soient optimaux.
- Les masques basés sur le SNR sont bien plus aisés à calculer que des masques fondés sur la maximisation du taux de reconnaissance.
- Les masques basés sur le SNR dans le cadre de la RAP avec données manquantes fournissent de très bons taux de reconnaissance. Ceci suggère que ces masques sont en fait une bonne approximation des masques optimisant le taux de reconnaissance.

2.2.1.3 Espace des observations

En théorie un masque peut être défini pour tout domaine de paramétrisation, en pratique, le domaine des paramètres acoustiques doit être consistant avec le critère de fiabilité. Le signal est donc paramétré dans un domaine composé de différentes dimensions où chacune d'entre elles est associée à des bandes de fréquences distinctes. Une telle paramétrisation correspond à l'analyse fréquentielle du signal opérée par notre système auditif d'une part, et est consistante avec la notion de fiabilité des paramètres acoustique fondée sur un seuillage du SNR. Nous considérerons donc uniquement dans la suite des paramétrisations basées sur une représentation fréquentielle du signal et en particulier la paramétrisation Mel spectrale.

2.2.1.4 Espace des indices acoustiques

Un des fondement de la théorie gestaltiste est qu'une observation ne doit pas être considérée de manière isolée, mais doit être évaluée dans sont contexte. Il est donc nécessaire de recourir à divers indices caractérisant une observation dans son contexte afin d'évaluer sa fiabilité. Outre la caractérisation du contexte, ces indices doivent permettre de discriminer les observations fiables des observations manquantes en accord avec le ou les critères de fiabilité retenus. Ces indices ne sont soumis à aucune contrainte, et par conséquent peuvent être de natures très variées. Nous évoquons au chapitre suivant plusieurs indices acoustiques proposés pour l'estimation de masques. Cependant nous citons ici, en guise d'exemple, les indices proposés par M.L. Seltzer [Seltzer 00, Seltzer 04].

L'harmonicité

En raison de la nature harmonique de la parole, la majorité de l'énergie d'un signal de parole réside aux niveaux de ses harmoniques. Le bruit additif (à l'exception de la musique ou de parole concurrente) ne partage généralement pas cette propriété. Par conséquent, lorsque qu'un bruit additif corrompt un signal de parole l'harmonicité du signal résultant est plus faible que celle de la parole seule. Plus la contribution du bruit est forte est plus l'harmonicité décroît. Les indices entrant dans cette catégorie sont le *comb filter ratio* et l'*autocorrelation peak ratio*. Le premier calcule le rapport entre l'énergie du signal présente aux harmoniques et l'énergie localisée aux fréquences inter-harmoniques. Le second est basé sur la fonction d'autocorrélation utilisée pour calculer la fréquence fondamentale d'un signal de parole. La fonction d'autocorrélation exhibe un pique important au niveau de la fréquence fondamentale de la parole. Un second pique apparaît ensuite mais de façon moins prononcée. Moins le signal étudié est périodique et moins ce second pique est prononcé. Par conséquent le rapport de la hauteur de ces piques est un bon estimateur de l'harmonicité.

Les propriétés de l'enveloppe spectrale de parole

Outre ses caractéristiques harmoniques, la parole possède une enveloppe spectrale distincte des autres sons. La majeure partie de l'énergie de la parole est concentrée entre 300 Hz et 4000 Hz. Cette enveloppe spectrale est modifiée en fonction de l'enveloppe spectrale du bruit additif. Le rapport entre les énergies de chaque coefficient spectraux et l'énergie totale d'une trame est alors utilisée. Seltzer propose également d'exploiter le rapport entre l'énergie comprise dans chaque bande de fréquence et une estimée de l'énergie du bruit dans cette même bande ou une estimée de l'énergie du bruit sur la trame complète. De plus, la parole exhibe une trajectoire fréquentielle caractéristique faite de piques (aux fréquences harmoniques) et de vallées (aux fréquences inter-harmoniques). La compression (logarithmique ou cubique) des énergies du signal usuellement appliquée rend les vallées beaucoup plus sensibles au bruit additif. Ces vallées sont de moins en

moins prononcées à mesure que l'intensité du bruit augmente. Une mesure de la profondeur de ces vallées, appelée *flatness*, est donc proposée.

La distribution des énergies

Beaucoup de signaux sonores de la vie courante, incluant la parole, sont vus comme des signaux super-gaussiens. Leur distribution possède un kurtosis plus élevé qu'un signal gaussien (pique plus prononcé et des queues plus aplaties). Lorsque deux signaux super-gaussiens sont combinés, le kurtosis du signal résultant chute. En supposant qu'un signal de parole et son homologue en condition bruitée ont des kurtosis différents, l'évaluation de cette différence constitue un indice de la présence du bruit ainsi que de son intensité.

2.2.2 Masque oracle

Si l'on résume les indices acoustiques à la seule valeur exacte du SNR, qui peut être calculée à partir de deux signaux stéréoscopiques d'une même phrase (un signal bruité et le second ne contenant que le signal correspondant à la parole), le masque produit est appelé masque *oracle*. Une deuxième possibilité consiste à calculer les masques oracles à partir des différents flux audio (un flux de parole et un flux de bruit par exemple) composant la scène auditive. Les masques oracles sont des masques exempts de toute erreur. Par conséquent ils sont optimaux en accord avec le critère de fiabilité fondé sur le seuillage du SNR. Par construction, les masques oracles sont discrets, car pour chaque coefficient spectral, la valeur du SNR est connue et donc la fiabilité des coefficients est déterminée avec certitude.
En situation réelle, une telle information est indisponible. En effet, seules les observations bruitées sont observables. Néanmoins, les masques oracles ont une grande utilité. Ils sont par exemple utilisés pour estimer le potentiel des systèmes de reconnaissance automatique de la parole avec données manquantes ou pour comparer les performances de différents systèmes. Ils font également office de masques de référence pour entraîner des estimateurs de masques ou pour éventuellement estimer la qualité des masques en terme d'erreurs de classification.

2.3 Reconnaissance de la parole sur des observations partielles

2.3.1 Le problème des données manquantes

Le problème général de la classification robuste est d'assigner à toute séquence de T observations bruitées $Y = (y_1, y_2, \ldots, y_T)$ la meilleure séquence d'états $Q = (q_1, q_2, \ldots, q_T)$ ayant généré Y. Dans le contexte de la reconnaissance avec données manquantes, certaines observations sont incertaines ou manquantes. Il est donc préférable de ne pas tenir compte de celles-ci pendant le décodage, ou d'adapter celui-ci afin de pouvoir traiter ces données. En de telles circonstances, le problème pour une classification probabiliste est que la vraisemblance $P(Y|Q)$ ne peut être

évaluée de manière classique :

$$P(Y|Q) = \pi_0 \prod_{t=1}^{T} a_{q_{t-1}q_t}.b_{q_t}(y_t) \tag{2.2}$$

avec :

$$b_{q_t}(y_t) = \sum_{j=1}^{N_\lambda} \lambda_j \, \mathcal{N}(Y; \mu_{q_t}, \Sigma_{q_t}) \tag{2.3}$$

$$\mathcal{N}(y_t; \mu_{q_t}, \Sigma_{q_t}) = \frac{1}{\sqrt{(2\pi)^M \ |\Sigma_{q_t}|}} \exp\Big(-\frac{1}{2}(y_t - \mu_{q_t})'\Sigma_j^{-1}(y_t - \mu_{q_t})\Big) \tag{2.4}$$

Les différentes notations utilisées par ces équations ont été définies au paragraphe décrivant la topologie des HMM page 13. Les termes génériques O et o représentant respectivement l'ensemble des observations ainsi qu'un vecteur d'observations au temps t ont été remplacés ici par Y et y_t car les observations correspondent à un signal de parole bruité.

Deux approches sont communément adoptées pour la classification en présence de données manquantes ou incertaines : *l'imputation* et *la marginalisation* de données.
L' *imputation* de données requiert un prétraitement à la reconnaissance durant lequel les données manquantes sont remplacées des estimations de leurs vraies valeurs (valeurs qui auraient dû être observées en absence de bruit). Ce prétraitement permet de fournir au module de décodage des vecteurs d'observations complets et donc d'utiliser l'équation 2.3.
La *marginalisation* de données ne nécessite pas de reconstruire les données incomplètes, mais repose sur une modification de l'algorithme de classification, plus précisément sur le calcul des vraisemblances des observations, de façon à ce que celui-ci puisse classifier les observations en présence de données manquantes ou incertaines.

Par la suite, nous considérons que nous disposons d'un masque de données manquantes pour tout ensemble d'observations Y. Par conséquent, les observations sont scindées en deux parties $Y = Y_p \cup Y_m$ contenant respectivement les données fiables et les données manquantes. Nous noterons $Y(t, f)$ le coefficient spectral observé au temps t dans la bande de fréquence f.

2.3.2 Imputation des données

L'objectif de l'imputation de données est d'estimer les valeurs $\hat{Y}_m$ des observations manquantes Y_m que l'on aurait observer en absence de bruit, afin de fournir au système de reconnaissance des vecteurs d'observations complets $\hat{X} = Y_p \cup \hat{Y}_m$ et lui permettre ainsi de procéder à la classification en utilisant la vraisemblance $P(\hat{X}|Q)$:

$$P(\hat{X}|Q) \ = \ P(\hat{Y}_m|Q) \ . \ P(Y_p|Q) \tag{2.5}$$

Une des contributions majeures dans le cadre de l'imputation de données est le travail de thèse réalisé par Bhiksha Raj Ramakrishnan [Raj 00]. Il propose plusieurs algorithmes d'estimation

des données manquantes ainsi que leur évaluation. Nous allons détailler dans les paragraphes suivants les principaux algorithmes d'imputation de données qu'il a proposés.

2.3.2.1 Modification du moteur de reconnaissance

L'approche connue sous le nom de *class-conditional imputation* reconstruit le signal de parole durant le processus de décodage. Cette approche tire son nom du fait que les estimées des données manquantes sont dépendantes des classes acoustiques considérées durant le décodage. Considérons une hypothèse de décodage selon laquelle le chemin correspondant dans le treillis des hypothèses se retrouve dans l'état q_t au temps t. Le vecteur d'observations au temps t est noté $Y(t) = Y_p(t) \cup Y_m(t)$. L'estimation $\widehat{Y_m(t)}$ des valeurs manquantes $Y_m(t)$ est donnée par :

$$\widehat{Y_m(t)} \;=\; \arg\max_{Y_m(t)} \; P\left(Y_m(t) \middle| Y_p(t), q_t \right) \tag{2.6}$$

$\widehat{Y_m(t)}$ prend donc les valeurs pour lesquelles la probabilité *a posteriori* de l'observation reconstruite est la plus forte pour les paramètres du modèle de l'état q_t. Les valeurs imputées maximisent donc indépendamment la vraisemblance de tous les chemins du treillis.

2.3.2.2 Reconstruction géométrique

La reconstruction géométrique du spectrogramme est l'approche la plus simple. Cette technique estime la valeur des données manquantes uniquement à partir des données fiables présentes dans le spectrogramme. Aucune autre source de connaissance n'est utilisée.

L'idée directrice de la reconstruction géométrique est que le spectrogramme présente une évolution temporelle et fréquentielle continue. Il semble alors possible de prédire les valeurs des données manquantes à partir des valeurs fiables en utilisant des algorithmes d'interpolation. Deux types d'interpolation sont considérés ici : l'interpolation fréquentielle I_f et l'interpolation temporelle I_t. Nous ne décrirons ici que les interpolations temporelles, les interpolations fréquentielles pouvant être formulées par analogie en substituant la dimension temporelle par la dimension fréquentielle.

Considérons par le suite, $Y^f(t) = (\ Y(1,f), Y(2,f), \ldots, Y(T,f)\ )$ représentant les valeurs des coefficients spectraux du signal observés dans la bande de fréquence f à chaque instant $t \in [1,T]$. Si l'on note $Y_p^f(t)$ et $Y_m^f(t)$ l'ensemble des données fiables et manquantes : $Y^f(t) = Y_p^f(t) \cup Y_m^f(t)$. Nous dénommons $Y^f(t)$ par le terme d' *enveloppe temporelle* du signal pour la bande de fréquence f.

Plaçons nous dans le cas ou l'on désire estimer les valeurs de données manquantes $Y_m^f(t_i)$ telles que $0 < t_1 \leq t_i \leq t_2 < T$. Supposons également que $Card(Y_p^f(t)) = N$ et donc $Card(Y_m^f(t)) = T - N$. Interpoler un ensemble de N points $Y_p^f(t_i)$ consiste à rechercher une fonction $\mathcal{F}(t)$ telle que $\mathcal{F}(t)$ passe au mieux par ces N points. Les données manquantes aux temps t_i sont alors remplacées par

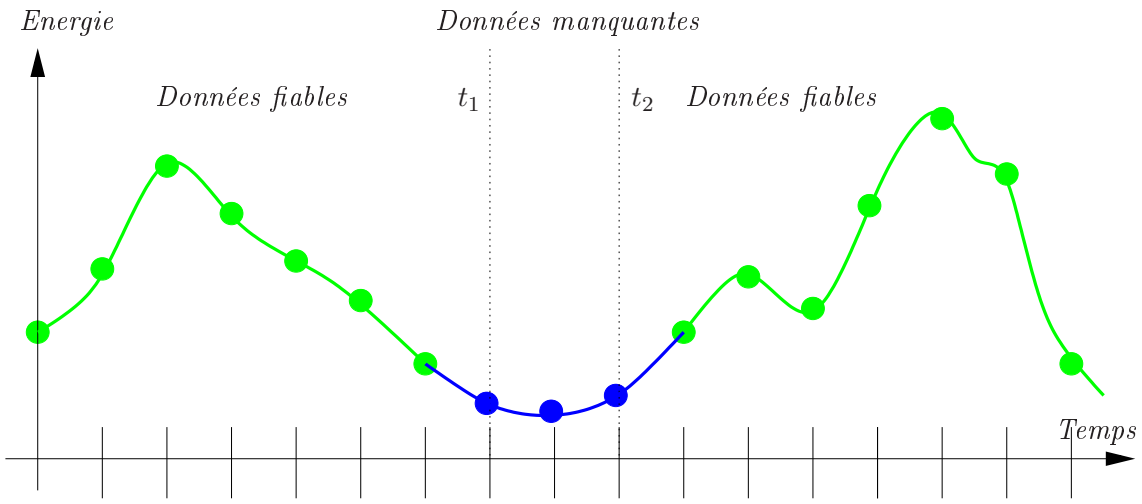


FIG. 2.1 – *Reconstruction d'un signal incomplet par interpolation temporelle. Les données fiables sont représentées par les courbes et points verts, les données reconstruites par des courbes et point bleus.*

leur estimée $\mathcal{F}(t_i)$ (FIG. 2.1). Plusieurs schémas d'interpolation ont été proposés : l'interpolation polynomiale avec le cas particulier de l'interpolation linéaire et l'interpolation rationnelle où la fonction d'approximation est définie comme le rapport de deux polynômes.

Interpolation polynomiale

L'interpolation polynomiale consiste à calculer les coefficients de l'unique polynôme $\mathcal{P}_{N-1}(t)$ de degré $N-1$ passant par les N points $(t, Y_p^f(t))$:

$$\mathcal{P}_{N-1}(t)\Big|_{I_t} = a_0 + a_1.t + a_2.t^2 + \ldots + a_{N-1}.t^{N-1} \tag{2.7}$$

L'algorithme utilisé est celui de Neuville qui propose une approche récursive du calcul des coefficients du polynôme de Lagrange :

$$\mathcal{P}_{N-1}(t)\Big|_{I_t} = \sum_{t_1 \leq t_i \leq t_2} \left\{ Y^f(t_i) \cdot \prod_{\substack{t_1 \leq t_j \leq t_2 \\ t_j \neq t_i}} \frac{(t - t_j)}{(t_i - t_j)} \right\} \tag{2.8}$$

Notons que l'interpolation linéaire est un cas particulier de l'interpolation polynomiale. Le polynôme est alors un polynôme de degré 1.

$$\begin{aligned}\mathcal{P}_1(t)\Big|_{I_t} &= Y^f(t_1-1) \ + \ \underbrace{\frac{Y^f(t_2+1)-Y^f(t_1-1)}{t_2-t_1+2}}_{\alpha} \ . \ (t-(t_1-1)) \\ &= \alpha.t \ + \ \underbrace{Y^f(t_1-1) \ - \ \alpha.(t_1-1)}_{\beta} \\ &= \alpha.t \ + \ \beta \end{aligned} \tag{2.9}$$

Interpolation rationnelle

L'interpolation rationnelle consiste à calculer les $L+M+1=N$ paramètres a_i et b_i de la fonction rationnelle unique $\mathcal{R}_M^L(t)$:

$$\begin{aligned}\mathcal{R}_M^L(t)\Big|_{I_t} &= \frac{\mathcal{N}_L(t)}{\mathcal{D}_M(t)} \\ &= \frac{1 \ + \ a_1.t \ + \ a_2.t^2 \ + \ \ldots \ + \ a_L.t^L}{b_0 \ + \ b_1.t \ + \ b_2.t^2 \ + \ \ldots \ + \ b_M.t^M}\end{aligned} \tag{2.10}$$

passant par les N points $(t, Y_p^f(t))$. Ici, $\mathcal{R}_M^L(t)$ peut être calculée efficacement par l'algorithme de Bulirsh-Stoer [Teukolsky 92]. L'algorithme de Bulirsh-Stoer est une procédure récursive qui construit les fonctions rationnelles d'ordre K à partir de fonctions rationnelles d'ordre $K-1$, sous la contrainte que l'ordre du polynôme dénominateur $\mathcal{D}_M(t)$ soit supérieur ou égale à l'ordre du polynôme numérateur $\mathcal{N}_L(t)$, c'est-à-dire $L \leq M \leq L+1$. Une fois les paramètres de la fonction rationnelle d'interpolation calculés, les valeurs des données manquantes $Y_m^f(t_i)$ aux temps t_i sont imputés par par les valeurs $\mathcal{R}_M^L(t_i)$. Il n'est cependant pas possible de recourir à l'interpolation pour les données se trouvant aux frontières du spectrogramme. Dans ce cas l'interpolation est remplacée par une extrapolation linéaire.

Raj Ramakrishnan a évalué ces algorithmes de reconstruction géométrique. Les résultats qu'il a reportés montrent que l'interpolation linéaire temporelle fournit les meilleures reconstructions.
Il apparaît que l'enveloppe fréquentielle ou temporelle ne peut être approximée par une fonction usuelle. Par conséquent, la reconstruction du spectrogramme obtenue par interpolation basée sur une fonction évoluée semble plus erronée qu'une reconstruction basée sur une simple interpolation linéaire.
De plus, l'interpolation temporelle permet une approximation par une fonction prototype de degré supérieur à ce qu'il est possible d'obtenir par interpolation fréquentielle car le nombre N de données fiables $Y_p^f(t)$ est plus important pour une enveloppe temporelle que pour une enveloppe fréquentielle.
En outre, il est plus fréquent de trouver des données manquantes aux frontières de l'enveloppe fréquentielle que de l'enveloppe temporelle. Par conséquent, une proportion plus importante de données sont estimées par extrapolation linéaire qui fournit une moins bonne estimation que

l'interpolation.

2.3.2.3 Reconstruction probabiliste

La reconstruction probabiliste repose sur un partitionnement $\mathcal{P}(X)$ de l'espace acoustique X du signal de parole seule, en K classes $\mathcal{C}_k$:

$$\mathcal{P}(X) \;=\; \sum_{k=0}^{K} c_k.\mathcal{G}_k(X;\Phi_k) \tag{2.11}$$

où $\mathcal{G}_k(X;\Phi_k)$ et c_k sont respectivement une représentation paramétrique de la distribution des vecteurs acoustiques du signal de parole seule et la probabilité *a priori* de la classe $\mathcal{C}_k$. Ce partitionnement $\mathcal{P}(X)$ est réalisé sur une base d'apprentissage non bruité. Généralement, $\mathcal{G}_k(X;\Phi_k)$ est modélisée par une loi normale :

$$\mathcal{G}_k(X;\Phi_k) \;=\; \mathcal{N}_k(\mu_k,\sigma_k) \tag{2.12}$$

Il est nécessaire dans un premier temps de déterminer la classe $\mathcal{C}_k$ d'appartenance du vecteur d'observations bruités $Y(t) \;=\; Y_p(t) \cup Y_m(t)$. Une fois la classes $\mathcal{C}_k$ identifiée, les données manquantes $Y_m(t)$ de ce vecteur sont réévaluées à partir des paramètres μ_k et σ_k de cette classe au maximum *a posteriori*. Le choix d'une métrique permettant d'identifier la classe d'appartenance des vecteurs acoustiques incomplets est donc prépondérant. Une première métrique reposant sur la distance marginale (*Cluster marginal reconstruction*) entre le vecteur d'observations $Y(t) = Y_p(t) \cup Y_m(t)$ et les classes $\mathcal{C}_k$ est proposée,

$$\begin{aligned} \mathcal{C}_k^{\star}(t) &= \arg\max_k \mathcal{N}_k(Y(t);\mu_k,\sigma_k) \\ &= \arg\max_k \left\{ c_k \;.\; \int_{-\infty}^{+\infty} P\Big(Y_p(t), Y_m(t) \Big| \mu_k,\sigma_k \Big)\; \mathrm{d}Y_m(t) \right\} \end{aligned} \tag{2.13}$$

Les résultats reportés par Raj Ramakrishnan sont similaires quelque soit le nombre de classes considérées. De meilleurs résultats sont obtenus par une pré-reconstruction du spectrogramme par interpolation temporelle (*Cluster time-interpolated reconstruction*) : $\hat{X}(t) = \widehat{Y_m(t)} \cup Y_p(t)$. Les vecteurs acoustiques $\hat{X}$ résultants de cette pré-reconstruction sont alors utilisés pour l'identification de la classe.

$$\mathcal{C}_k^{\star}(t) \;=\; \arg\max_k \mathcal{N}_k(\hat{X}(t);\mu_k,\sigma_k)$$

De manière générale les résultats obtenus par la reconstruction probabiliste sont meilleurs qu'en reconstruction géométrique.

2.3.2.4 Reconstruction statistique

Cette approche consiste à prendre en compte les corrélations entre les différents coefficients spectraux pour reconstruire le signal. Considérons un ensemble de vecteurs d'observations $Y =$

$Y_m \cup Y_p$. Notons μ_m et μ_p les vecteurs moyens des données manquantes et fiables, Σ_{pp} la matrice de covariance des données fiables et Σ_{mp} la matrice de covariance croisée des données manquantes et fiables. L'estimation des données manquantes d'un vecteur d'observations $Y(t)$ est alors :

$$\widehat{Y_m(t)} \;=\; \mu_m \;+\; \Sigma_{mp} \,.\, \Sigma_{pp}^{-1} \,(Y_p(t) \;-\; \mu_p) \tag{2.14}$$

Il est plus facile pour des raisons de complexité de reconstruire individuellement (*Covariance individual reconstruction*) les données manquantes présentes dans un vecteur d'observations $Y(t)$. Cependant de meilleurs résultats sont reportés lorsque que cette estimation est faite conjointement (*Covariance joint reconstruction*) pour l'ensemble des données manquantes de ce vecteur.

2.3.2.5 Reconstruction en présence de données incertaines

Nous avons présenté jusqu'ici des algorithmes de reconstruction de spectrogrammes incomplets. Aucune hypothèse *a priori* sur les valeurs pouvant être imputées aux données manquantes n'est utilisée. Cependant, nous avons vu que les énergies des différentes sources sonores composant un signal peuvent être considérées comme additives dans le spectre de puissance. Il est alors possible d'exploiter cette propriété pour définir une plage de valeurs possibles pour chaque donnée manquante. En particulier, toute contribution de l'énergie de la parole $X(t,f)$ pour une observation bruitée $Y(t,f)$ suit la relation :

$$0 \leq X(t,f) \leq Y(t,f) \tag{2.15}$$

Cette plage constitue l'intervalle d'incertitude sur les valeurs qui auraient dû être observées en l'absence de bruit. On parle alors de données *incertaines*. Des algorithmes de reconstruction qualifiée de *bornés* ont donc été proposés pour traiter des données incertaines et non plus manquantes : *bounded class-conditional imputation*, *bounded cluster marginal reconstruction* et *bounded covariance-based reconstruction*. Ces algorithmes sont des raffinements de ceux précédemment exposés, imposant que les valeurs estimées des données incertaines respectent l'inégalité 2.15. Ces schémas de reconstruction ne sont pas détaillés ici. Nous invitons les lecteurs à consulter [Raj 00] pour de plus amples informations.

2.3.3 Marginalisation des données

Le décodage de la parole par des modèles HMM revient à déterminer la meilleure séquence d'états $Q^\star = (q_1, q_2, \ldots, q_T)$ pouvant engendrer la séquence d'observations $Y = (y(1), y(2), \ldots, y(T))$:

$$\begin{aligned} Q^\star &= \arg\max_Q P(Q|Y) \\ &= \arg\max_Q \frac{P(Y|Q)\;P(Q)}{P(Y)} \end{aligned}$$

$P(Y)$ étant constant pour tout Q :

$$\begin{aligned} Q^\star &= \arg\max_Q \; P(Y|Q)\; P(Q) && (2.16)\\ &= \arg\max_Q \; \left(\prod_{t=1}^{T} P(y_t|q_t)\right) . \left(\pi_0 . \prod_{t=1}^{T} P(q_t|q_{t-1})\right) \\ &= \arg\max_Q \; \pi_0 \prod_{t=1}^{T} a_{q_{t-1}q_t} . b_{q_t}(y_t) && (2.17) \end{aligned}$$

Nous avons vu qu'il n'est pas efficace, en présence de données manquantes ou incertaines, d'évaluer la probabilité $P(Y|Q)$ directement par l'équation 2.17 et plus particulièrement le terme $b_{q_t}(y_t)$ qui est la vraisemblance de l'observation y_t au regard des paramètres de l'état q_t. Sous l'hypothèse d'additivité des énergies dans le spectre de puissance $Y = X + N$ avec X et N étant respectivement les énergies de la parole seule et du bruit, le décodage optimal est :

$$\begin{aligned} Q^\star &= \arg\max_Q \; P(X|Q)\; P(Q) \\ &= \arg\max_Q \; \left(\prod_{t=1}^{T} P(x_t|q_t)\right) . \left(\pi_0 . \prod_{t=1}^{T} P(q_t|q_{t-1})\right) \\ &= \arg\max_Q \; \pi_0 \prod_{t=1}^{T} a_{q_{t-1}q_t} . b_{q_t}(x_t) && (2.18) \end{aligned}$$

Cependant, en condition de test l'énergie du signal de parole X n'est pas observable, seules les énergies du signal bruité Y le sont. Nous allons voir dans le paragraphe suivant comment estimer $P(X|Q)\; P(Q)$ à partir des seules observations bruitées Y.

2.3.3.1 Décodage en présence de données manquantes ou incertaines

Le décodage de Viterbi fournit comme résultat de classification une séquence d'états. Cette séquence est celle dont la probabilité *a posteriori* conditionnée par les observations Y et les paramètres des HMM Θ est la plus forte.

$$Q^\star = \arg\max_Q \; P(Q|Y,\Theta)$$

Considérant que certaines observations peuvent être manquantes ou incertaines cette équation peut être reformulée comme suit :

$$\begin{aligned} Q^\star &\propto \arg\max_Q \; E[\; P(Q|X,\Theta) \mid X \sim s(X)\;] \\ &= \arg\max_Q \; \int_\Lambda P(Q|X,\Theta) . s(X)\; \mathrm{d}X \end{aligned}$$

En appliquant la règle de Bayes, il vient :

$$Q^\star = \arg\max_Q \; P(Q|\Theta) . \int_\Lambda \frac{P(X|Q,\Theta)}{P(X|\Theta)} . s(X)\; \mathrm{d}X \qquad (2.19)$$

Nous venons d'introduire le nouveau terme $s(X) = P(X|\Phi)$, qui peut être défini comme la fonction de densité de probabilité de X définie sur Λ et conditionnée par un ensemble Φ de connaissances (critères, propriétés et/ou observations) permettant d'inférer $s(X)$. Cet ensemble de connaissances Φ peut être divisé en trois catégories :

1. Les observations non bruitées utilisées pour entraîner les modèles acoustiques, modélisées par la densité de probabilité $P(X|X_{Tr}) = P(X|\Theta)$, ou Θ est l'ensemble des paramètres des modèles acoustiques.
2. Les observations bruitées Y.
3. Un ensemble de connaissances additionnelles κ. Ces connaissances peuvent être très variées. On peut évoquer l'estimation du SNR, des contraintes sur le domaine de définition de $s(X)$, l'intervalle d'incertitude des observations, etc.

On peut alors exprimer $s(X)$ comme étant la loi de probabilité de X conditionnée par ces trois ensembles de connaissance :

$$s(X) = P(X|X_{Tr}, Y, \kappa) \tag{2.20}$$
$$s(X) = P(X|\Theta, Y, \kappa) \tag{2.21}$$

En appliquant la règle de Bayes et en postulant l'indépendance entre X_{Tr} et Y (c'est à dire que le bruit est décorrélé du signal de la parole) :

$$P(X|X_{Tr}, Y, \kappa) = \frac{P(X|\Theta).P(X|Y, \kappa)}{P(X)} \tag{2.22}$$

On peut remplacer le terme $\frac{1}{P(X)}$ par une constante α car cette densité de probabilité est plate.

$$s(X) = \alpha \,.\, P(X|Y, \kappa) \,.\, P(X|\Theta) \tag{2.23}$$

Par conséquent, en remplaçant $s(X)$ dans l'équation 2.19 et puisque $Q^\star$ ne dépend pas de α, il vient :

$$Q^\star = \arg\max_Q \underbrace{P(Q|\Theta)}_{(1)} \,.\, \underbrace{\int_\Lambda P(X|Q, \Theta) \,.\, P(X|Y, \kappa) \, \mathrm{d}X}_{(2)} \tag{2.24}$$

Nous retrouvons ici un premier terme (1) correspondant à la probabilité a priori de la séquence d'état Q. Par contre, le terme (2) correspondant initialement à la probabilité $P(Y|Q)$ que les observations aient été générées par Q est ici remplacé par l'espérance de cette probabilité sur Λ du fait que certaines observations peuvent être manquantes ou incertaines.

L'équation 2.25 est la reformulation de l'équation 2.24 dans le cadre d'un décodage à partir de l'algorithme de Viterbi décrit au paragraphe 1.1.4.2 (page 14).

$$\begin{aligned} Q^\star &= \arg\max_Q \left(\pi_0 \,.\, \prod_{t=1}^{T} P(q_t|q_{t-1})\right) \,.\, \left(\prod_{t=1}^{T} \int_\Lambda P(x_t|q_t) \,.\, P(x_t|Y,\kappa)\ \mathrm{d}x_t\right) \\ &= \arg\max_Q \left(\pi_0 \,.\, \prod_{t=1}^{T} a_{q_{t-1}q_t}\right) \,.\, \left(\prod_{t=1}^{T} \int_\Lambda b_{q_t}(x_t) \,.\, P(x_t|Y,\kappa)\ \mathrm{d}x_t\right) \\ &= \arg\max_Q \ \pi_0 \,.\, \prod_{t=1}^{T} \left[a_{q_{t-1}q_t} \,.\, \int_\Lambda b_{q_t}(x_t) \,.\, P(x_t|Y,\kappa)\ \mathrm{d}x_t\right] \end{aligned} \tag{2.25}$$

L'équation 2.25 implique d'utiliser des HMM possédant des matrices de covariance diagonales. En effet il est complexe et coûteux d'évaluer l'expression $\int_\Lambda b_{q_t}(x_t) \,.\, P(x_t|Y,\kappa)\ \mathrm{d}x_t$ au niveau de la trame en tenant compte des corrélations existantes entre les différents coefficients. L'utilisation de matrices de covariance diagonales résulte de l'hypothèse d'indépendance des coefficients spectraux d'une même trame. Bien que cette hypothèse soit fausse pour le signal de la parole, elle offre un cadre théorique simplifié permettant une mise en œuvre plus aisée de la marginalisation de données. Finalement, sous cette hypothèse d'indépendance des coefficients, l'équation générale de la reconnaissance de la parole avec données manquantes ou incertaines est :

$$Q^\star = \arg\max_Q \ \pi_0 \,.\, \prod_{t=1}^{T} \left[a_{q_{t-1}q_t} \,.\, \prod_{f=1}^{F} \underbrace{\int_\Lambda b_{q_t}(x(t,f)) \,.\, P(x(t,f)|Y,\kappa)\ \mathrm{d}x(t,f)}_{\text{Probabilité d'émission marginale}}\right] \tag{2.26}$$

où F est le nombre de bandes de fréquences du spectrogramme. Notons que que l'équation 2.26 ne fait pas de distinction explicite entre données manquantes et données fiables, les vraisemblances de tous les coefficients sont exprimée de la même manière. Cette distinction est pourtant présente implicitement dans le terme $P(x(t,f)|Y,\kappa)$. Cette loi de probabilité est propore à chacun des coefficients spectraux. Elle caractérise l'interprétation du masque faite par le système de reconnaissance.

Par exemple, considérons la reconnaissance de la parole en situation maîtrisée, c'est-à-dire qu'aucun bruit ne vient perturber le signal : $Y = X$. Au sens du SNR, aucun coefficient n'est masqué. La certitude sur la pertinence des observations est donc totale. En une telle situation, la meilleure solution pour modéliser $P(X|Y,\kappa)$ est alors d'utiliser une fonction de Dirac. Une fonction de Dirac $\delta(.)$ vaut 1 si son paramètre est nul et 0 sinon. Il est ainsi possible de reformuler l'équation 2.26

par :

$$\begin{aligned}
Q^\star &= \arg\max_Q \ \pi_0 \,.\, \prod_{t=1}^{T} \left[a_{q_{t-1}q_t} \,.\, \prod_{f=1}^{F} \int_\Lambda b_{q_t}(x(t,f)) \,.\, \delta(x(t,f) - y(t,f)) \, \mathrm{d}x(t,f) \right] \\
&= \arg\max_Q \ \pi_0 \,.\, \prod_{t=1}^{T} \left[a_{q_{t-1}q_t} \,.\, \prod_{f=1}^{F} b_{q_t}(y(t,f)) \right] \\
&= \arg\max_Q \ \pi_0 \,.\, \prod_{t=1}^{T} \left[a_{q_{t-1}q_t} \,.\, \prod_{f=1}^{F} b_{q_t}(x(t,f)) \right] \\
&= \arg\max_Q \ \pi_0 \,.\, \prod_{t=1}^{T} a_{q_{t-1}q_t} \,.\, b_{q_t}(x_t) \qquad (2.27)
\end{aligned}$$

L'équation 2.27 est exactement l'équation de reconnaissance de la parole sans bruit (équation 2.18). L'équation 2.26 est une généralisation de l'équation 2.18 classiquement utilisée en reconnaissance de la parole.

2.3.3.2 Schémas de marginalisation

Soit $m(t,f)$ une variable aléatoire prenant valeur dans $[0,1]$. Cette variable vaut 1 si le coefficient spectral $y(t,f)$ est manquant ou incertain et 0 sinon. Un masque de données manquantes M fournit alors la probabilité $P(m(t,f)=1)$ (ou $1-P(m(t,f)=0)$) pour chaque coefficient spectral $y(t,f)$. Dans un soucis de clarté, nous noterons abusivement $x(t,f)$, $y(t,f)$ et $m(t,f)$ par x, y et m.

Marginalisation non-bornée

La technique de marginalisation le plus simple se place dans le contexte de données manquantes par opposition aux données incertaines. En effet, dans ce contexte, les données fiables sont considérées comme étant suffisamment représentatives du signal de la parole. Le moteur de reconnaissance les traite donc de manière classique (équation 1.9). Ceci est une conséquence directe de l'hypothèse de dominance. L'hypothèse de dominance postule qu'un coefficient spectral est dominé soit par l'énergie de la parole, soit par l'énergie du bruit. Lorsque le spectre de puissance est compressé par une fonction, ce qui est souvent le cas avec $\log(.)$ ou $\sqrt[3]{.}$, l'énergie observée peut être imputée soit au signal de la parole, soit au signal du bruit :

$$y \approx \max(x,n)$$

Par contre les données masquées sont considérées comme étant totalement manquantes, c'est à dire qu'elles sont inutilisable par le moteur de reconnaissance. En outre, cette technique de marginalisation fait l'hypothèse qu'il n'existe pas de critère permettant d'estimer les valeurs de ces données qui auraient dues être observées en absence de bruit. Une telle marginalisation est

appelée *Full marginalisation* (Fmarg) [Cooke 96]. Elle est caractérisée par :

$$\begin{aligned}\int_{\Lambda} b_{q_t}(x) \;.\; P(x|Y,\kappa)\;\mathrm{d}x &= P(m=0)\;.\;b_{q_t}(x) \;+\; P(m=0)\;.\;\underbrace{\int_{-\infty}^{\infty} b_{q_t}(x)\;\mathrm{d}x}_{=1} \\ &= P(m=0)\;.\;b_{q_t}(x) \;+\; P(m=0) \end{aligned} \tag{2.28}$$

La full marginalisation consiste donc a reconnaître la parole à partir uniquement des données fiables du spectrogramme. Les données manquantes n'influent pas le résultat car la marginale de leur vraisemblance sur $]-\infty,\infty[$ est égale à 1 quelques soit l'état considéré.

Marginalisation bornée

La marginalisation bornée exploite le fait que les valeurs du spectrogramme sont des énergies, et par conséquent celles-ci sont positives. Si l'on note x l'énergie de la parole :

$$0 \;\leq\; x \tag{2.29}$$

De plus si l'on considère l'additivité des énergies de différents signaux (le signal de la parole compris) composant une scène auditive, une borne supérieure sur les valeurs des données manquantes peut être fixée. En effet, un coefficient bruité y peut être exprimé comme la somme des énergies de la parole x et du bruit n :

$$y \;=\; x \;+\; n$$

Les valeurs x et n étant toutes deux positives :

$$x \;\leq y \tag{2.30}$$

Les équations 2.29 et 2.30 définissent un intervalle de valeurs pouvant être considérées pour les données masquées. On parle alors ici de données incertaines.

Une première formulation de la marginalisation bornées est basée sur une modélisation de $P(x|Y,\kappa)$ par un mélange de deux fonctions.
La première d'entre elles est la fonction de Dirac. Une fonction de Dirac $\delta(.)$ vaut 1 si son paramètre est nul et 0 sinon.
La deuxième est une fonction de densité de probabilité uniforme $u(a,b)$. Soit une variable aléatoire z régie par une fonction de densité de probabilité $u(a,b)$, $P(z=c) = \frac{1}{b-a},\ \forall c \in [a,b]$ et $P(z=c) = 0,\ \forall c \notin [a,b]$.
Finalement, $P(x|Y,\kappa)$ est exprimée par :

$$P(x|Y,\kappa) \;=\; P(m=0)\;.\;\delta(x-y) \;+\; P(m=1)\;.\;u(0,y) \tag{2.31}$$

d'où :

$$\begin{aligned}
\int_\Lambda b_{q_t}(x) \,.\, P(x|Y,\kappa)\ \mathrm{d}x &= \int_\Lambda b_{q_t}(x) \,.\, \Big(P(m=0) \,.\, \delta(x-y) \ +\ P(m=1) \,.\, u(0,y)\Big)\ \mathrm{d}x \\
&= P(m=0) \,.\, \int_\Lambda b_{q_t}(x) \,.\, \delta(x-y)\ \mathrm{d}x \ + \\
&\quad P(m=1) \,.\, \int_\Lambda b_{q_t}(x) \,.\, u(0,y)\ \mathrm{d}x \\
&= P(m=0) \,.\, b_{q_t}(y) \ +\ \frac{P(m=1)}{y} \,.\, \int_0^y b_{q_t}(x)\ \mathrm{d}x \qquad (2.32)
\end{aligned}$$

Nous désignerons cette marginalisation par *Uniform-Dirac marginalisation* (UDmarg) [Vizinho 99, Barker 01b].

Considérons maintenant le critère de fiabilité utilisé pour partitionner le spectrogramme en données manquantes et données fiables. Ce critère est basé sur le seuillage du SNR_{local}. Tout coefficient spectral dont le SNR_{local} est inférieur à un seuil γ est considéré comme manquant.

$$\begin{aligned}
P(m=1) &= P(SNR_{local} \ <\ \gamma) \\
&= P(10\ \log_{10}\big(\frac{x}{n}\big) \ <\ \gamma) \\
&= P(x \ <\ x_{SNR-\gamma}) \qquad (2.33)
\end{aligned}$$

Nous définissons la valeur $x_{SNR-\gamma}$ comme l'énergie du signal de parole sous l'hypothèse que le SNR soit égal à γ dB. Inversement tout coefficient spectral dont le SNR_{local} est supérieur ou égal à γ est considéré comme fiable.

$$\begin{aligned}
P(m=0) &= P(SNR_{local} \ \geq\ \gamma) \\
&= P(x \ \geq\ x_{SNR-\gamma}) \qquad (2.34)
\end{aligned}$$

La combinaison des inéquations 2.29 et 2.33 ainsi que 2.30 et 2.34 donne :

$$P(m=1) = P(0 \ \leq\ x \ <\ x_{SNR-\gamma}) \qquad (2.35)$$

$$P(m=0) = P(x_{SNR-\gamma} \ \leq\ x \ \leq\ y) \qquad (2.36)$$

Ces deux équations, 2.35 et 2.36, définissent respectivement des intervalles de marginalisation pour les données manquantes et les données fiables. La probabilité marginale d'émission d'une

observation y pour un état q_t est alors :

$$\begin{aligned}
\int_\Lambda b_{q_t}(x).P(x|Y,\kappa)\ \mathrm{d}x &= \int_\Lambda b_{q_t}(x).\Big(P(m=0).u(0,x_{SNR-\gamma}) + P(m=1).u(x_{SNR-\gamma},y)\Big)\mathrm{d}x \\
&= P(m=0)\ .\ \int_\Lambda b_{q_t}(x)\ .\ u(0,u(x_{SNR-\gamma}))\ \mathrm{d}x\ + \\
&\quad P(m=1)\ .\ \int_\Lambda b_{q_t}(x)\ .\ u(u(x_{SNR-\gamma}),y)\ \mathrm{d}x \\
&= \frac{P(m=0)}{x_{SNR-\gamma}}\ .\ \int_0^{x_{SNR-\gamma}} b_{q_t}(x)\ \mathrm{d}x\ + \\
&\quad \frac{P(m=1)}{y\ -\ x_{SNR-\gamma}}\ .\ \int_{x_{SNR-\gamma}}^{y} b_{q_t}(x)\ \mathrm{d}x
\end{aligned} \tag{2.37}$$

Cette marginalisation est appelée *Uniform-Uniform marginalisation* (UUmarg) et fut proposée pour le première fois par Morris [Morris 01a].

Il est également possible de représenter $P(x|Y,\kappa)$ par n'importe quelle loi de probabilité. Cependant il est nécessaire de disposer de données par estimer cette loi. Ceci peut se faire sur une base de donnée ou encore sur les données fiables du signal à décoder si celles-ci sont suffisamment nombreuses. L'utilisation d'une loi normale est proposée (*Gaussian marginalisation* : Gmarg) également par Morris [Morris 01b].

2.3.3.3 Marginalisation des coefficients de vitesse et d'accélération

Nous avons vu au paragraphe 1.1.4.3 que la paramétrisation spectrale est souvent étendue par des coefficients dynamiques (dérivées premières Δ et secondes $\Delta\Delta$). Soit $\Delta y(t,f)$ et $\Delta\Delta y(t,f)$ les dérivées premières et secondes du signal au temps t dans la bande de fréquence f. Le calcul de ces valeurs est le suivant :

$$\Delta y(t,f) = \frac{\sum_{\theta=1}^{N} \theta.\big(y(t+\theta,f)\ -\ y(t-\theta,f)\big)}{2.\sum_{\theta=1}^{N}\theta^2} \tag{2.38}$$

$$\Delta\Delta y(t,f) = \frac{\sum_{\theta=1}^{N} \theta.\big(\Delta y(t+\theta,f)\ -\ \Delta y(t-\theta,f)\big)}{2.\sum_{\theta=1}^{N}\theta^2} \tag{2.39}$$

L'application d'un seuillage du SNR_{local} comme critère de fiabilité n'a évidemment pas de sens pour les coefficients dynamiques. Une solution simple fut proposée par Barker [Barker 00]. Il propose de considérer un coefficient dynamique comme manquant si au moins un des coefficients contribuant à son calcul est manquant. Un tel critère ne permet pas d'inférer un intervalle de marginalisation pour les données comme c'est la cas pour les coefficients statiques. Les coefficients dynamiques sont alors traités par la technique de *full marginalisation*.

2.3.3.4 Évaluation

Nous proposons dans ce paragraphe une évaluation de la marginalisation de données. Cette évaluation est effectuée sur la base de données standardisée Aurora 2 (paragraphe 5.2.1.2) dédiée à l'évaluation de systèmes de reconnaissance robuste de la parole. Cette base est constituée de phrases correspondant à des suites de chiffres prononcées en anglais. Nous avons utilisé un système de reconnaissance basé sur HTK (HMM ToolKit) développé à l'université de Cambridge. Les calculs des vraisemblances des observations correspondant aux techniques de marginalisation Fmarg, UDmarg et UUmarg ont été implémentées. La reconnaissance repose sur l'utilisation de masques oracles au sein desquels nous avons introduit aléatoirement une certaine proportion d'erreurs. Les vecteurs d'observations sont constitués de 32 coefficients Mel-spectraux compressés cubiquement et complétés par leur dérivée première.

La figure 2.2 donne pour chaque proportion d'erreurs sur les masques les taux de reconnaissance en mots obtenus avec chacune des 3 techniques de marginalisation mentionnées ci-dessus. Ces taux de reconnaissance représentent les taux moyens obtenus sur l'ensemble des conditions acoustiques proposés par la base de test A d'Aurora 2.

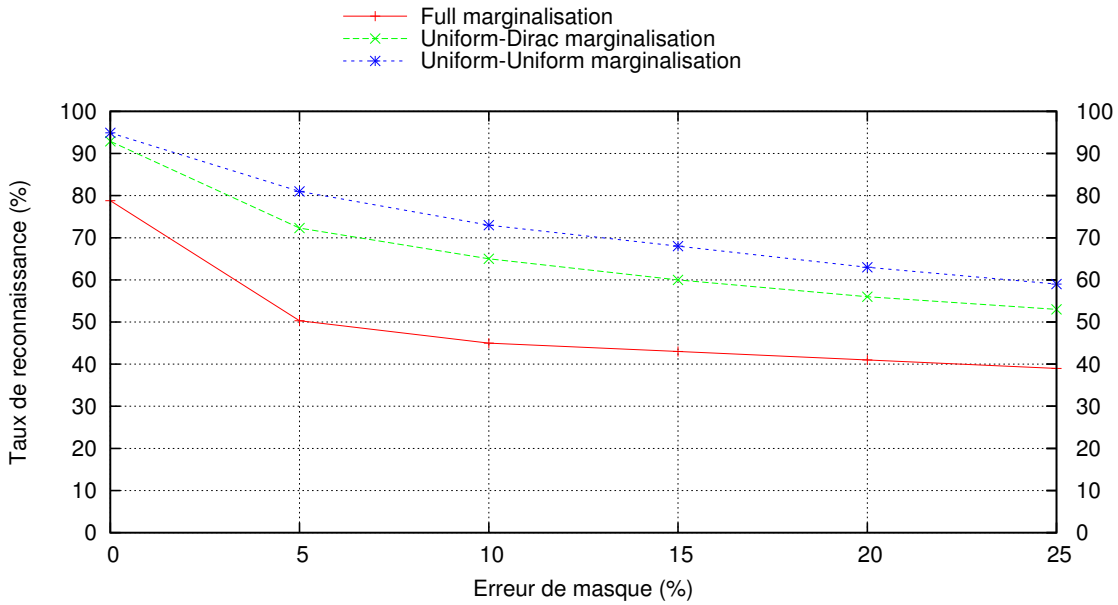


FIG. 2.2 – *Évaluation et comparaison de 3 techniques de marginalisation de données : Full marginalisation, Uniform-Dirac marginalisation et Uniform-Uniform marginalisation.*

Ces résultats illustrent le bénéfice apporté par l'utilisation de fonctions de densité de probabilité dérivées des propriétés du spectrogramme ainsi que du critère de fiabilité des données basé sur le seuillage du SNR_{local}. En effet de bien meilleurs résultats sont atteints avec les techniques de marginalisation UDmarg et UUmarg comparativement à Fmarg.

De plus, la technique UUmarg semble plus robuste aux erreurs de masques que UDmarg. Une

explication possible est la suivante. D'une part, tout comme Fmarg, UDmarg exploite l'hypothèse de dominance. La contribution du bruit dans les coefficients spectraux non masqués n'est peut-être pas aussi négligeable qu'il n'y parait pour la tâche de reconnaissance. D'autre part, UUmarg fournit des intervalles de marginalisation plus petits que ceux de UDmarg. Ceci se traduit par une meilleure évaluation des probabilités marginales, c'est à dire que celles-ci sont plus discrimantes lorsqu'elles sont calculées sur des intervalles plus restreints. Cette dernière remarque explique la différence de performance entre UDmarg et Fmarg qui diffèrent seulement par le fait que les données manquantes sont marginalisées sur un intervalle fini pour UDmarg et infini pour Fmarg.

Les taux de reconnaissance obtenus en absence d'erreur illustrent bien le fort potentiel de la reconnaissance automatique de la parole avec données manquantes. Les masques oracles fournissent de très bons taux de reconnaissance malgré la présence de conditions acoustiques fortement dégradées par le bruit. Bien entendu ces performances constituent une limite des performances atteignables et ne reflètent donc pas les performances d'un système de reconnaissance avec données manquantes utilisé en conditions réelles.

L'évolution des taux de reconnaissance en fonction de la proportion d'erreurs de masque souligne l'importance de la qualité des masques. En effet les performances se dégradent fortement dès la présence d'erreurs de masque même en de petites quantités Ce phénomène s'accentue en général avec l'utilisation de masque estimé. Les erreurs de masques sont introduites aléatoirement pour cette expérience et sont donc réparties assez uniformément sur l'ensemble du plan temps-fréquence. Les erreurs commises sur les masques estimés ont tendance à être localisées en des régions homogènes, ce qui est bien plus pénalisant pour la reconnaissance.

2.4 Conclusion

Nous avons présenté dans ce chapitre les principes de la reconnaissance automatique de la parole avec données manquantes. Nous avons montré, dans un premier temps, que le phénomène de masquage est présent dans notre système d'audition. Les différents travaux menés dans cette voie n'ont pu mettre en évidence la manière dont nous masquons certaines parties du signal, mais ils ont, en revanche, montré l'existence de ce principe. Dans un second temps, ce principe de masquage est transposé au problème de la reconnaissance automatique robuste de la parole. Nous nous sommes efforcés de définir de la façon la plus générale les masques de données manquantes. Ceux-ci peuvent être considérés comme l'application de critères sur les paramètres du signal permettant d'inférer leur fiabilité au regard du processus de reconnaissance.

Le critère le plus utilisé est basé sur le seuillage du SNR_{local} qui est une mesure de la quantité de bruit présent dans un signal. Bien que ce critère soit très intuitif, il constitue une des principales limitations de la reconnaissance de la parole avec données manquantes. En effet, celui-ci

contraint le système de reconnaissance à opérer dans le domaine spectral qui est un domaine de paramétrisation peu robuste au bruit.

Les différents algorithmes de reconnaissance en présence de données manquantes ou incertaines ont été présentés. Ils se décomposent en deux grandes familles : l'imputation et la marginalisation de données initiées respectivement par des travaux effectués à l'université de Carnegie-Mellon et l'université de Sheffield. Bien que les algorithmes au sein d'une même famille partagent la même philosophie, les approches du problème et leur mise en œuvre diffèrent. Nous avons fournit pour chacune d'elles les différentes variantes qui les composent. La figure 2.3 propose un arbre de classification des différentes approches de reconnaissance avec données manquantes mentionnées dans ce chapitre.

Ces deux familles peuvent s'apparenter à des approches robustes mentionnées au chapitre 1. En effet, l'imputation de données peut être vue comme une technique de débruitage. Cependant ces deux approches diffèrent par le fait que l'imputation de données débruite seulement les régions spectrales masquées. La marginalisation s'apparente de son côté au décodage incertain. Elle est caractérisée par le fait que les intervalles d'incertitudes sont le plus souvent modélisés par des distributions de probabilités uniformes, et que ces intervalles sont directement corrélés avec les différentes notions, critères et propriétés utilisés pour inférer le masque de données manquantes ou incertaines.

Nous avons proposé, dans le cadre de la marginalisation, une évaluation comparative de trois techniques de marginalisation. Les résultats obtenus soulignent le fort potentiel de la reconnaissance avec données manquantes mais aussi l'importance de la qualité des masques. En effet, la reconnaissance est très sensible aux erreurs commises sur les masques. L'estimation des masques constitue donc à l'heure actuelle un enjeu important et est à l'origine de nombreux travaux. Nous allons dans le chapitre suivant dresser l'état de l'art concernant l'estimation de masques de données manquantes.

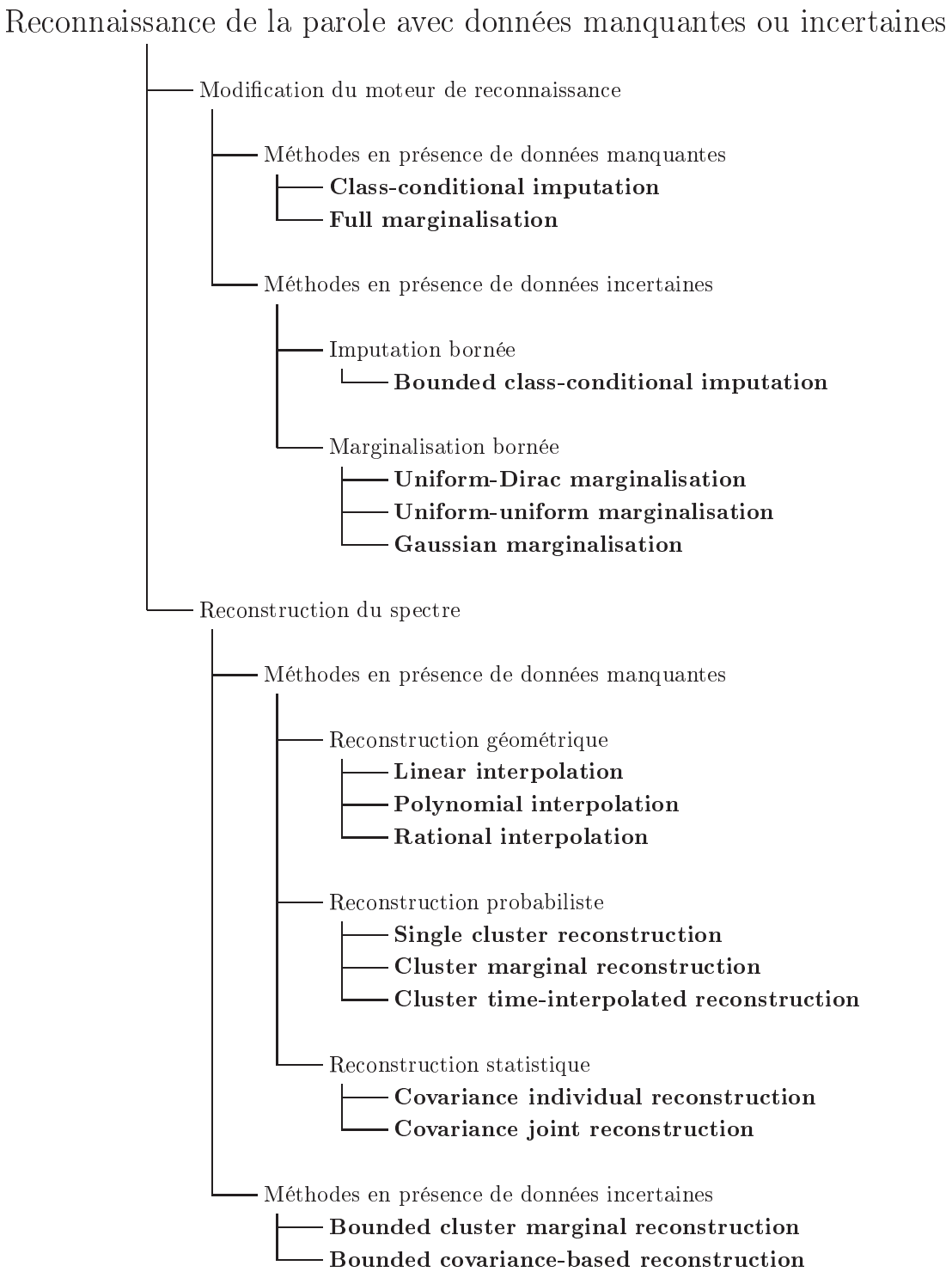


FIG. 2.3 – *Arbre de classification des algorithmes de reconnaissance automatique de la parole en présence de données manquantes ou incertaines (adapté de [Raj 00]).*

Chapitre 3

Estimation des masques : état de l'art

« Si nous avons chacun un objet et que nous les échangeons, nous avons chacun un objet. Si nous avons chacun une idée et que nous les échangeons, nous avons chacun deux idées. »

- proverbe chinois -

Sommaire

3.1 Introduction

Ce Aurora 2itre est dédié à l'estimation des masques de données manquantes. L'objectif est de recenser les principales techniques pouvant contribuer à identifier les données manquantes présentes dans un signal bruité. Ces techniques sont présentées en deux grandes classes : d'une part, les méthodes s'inspirant de l'oreille humaine et, d'autre part, les méthodes fondées sur le traitement du signal acoustique et sur des modèles stochastiques.

La première classe d'algorithmes recense les approches issues du domaine CASA (Analyse de Scènes Auditives Computationnelle). L'analyse de scène auditive computationelle se propose de modéliser l'aptitude de l'oreille humaine à structurer notre environnement sonore. En effet notre système auditif, contrairement aux systèmes de reconnaissance automatique de la parole, possède la capacité d'identifier tout ou partie des sources sonores composant notre environnement acoustique. De plus, nous pouvons aisément focaliser notre attention sur une source sonore particulière même si celle-ci est noyée dans un environnement acoustique sévère. Les travaux issus de CASA ont donc pour objectif d'étudier et de modéliser les différents mécanismes mis en œuvre par notre système auditif dans le but de les transposer à la RAP.

Les algorithmes constituant la deuxième classe abordent le problème d'estimation des masques exclusivement d'un point de vue du traitement du signal. Au lieu de modéliser la structure et le comportement de notre système auditif, ces approches considèrent le signal acoustique comme un ensemble d'observations dont certaines sont erronées. La détection de ces dernières est basée sur des considérations statistiques (application de filtres par exemple) ainsi que sur nombre d'outils mathématiques.

L'état de l'art proposé dans ce Aurora 2itre ne considère pas les méthodes exploitant plus d'un microphone pour l'acquisition du signal. Certes, les indices de localisation des différentes sources sonores pouvant être extraits dans une telle situation sont très importants et ont prouvé à plusieurs reprises leur utilité. Cependant l'objectif qui est le nôtre est de proposer des solutions afin d'améliorer la robustesse au bruit des systèmes RAP. La finalité de cet axe de recherche est de pouvoir, à terme, utiliser la RAP de façon quotidienne via des systèmes électroniques embarqués (téléphonie mobile, industrie automobile,...). Pour des raisons d'ergonomie, de mise en œuvre ou même économiques, ces systèmes sont munis dans la majorité des cas d'un seul microphone. Ce travail bibliographique est donc focalisé exclusivement sur des approches basées sur un seul et unique microphone.

3.2 Analyse computationnelle de scène auditive

3.2.1 Principes de base

L'analyse de scène auditive (ASA) étudie, d'une part, notre capacité à analyser notre environnement acoustique pour identifier chacun des objets ou entités élémentaires contribuant au signal perçu et, d'autre part, notre capacité à focaliser notre attention sur une ou plusieurs de ces sources sonores. L'analyse computationnelle de scène auditive (CASA) vise à développer et à mettre en œuvre des modèles et algorithmes ayant le même objectif, c'est à dire identifier chaque source sonore, extraire les flux sonores d'une ou plusieurs de ces sources et se focaliser sur ceux-ci dans le but de les décoder (détection d'évènements, identification du locuteur principal, etc). Dans le contexte de la reconnaissance automatique de la parole avec données manquantes, le rôle des approches CASA est limité à l'estimation des masques. Le but est d'identifier et isoler les régions spectrales dominées par le signal de la parole du locuteur principal. L'identification de formes masquantes permet de mieux comprendre la scène auditive et ainsi mettre en évidence des processus de regroupements perceptifs des évènements sonores (Fig. 3.1).

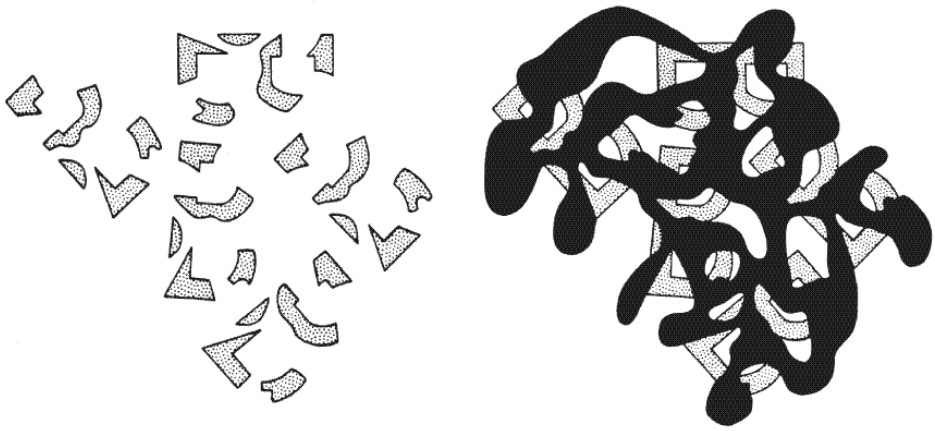


Fig. 3.1 – *Analyse de scène visuelle. A gauche, les fragments sont inorganisés. A droite, la présence d'une forme masquante permet leur regroupement perceptif. L'analyse de scènes auditives cherche des principes analogues pour l'organisation du monde sonore (d'après [Bregman 90]).*

Une première approche pour estimer les masques de données manquantes est d'imiter la capacité qu'a l'oreille humaine à se focaliser sur un locuteur en particulier, même avec des bruits très difficiles comme de la parole concurrente. Beaucoup d'études [Bronkhorst 00, de Cheveigne 00, Cooke 01a] ont essayé de comprendre les mécanismes permettant à l'oreille humaine d'atteindre un tel potentiel. Le livre de Bregman [Bregman 90] définit les principes fondamentaux sur lesquels la plupart des systèmes CASA s'appuient. Ces principes reprennent en partie les concepts gestaltistes afin de modéliser et structurer notre environnement sonore.

Bregman a montré qu'à partir de différents indices ou caractéristiques d'un signal, différents sons peuvent être perçus comme provenant d'une seule et même source sonore. Ceci montre qu'il existe au sein du cerveau humain des processus capables d'identifier différents évènements sonores et de les fusionner afin qu'il soient perçus comme un unique son ou comme plusieurs sons mais provenant d'une même source. Sa théorie propose deux classes d'organisation et de fusion des évènements acoustiques. La première est la fusion d'évènements simultanés (synchronicité d'attaque). Par exemple, plusieurs signaux sinusoïdaux partageant les mêmes caractéristiques d'harmonicité sont fusionnés pour être perçus comme un unique son. Par contre, si ceux-ci ne partagent plus les mêmes caractéristiques d'harmonicité ou s'ils sont décalés dans le temps, alors ces signaux sont perçus comme plusieurs sons. La seconde classe d'organisation est l'organisation séquentielle des évènements acoustiques. Des principes dérivés de la théorie gestaltiste postulent que des évènements acoustiques successifs sont assignés à des flux audio différents. Bregman suggère qu'une telle organisation des évènements acoustiques est « pré-cablée » dans notre appareil auditif. Il propose en outre un autre processus de groupement et de partitionnement des évènements acoustiques qui est basé sur des schémas appris au préalable. Contrairement aux processus d'organisation et de fusion fondés sur des caractéristiques organisationnelles et structurelles du signal, ces schémas sont spécifiques aux différents types de sources sonores, comme par exemple la parole. Ceci suggère que notre système auditif dispose de techniques de groupement d'évènements acoustiques spécifiques à des signaux issus de sources sonores connues. Notre système auditif serait alors capable d'apprendre des schémas d'organisation et de fusion spécifiques à différentes sources sonores.

3.2.2 Bref survol des systèmes CASA

Les systèmes CASA les plus significatifs sont présentés dans [Brown 05]. Les premiers systèmes CASA [Weintraub 85, Brown 94] étaient basés sur une approche ascendante (bottom-up), pour laquelle des indices acoustiques sont mis en évidence afin de grouper ou de séparer les contributions des différentes sources sonores composant une scène auditive. Par exemple le système proposé par Brown propose un regroupement d'évènements acoustiques ayant une réalisation quasi-simultanée, c'est à dire qui possèdent à peu près les mêmes dates de début (onset) et de fin (offset). Ce système exploite également des critères d'autocorrélation ainsi que des caractéristiques sur l'évolution du signal dans chaque bande de fréquence. Les composants partageant les mêmes caractéristiques sont alors considérés comme étant issus de la même source sonore. Un autre système CASA ascendant qui exploite la continuité du timbre de la voix (caractéristique fréquentielle) a été proposé dans [Masuda-Katsuse 99]. Masuda-Katsuse et ses collègues montrent que la continuité observée dans l'évolution temporelle du signal spectral est un indice fort qui est utilisé dans l'approche de regroupement séquentiel décrit par Bregman. Bien que ce critère fut nié par Bregman, les auteurs de [Masuda-Katsuse 99] ont mis en œuvre des expériences psychophysiques qui confirment son existence. Leur système exploite également l'harmonicité du signal pour grouper les fréquences de différents flux audio concurrents. Un regroupement séquen-

tiel des flux audio est ensuite effectué à l'aide d'un modèle auto-regressif du second ordre. La modulation d'amplitude peut également être employée pour estimer les masques de données manquantes dans les parties de hautes fréquences du spectre [Hu 04]. En effet, Les filtres utilisés pour les hautes fréquences couvrent une grande largeur de bande et les harmoniques qu'ils capturent créent des phénomènes de battement possédant une modulation d'amplitude propre.

La conception d'un système générique exploitant ces procédures de fusion et de partitionnement des différents composants d'une scène auditive n'est cependant pas proposée dans ces travaux. On peut toutefois évoquer les systèmes fondés sur le modèle du tableau noir (blackboard), les multi-agents ou encore les systèmes neuromimétiques. Ces trois architectures sont présentées dans les paragraphes suivants.

3.2.2.1 Architecture de tableau noir

Une architecture dite de tableau noir est issue du domaine de l'intelligence artificielle et a pour but de faire coopérer plusieurs sources de connaissances disponibles sous forme d'agents. Une telle architecture fut utilisée pour la première fois par R. Reddy et son équipe de Carnegie-Mellon University dans le système HEARSAY II [Reddy 80].

Godsmark et Brown [Godsmark 99] ont fondé leurs travaux sur le fait que l'organisation des évènements acoustiques au sein de l'oreille humaine est sensible au contexte acoustique et est rétroactive. Ceci suggère que la décision de grouper, ou non, certaines régions spectrales du signal peut être retardée avec l'espoir qu'un indice désambiguïsant émerge par la suite. Ils proposent ainsi de laisser interagir les principes de regroupement sur une fenêtre glissante de 300 millisecondes, puis de décider d'une organisation des évènements acoustiques au terme de cette fenêtre. De cette manière, chaque décision d'organisation de la scène auditive prise à chaque instant est fondée sur l'ensemble des indices collectés pendant 300 millisecondes précédant cet instant.

Ce système a été validé sur des signaux musicaux. Il est fondé sur une architecture de tableau noir, décomposée en huit étapes, s'enchaînant de manière ascendante. La décision est ainsi construite à partir d'indices à faible granularité ($1^{\text{ème}}$ étape) à partir desquels des hypothèses successives sont émises, qui permettent elles-mêmes d'inférer de nouvelles hypothèses. Finalement, la dernière étape fournit l'organisation de la scène acoustique en terme de phrases mélodiques. L'organisation de ce système est la suivante :

1. Le signal est paramétré de façon à révéler à chaque instant les fréquences dominantes (de plus forte énergie) pour former des traits mélodiques (« synchrony strands ») (Fig. 3.2). Le regroupement de différentes fréquences dominantes est fondé sur le principe de continuité temporelle, de leur proximité fréquentielle ainsi que sur la cohérence entre leurs amplitudes respectives. Ce dernier principe est employé pour empêcher plusieurs signaux, issus d'instruments différents et qui joueraient des sons harmoniques, d'être regroupés.

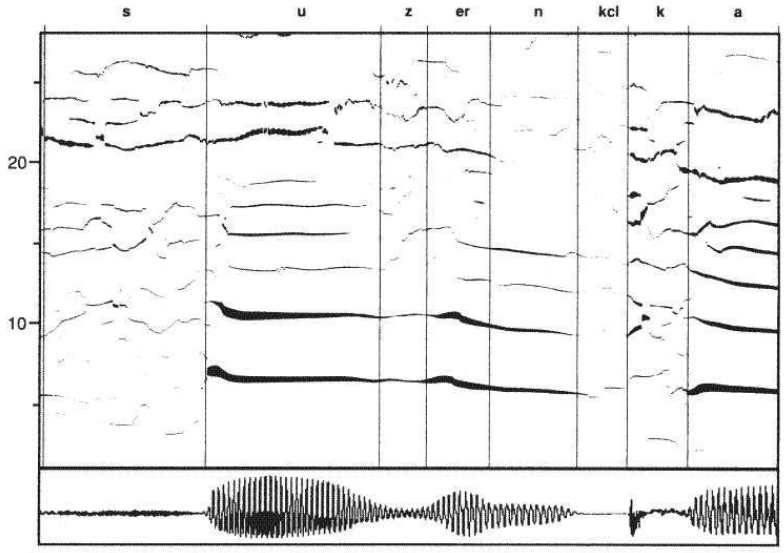


FIG. 3.2 – *Représentation de type « synchrony strands », de Cooke, en réponse à une portion de parole (en bas). Dans les basses fréquences chaque* strands *correspond à une harmonique, dans les hautes fréquences il correspond à un formant (d'après [Cooke 93]).*

2. Chaque trait mélodique peut être décrit par des dates de début et de fin ainsi que par sa trajectoire dans le plan temps-fréquence. Ces caractéristiques sont mentionnées sous le terme traits caractéristiques (« feature strands »).
3. Les traits mélodiques partageant des traits caractéristiques similaires sont regroupés. La structuration acoustique résultante de ce regroupement de traits mélodiques correspondent dans cette étape à l'identification de note.
4. Les différentes hypothèses formulées à l'étape précédente sont évaluées par des experts et triées par probabilité décroissante. Un score calculé à partir des différentes règles de regroupement (proximité fréquentielle, continuité temporelle, etc) est associé à chaque hypothèse. Les hypothèse ayant les scores les plus élevées sont les hypothèses les plus probables.
5. La fréquence fondamentale ainsi que le timbre sont extraits pour chaque note.
6. Un nouveau regroupement est ensuite effectué. Le regroupement des différentes notes successives s'appuie sur la proximité de leur fréquence fondamentale ainsi que sur la similitude de leur timbre. La continuité de la fréquence fondamentale reflète la trajectoire d'une mélodie dans le plan temps-fréquence. Le timbre est propre à chaque type d'instrument. C'est en quelque sorte sa signature acoustique. Ces nouveaux groupes correspondent ici à des lignes mélodiques.
7. Ces nouvelles hypothèses sont également évaluées et un score leur est attribué.
8. La dernière étape consiste à identifier les différentes mélodies à partir de caractéristiques spécifiques à chaque instrument. Ces experts affaiblissent ou renforcent les hypothèses

émises plus tôt.

Ellis propose un système CASA [Ellis 96] reposant sur une architecture de tableau noir prédictif, fondé sur des modèles internes du monde. Les différents agrégats possibles d'objets composant ces modèles expliquent et prévoient le signal observé. Les modèles du monde contiennent trois classes élémentaires de sons. Chaque modèle est en fait une composition structurelle de sons élémentaires dont les paramètres représentent la contribution et l'organisation de ces sons. A chaque instant t un modèle du monde expliquant la scène auditive est prédit de manière probabiliste à partir du modèle du monde déterminé à l'instant $t-1$ et de l'ensemble des observations (l'enveloppe énergétique par exemple) à l'instant t. Parallèlement, le meilleur modèle correspondant à ces observations est déterminé. Si le modèle prédit est suffisamment proche du modèle effectivement observé, celui-ci est mis à jour avec les nouveaux paramètres. Dans le cas contraire, le nouveau modèle est activé ou le modèle prédit est désactivé afin de prendre en compte l'incohérence de ceux-ci.

3.2.2.2 Architecture multi-agents

Nakatani [Nakatani 02] propose une architecture CASA très flexible fondée sur l'utilisation d'agents. Ce système utilise également des indices d'harmonicité. Un générateur d'agents détecte les débuts et fins de structures harmoniques et génère un agent "suiveur" pour chacune d'entre elles. La détection de structures harmoniques est effectuée par des agents « détecteurs » d'harmonicité. Chacun d'eux évalue l'harmonicité du signal pour une fréquence fondamentale fixée. L'agent « détecteur » renvoyant le score d'harmonicité le plus fort génère un agent « suiveur ». Cet agent « suiveur » est chargé de regrouper temporellement les structures harmoniques successives sur la base de la proximité de leur fréquence fondamentale. Les différentes structures harmoniques détectées sont ensuite soustraites du signal. Le signal résiduel peut ensuite être traité de nouveau. Un agent de contrôle est également mis en œuvre afin d'éliminer des agents « suiveurs » redondants.

Cet ensemble d'agents, dédié à une tâche précise comme l'extraction d'indices d'harmonicité, forme une *agence* (agency). D'autres agences ont été proposées, chacune d'elles exploitant respectivement des indices visuels, de localisation des sources sonores et de séparation binaurale ou monaurale. Chaque agence produit sa propre représentation de la scène auditive, et différents types d'interaction inter-agence sont mis en œuvre pour fusionner ces différentes représentations. Par exemple les représentations résultantes de deux agences fondées sur une séparation monaurale peuvent être fournies comme entrées à une agence procédant à une séparation de source binaurale. Pour illustrer ces interactions entre différentes agences, les auteurs ont enrichi leur système de façon à prendre en compte simultanément des indices d'harmonicité et localisation de sources.

La figure 3.3 décrit l'architecture d'un système CASA multi-agents Ipanema proposé par Kashino et Murase [Kashino 97]. Ce système fut évalué dans le cadre de l'identification d'instruments et

de leur mélodie à partir d'un signal de musique polyphonique. Les auteur reportent des résultats encourageants avec un taux d'identification de 66 %.

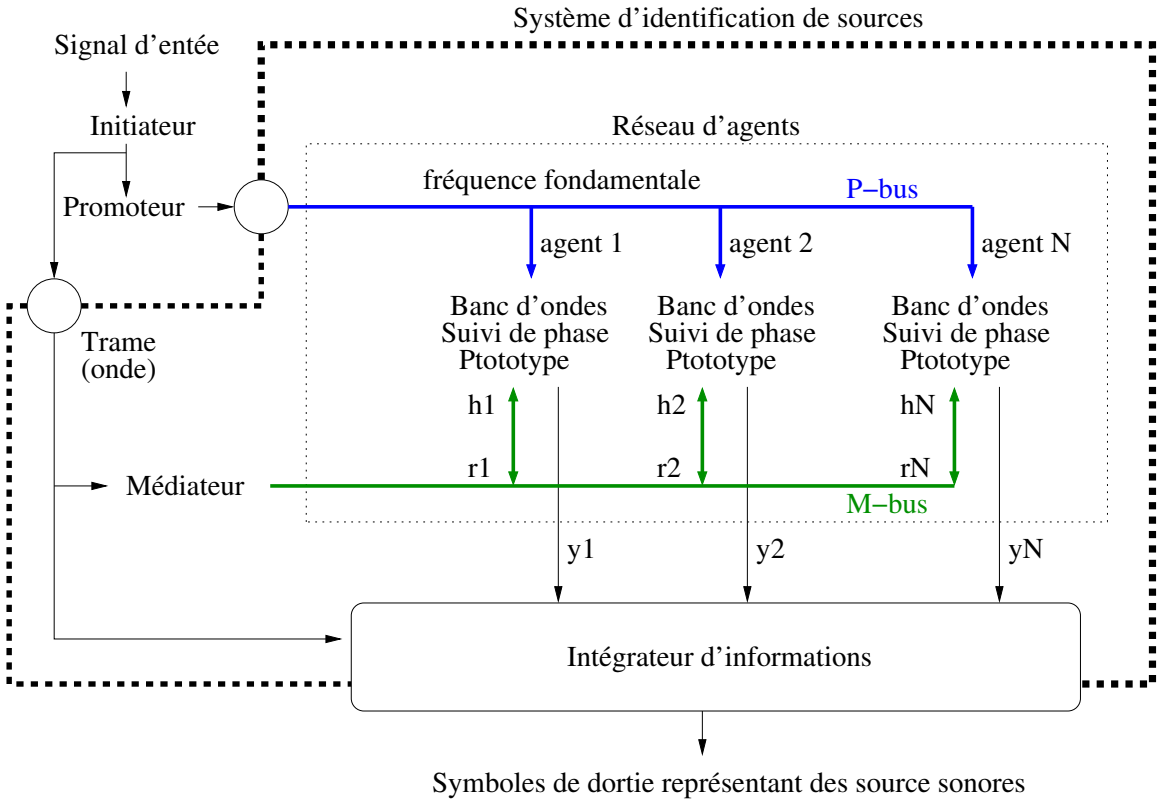


FIG. 3.3 – *Exemple d'architecture d'un système CASA multi-agents : le système Ipanema d'analyse musicale. Chaque agent est spécialisé dans la traque d'un aspect du signal, sous le contrôle d'un agent « médiateur ».*

3.2.2.3 Architecture neuromimétique

Wang et Brown [Wang 99] ont développé un réseau de neurones oscillants pour la ségrégation de flux audio. Un signal de parole bruité est traité en 4 étapes principales. Dans la première étape, l'activité du nerf auditif est simulée en passant le signal observé à un modèle du système auditif périphérique (filtres cochléaires et cellules ciliées de l'organe de Corti). Une représentation parcellaire de la scène est alors construite à partir d'un corrélogramme ainsi que des corrélations entre les différentes bandes de fréquences. Le réseau de neurones oscillants groupe ensuite différents composants acoustiques mis en évidence à l'étape précédente. Finalement, les signaux de la parole et du bruit sont reconstruits en accord avec la séparation de sources obtenue avec le réseau de neurones. Ce système améliore les SNR d'un signal de la parole corrompu par un bruit de fond mais ne surpasse pas les performances obtenues en séparation aveugle de sources lorsque le signal de la parole est corrompu par plusieurs sources de bruit différentes [van der Kouwe 01].

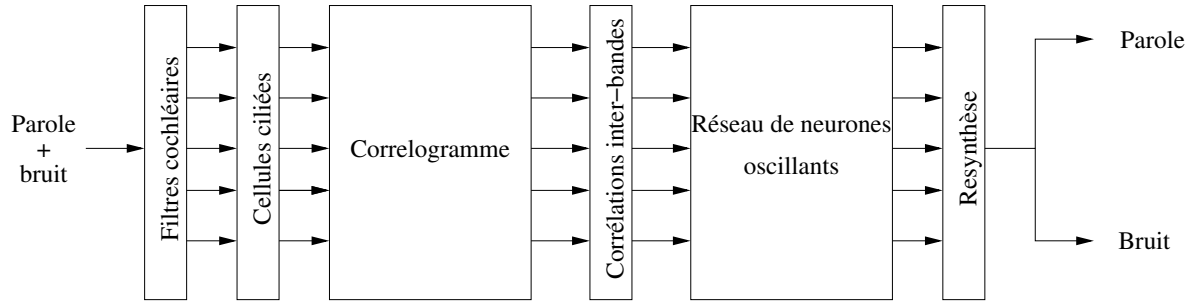


FIG. 3.4 – *Représentation schématique (reproduite à partir l'article original) du système proposé par Wang et Brown [Wang 99].*

Une autre architecture de réseaux de neurones a été proposée : le réseau cortronique [Sagi 01]. Cette architecture est proposée pour résoudre le problème de la séparation de sources lorsque le bruit est un autre signal de parole (le problème *cocktail party*). Ce système est basé sur la modélisation par un réseau de neurones à 3 couches d'une mémoire associative. La première couche sert d'encodeur du signal, la seconde traite les indices extraits par la première et finalement la troisième et dernière couche de neurones traite les hypothèses de mots formulées à partir de la couche précédente.

Différents concepts importants concernant le développement de systèmes CASA neuromimétiques sont discutés dans [Haykin 05]. Les auteurs proposent, dans le cadre du problème *cocktail party*, une architecture exploitant une nouvelle idée : l'audition active. Cette architecture est composée de 4 couches de neurones :

1. **Localisation** : détermination spatiale des différentes sources sonores. La localisation de sources peut être réalisée par un banc de microphones par exemple.
2. **Séparation et identification** : séparation des différentes sources sonores et focalisation sur l'une d'elles (celle correspondant au signal d'intérêt).
3. **Suivi** : l'évolution du signal d'intérêt est suivie et les futures évolutions prédites.
4. **Apprentissage** : les modèles internes sont continuellement mis à jour afin de tenir compte de l'évolution du monde perçu.

3.3 Traitement du signal et modèles statistiques

Les méthodes d'estimation de masques de données manquantes qui sont présentées dans cette partie ne sont pas fondées sur les principes de Bregman et ne s'inspirent pas du système auditif humain. Elles utilisent des concepts issus du traitement du signal et de la statistique.

3.3.1 Séparation basée sur le SNR local

3.3.1.1 Estimation du SNR

Comme nous l'avons mentionné au Aurora 2itre 2, la notion de données manquantes en reconnaissance automatique de la parole est fortement liée au SNR local. Par conséquent, la plupart des études menées en reconnaissance de la parole avec données manquantes proposent des méthodes d'estimation du SNR local permettant d'inférer les masques. Pratiquement, tout algorithme d'estimation du SNR local peut être employé dans ce but.

Le plus simple de ces algorithmes estime le SNR local à partir d'un modèle paramétrique du bruit, généralement calculé sur les segments du signal ne comportant pas de parole, à la manière de la soustraction spectrale [Cooke 97, Renevey 01a, Renevey 01b]. Cependant cette solution simple souffre d'un manque de robustesse vis-à-vis des bruits non stationnaires. Par conséquent, l'utilisation d'algorithmes plus sophistiqués est requise, comme par exemple les *Taylor Vector Series* (VTS) [Raj 00].

Dupont et Ris [Dupont 01] ont évalué et comparé trois méthodes d'estimation du SNR local : partitionnement des énergies (*energy clustering*), les histogrammes de Hirsh (*Hirsh histograms*) et suivi de l'enveloppe spectrale de faible énergie (*low-energy envelop tracking*). Bien que ces techniques améliorent la précision du SNR local, le gain n'est cependant pas suffisant en présence de bruit non-stationnaire.

3.3.1.2 Indices additionnels : l'harmonicité et modulation d'amplitude

D'autres indices acoustiques additionnels, tels que l'harmonicité ou la modulation d'amplitude [Hu 04], sont calculés et combinés avec des modèles de bruit.

Harmonicité

Les mesures d'harmonicité du signal ont été étudiées pour estimer les masques de données manquantes au niveau du vecteur ou du coefficient acoustique. Dans le cadre de l'estimation de masque au niveau du vecteur acoustique, une approche connexe, appelée parole utile (*usable speech*), vise à identifier et extraire les trames du signal les moins bruitées, c'est-à-dire utiles pour des applications de traitement de la parole (identification du locuteur ou reconnaissance de la parole). Dans ce contexte les deux mesures suivantes ont été proposées [Yantorno 03] :

1. **Adjacent Pitch Period Comparison** (APPC) : Cette mesure compare la périodicité de fréquences fondamentales adjacentes sur des portions du signal contenant de la parole voisée. La comparaison s'appuie sur une distance euclidienne. Cette distance est généralement proportionnelle à la quantité de bruit.
2. **Spectral Autocorrelation Peak Valley Ratio-Residual** (SAPVR-Residuelle) : Cette mesure détecte les portions de parole utile en examinant la structure d'autocorrélation

de la FFT du codage prédictif linéaire (LPC) résiduel [Chandra 02]. Quand il y a peu d'interférence entre les différents signaux composant la scène auditive, une représentation faite de crêtes et de vallées apparaît, et quand l'interférence est forte, les crêtes tendent à s'aplatir et les vallées à se remplir. Le succès de la mesure SAPVR-Residuelle dépend fortement de l'identification de ces crêtes et vallées, il est donc essentiel d'utiliser des algorithmes robustes de détection de crêtes.

Chandra et Yantorno [Chandra 02] fusionnent ces deux mesures. Une analyse en composantes indépendantes (ICA) est d'abord effectuée sur ces deux mesures pour éliminer l'information redondante et améliorer la qualité de la fusion [Hall 92]. La fusion est ensuite réalisée par un système d'erreurs au moindre carré du troisième ordre, ainsi que par un classifieur bayésien.

Lorsque les mesures d'harmonicité sont appliquées à chaque coefficient spectral, l'hypothèse sous-jacente est que les coefficients partageant les mêmes caractéristiques harmoniques peuvent être considérés comme issus de la même source sonore [de Cheveigne 95]. Les coefficients ne faisant pas partie de ce regroupement ont une forte probabilité d'être masqués. Par conséquent, Barker et ses collègues [Barker 01b] ont proposé de combiner un masque d'harmonicité avec le masque basé sur l'estimation du SNR local. Le masque d'harmonicité qu'ils proposent est obtenu en extrayant du corrélogramme les coefficients correspondant à la fréquence fondamentale. Ces coefficients sont ensuite mis à l'échelle par une fonction sigmoïdale afin de représenter un masque d'harmonicité (chaque valeur est un scalaire compris entre 0 et 1). Le masque résultant est alors une somme pondérée du masque d'harmonicité et du masque basé sur le SNR. Les poids de la somme sont inférés à partir du degré de voisement calculé pour chaque trame du signal.

Un cas plus difficile se produit lorsque le bruit est lui aussi harmonique. Une estimation simple de l'harmonicité du signal n'est pas suffisante, car elle ne permet pas de distinguer plusieurs harmoniques. L'utilisation d'algorithmes de détection et de suivi de plusieurs fréquences fondamentales simultanées est alors nécessaire [Parsons 76, de Cheveigne 93]. Une décomposition sinusoïdale du signal peut être utile pour cet objectif [Tolonen 00, Karjalainen 99]. Une fois le signal harmonique décomposé, ses composantes sinusoïdales peuvent être regroupées en une ou plusieurs sources. Le regroupement des composantes sinusoïdales est basé sur le rapport de leur fréquence. Virtanen et Klapuri [Virtanen 02] ont étendu ce principe et l'ont intégré au sein d'un système de séparation de sources. La fréquence fondamentale dominante du signal est détectée puis retranchée du signal. Cette opération est alors répétée sur le signal résiduel [Nakatani 02].

Modulation d'amplitude

Tchorz et Kollmeier [Tchorz 02] ont proposé une approche pour estimer le SNR local fondée sur des résultats neurophysiologiques obtenus dans le cadre d'études du système auditif chez les mammifères. Leur système exploite la modulation d'amplitude du signal étudié. Cette information est représentée par un spectrogramme de modulation d'amplitude (AMS) contenant

la modulation d'amplitude calculée pour chaque centre de bande de fréquences. Un réseau de neurones est alors entraîné sur une base d'AMS générée à partir de signaux de parole corrompus par du bruit. Une estimation du SNR est fournie au réseau comme référence. Finalement, une fois entraîné, le réseau de neurones estime le SNR d'un signal à partir de ses AMS. Les expériences menées ont montré la robustesse de cette approche vis-à-vis des bruits non-stationnaires. Les auteurs ont également également une extension de l'algorithme pour estimer le SNR local dans chaque bande de fréquence [Tchorz 01].

3.3.2 Réseaux de neurones

Potamitis et ses collègues [Potamitis 00a] utilisent un réseau de neurones pour détecter les trames de parole utile en présence de bruit impulsif. Les auteurs ont proposé une version de ce système dans le cadre de l'estimation de masque de données manquantes [Potamitis 00b]. Le spectre du signal est divisé en différentes bandes de fréquences. Un réseau de neurones de type *time delay* (Time Delay Neural Network : TDNN) est entraîné sur de la parole seule afin de prédire les observations futures en fonction des observations courantes. L'erreur de prédiction du TDNN est évaluée sur un corpus de développement. Durant la phase de test, les coefficients observés trop éloignés de leur estimation sont alors masqués. Finalement, un second réseau de neurones est utilisé pour inférer les valeurs des données manquantes.

3.3.3 Séparation de sources

Nous avons mentionné au Aurora 2itre **??**, paragraphe 2.2.2, que les masques oracles peuvent être calculés à partir des différents flux audio composant une scène auditive. Il est possible d'estimer les masques réels de la même manière, en utilisant des algorithmes de séparation de sources à la place d'un système CASA en amont de l'estimation des masques. Ces algorithmes de séparation de sources fournissent les différents signaux issus des différentes sources sonores composant la scène auditive. Ensuite, le signal du locuteur principal seul peut être reconnu, mais la séparation des signaux n'est pas optimale, et des phénomènes de distorsion peuvent dégrader le signal, ce qui pénalise fortement la reconnaissance. Le signal de la parole extrait par ces algorithmes n'est donc pas de qualité suffisante pour obtenir de bonnes performances de reconnaissance, et il est souvent préférable d'estimer des masques de données manquantes à partir des flux de la parole et du bruit, puis de reconnaître les mots prononcés par une des technique de reconnaissance en présence de données manquantes. La figure 3.5 illustre une telle approche.

La séparation de sources est un problème recouvrant plusieurs domaines de recherche en traitement de la parole : CASA, séparation aveugle de sources, séparation de signaux de parole concurrents [Quatieri 90, Yen 99]. Beaucoup d'algorithmes de séparation de sources aveugles exploitent un banc de microphones, ce qui permet de séparer les sources par leur localisation. Lorsqu'un seul microphone est considéré, Potamitis et ses collègues [Potamitis 01] proposent deux approches fondées sur une analyse en composante indépendante du signal :

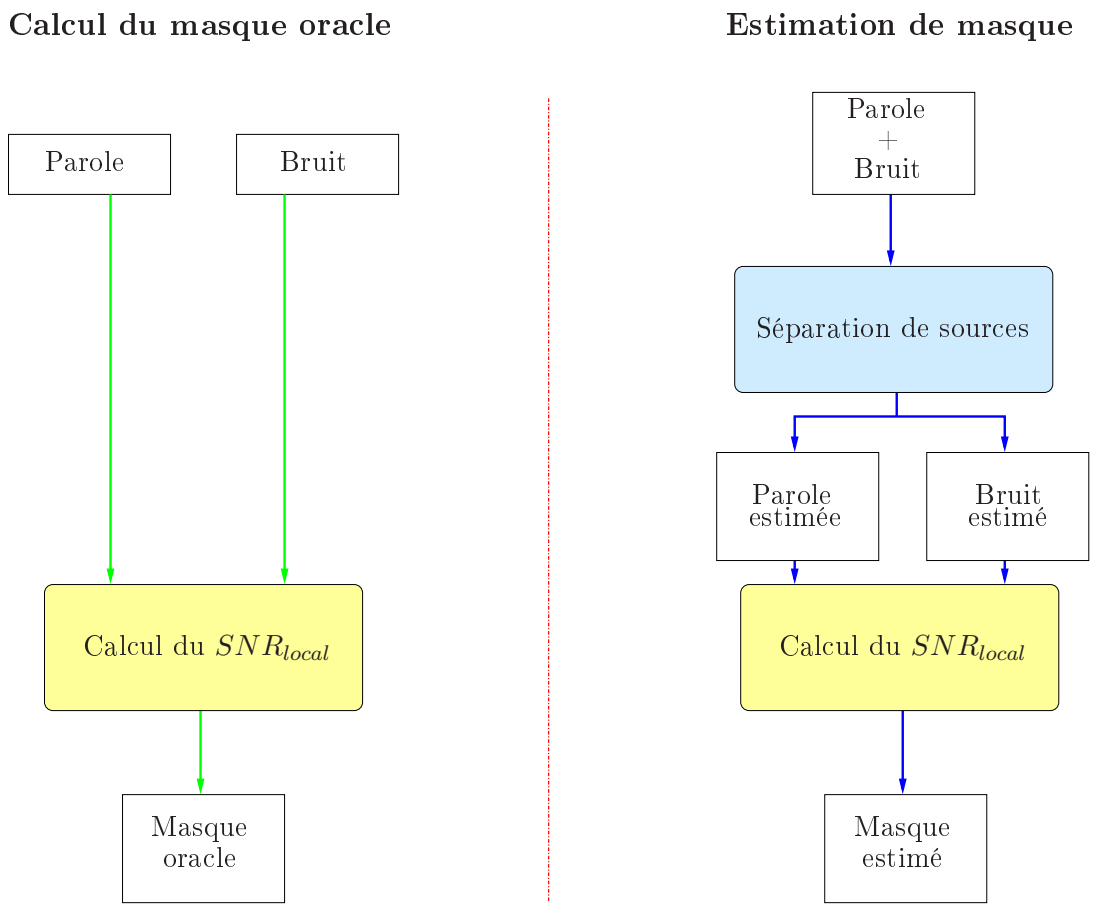


FIG. 3.5 – *Utilisation d'algorithmes de séparation de sources pour l'estimation de masques de données manquantes.*

Sparse Code Shrinkage

L'ICA fournit une matrice de projection W (domaine temporel $\rightarrow$ domaine des composantes indépendantes) du signal, calculée sur une base de données de parole seule, qui est par la suite orthogonalisée.

$$Z_i \quad = \quad W \; . \; X_i \tag{3.1}$$

$X = [x_1, x_2, \ldots, x_T]$ est le signal temporel, $Z = [z_1, z_2, \ldots, z_T]$ est le même signal mais exprimé dans le référentiel des composantes indépendantes, T est l'indice de l'échantillon et $W \in \mathcal{R}^{F \star N}$ avec N et F étant respectivement le nombre de composantes indépendantes et la longueur de la fenêtre d'analyse. Considérant l'indépendance des composantes z_i de Z, une loi de probabilité $p(z)$ est estimée sur cette même base de données de parole seule.

Soit Y un signal composé de deux signaux de parole X_1 et X_2. Ce signal est dans un premier

temps exprimé dans le domaine de projection :

$$Z = W.Y \tag{3.2}$$
$$= W.X_1 + W.X_2 \tag{3.3}$$
$$= U + V \tag{3.4}$$

Soit $U = [u_1, \ldots, u_T]$ la représentation du signal de la parole X_1 et $V = [v_1, \ldots, v_T]$ la représentation du signal de la parole X_2 dans le domaine de projection :

$$U^\star = \arg\max_U \prod_{i=1}^{N} p(z = u_i) \; . \; \prod_{i=1}^{N} p(z = v_i) \tag{3.5}$$
$$V^\star = Z - U^\star \tag{3.6}$$

Par la suite X_1 et X_2 peuvent être reconstruits :

$$X_1 = W^{-1} \; . \; U^\star \tag{3.7}$$
$$X_2 = W^{-1} \; . \; V^\star \tag{3.8}$$

Variational Bayes Approximation

Une matrice de démixage est estimée ici pour chaque trame du signal bruité. Chaque source est caractérisée par une probabilité *a priori* modélisée par un mélange de lois normales. Les paramètres du modèle de séparation sont ensuite inférés en minimisant la fonction de divergence de Kullback-Leibler entre l'approximation de la distribution des probabilités *a posteriori* et la vraie probabilité *a posteriori* $p(X|Y)$.

Bach et Jordan [Bach 05] proposent une alternative à l'ICA. Ils proposent d'entraîner un classifieur de données spectrales qui aura en charge de partitionner les spectrogrammes en deux classes, chacune d'elles correspondant aux régions dominées par le signal d'un des deux locuteurs. Ce classifieur est une matrice d'affinité qui encode les relations topologiques entre différents indices spectraux. Ces indices sont : la continuité temporelle et fréquentielle, la co-évolution temporelle de l'enveloppe spectrale de plusieurs bandes de fréquences, l'harmonicité et le timbre. Les paramètres de ce modèle sont entraînés sur des mélanges de signaux de parole issus de plusieurs locuteurs.

3.3.4 Modèles statistiques

L'approche la plus intuitive pour résoudre le problème d'estimation de masque de données manquantes à partir de modèles statistiques est de calculer la probabilité :

$$p(Y|M) \tag{3.9}$$

où Y et le signal bruité et M le masque associé à Y.

C'est l'approche suivie par Seltzer [Seltzer 00] et Raj Ramakrishnan [Raj 00]. Raj Ramakrishnan construit un classifieur bayésien pour chaque bande de fréquence. La classification est faite à partir de vecteurs d'observations composés de 5 indices spectraux caractérisant un coefficient spectral $Y(t, f)$ particulier.

$$\overrightarrow{Y(t,f)} = \begin{bmatrix} Y(t,f) \\ Y(t+1,f) - Y(t-1,f) \\ Y(t,f+1) - Y(t,f-1) \\ Y(t+1,f+1) - Y(t-1,f-1) \\ Y(t-1,f-1) - Y(t-1,f+1) \end{bmatrix}$$

Seltzer propose d'autres indices que Raj Ramakrishnan. L'architecture de son classifieur est également basée sur la construction de deux GMM (modèle à mélange de gaussiennes) par bande de fréquences : un pour les données fiables et le second pour les données manquantes. Les paramètres qu'ils proposent sont des caractéristiques du signal de la parole sur lesquels l'effet d'un bruit additif est connu :

Comb filter ratio : qui représente la proportion de l'énergie résidant en des fréquences harmoniquement corrélées. Ceci constitue une alternative au masque d'harmonicité de Barker.

Autocorrelation peak ratio : qui évalue la périodicité du signal. Cette mesure est corrélée au SNR sous l'hypothèse que le bruit ne soit pas lui-même harmonique.

Sub-band energy to full-band energy ratio : qui représente la forme de l'enveloppe du spectre de la parole.

Kurtosis : qui mesure la gaussianité du signal (un signal de parole seule a un kurtosis plus important qu'un signal de parole bruitée).

Flatness of the spectrales valleys : qui est corrélée au SNR. En effet, plus le bruit est important (plus le SNR est faible) et plus les vallées de l'enveloppe spectrale s'atténuent. Ceci découle directement de l'observation que les coefficients spectraux de faible énergie sont plus vulnérables au bruit que les coefficients hautement énergétiques.

Sub-band energy to sub-band noise floor : cet indice est également corrélé au SNR.

SNR estimation : Ce paramètre est l'estimation du SNR obtenue à partir de la soustraction spectrale.

Seltzer propose d'entraîner ses modèles de masques sur une base de données de parole seule corrompue par un bruit blanc dans le but d'améliorer la robustesse des modèles vis-à-vis d'environnements inconnus. Kim et ses collègues [Kim 05] ont remarqué que la robustesse des modèles de masques aux environnements inconnus n'est pas suffisante. Ils ont souligné le fait qu'entraîner ces modèles sur du bruit blanc n'est valide que si l'estimation du masque de chaque coefficient est indépendante des estimations des masques des coefficients adjacents, ce qui n'est pas le cas dans le domaine spectral. Ils proposent alors, comme alternative, de modéliser les variations spectrales du signal interférant en bruitant chacune des bandes de fréquences aléatoirement par du

bruit coloré de durée variable. Les résultats reportés montrent que cette méthode d'apprentissage est encourageante. Cette approche fut ensuite améliorée en entraînant des modèles de masques indépendants de la bande de fréquence [Kim 06].

3.3.5 Masque comme produit de la reconnaissance

Nous avons exposé dans les paragraphes précédents nombre de techniques dédiées ou transposables à l'estimation de masques de données manquantes. Ces techniques sont considérées dans le cas où le masque est estimé avant le processus de reconnaissance et utilisé par ce dernier comme une source de connaissances additionnelles (autres que les observations). Plus récemment, des études ont montré que les masques de données manquantes peuvent être produits durant l'étape de décodage. Ce sont ces approches que nous présentons dans ce paragraphe.

Quand deux modèles pour la parole et pour le bruit (ou de la parole concurrente) sont disponibles, il est possible de combiner ces deux modèles et de rechercher la meilleure séquence d'états qui maximise la probabilité *a posteriori* des observations dans l'espace des états combinés. De ce point de vue, Roweis [Roweis 03] (Fig. 3.6) propose une approche basée sur le traitement du signal en sous-bandes de fréquences pour séparer deux signaux de parole concurrente. En accord avec l'hypothèse de dominance, il suppose que tout coefficient spectral est dominé par l'énergie du signal d'un seul locuteur.

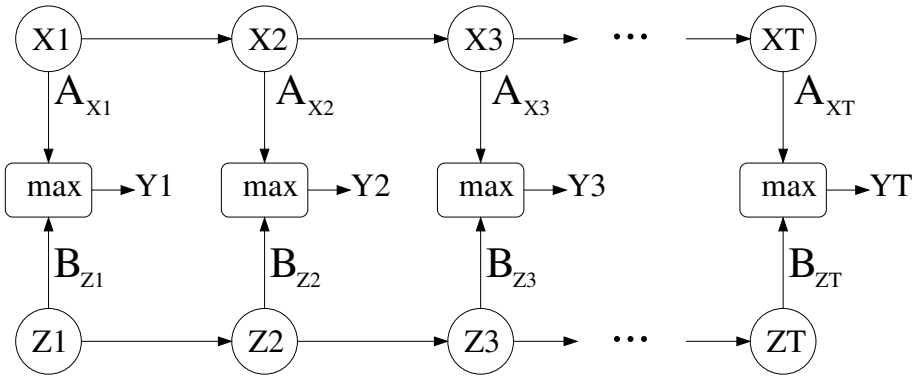


Fig. 3.6 – *HMM factoriel proposé dans [Roweis 03] : deux chaînes de Markov X_t et Z_t évoluent indépendamment, chacune représentant la réalisation acoustique d'une source sonore. Les observations bruitées Y_t sont alors considérées comme étant réalisées conjointement par ces deux chaînes de Markov.*

Le décodeur multi-sources proposé par Barker [Barker 01a] (Fig. 3.7) combine explicitement l'utilisation de regroupement d'observations et une architecture reposant sur les modèles acoustiques. Plus simplement, des principes de regroupement d'observations sont mis en œuvre de manière à constituer des fragments du spectre cohérents où les coefficients d'un même fragment

sont considérés comme étant issus de la même source sonore (approche ascendante). Le décodage est effectué en considérant toutes les hypothèses de regroupement de fragments. Le regroupement des fragments est donc guidé par les modèles acoustiques (approche descendante). Ce système s'apparente aux modèles ASA appelés *glimpsing model* [Cooke 03a, Cooke 03b]. Les *glimpsing models* s'inscrivent dans une représentation parcellaire du signal de la parole, où chaque parcelle est une région homogène du plan temps-fréquence issue d'une même source sonore.
L'identification des fragments spectraux cohérents, ainsi que la mise en œuvre de ce système dans le contexte de la reconnaissance de la parole avec données manquantes est présentée dans [Baker 05].

Certaines améliorations de ce système ont depuis été proposées [Barker 06]. Les différents fragments cohérents du signal sont exprimés de manière *soft* alors que ceux-ci l'étaient de manière binaire auparavant. Une nouvelle technique de calcul des vraisemblances des coefficients dynamiques a été proposée ainsi que l'utilisation d'un GMM afin d'estimer la probabilité *a priori* d'une observation. Cette distribution de probabilité est utilisée en lieu et place de la distribution uniforme auparavant utilisée lors de la marginalisation des données. Le nouveau système a été évalué dans le cadre du problème du *cocktail party*. Les bons résultats obtenus montrent la capacité d'un tel modèle à reconnaître un signal de parole en présence de parole concurrente.
Plus récemment, un nouveau processus de génération des fragments a été présenté [Christensen 07]. Ce nouveau processus s'appuie sur un suivi de fréquences fondamentales ainsi que sur des indices de localisation de sources (utilisation de 2 microphones).

La combinaison d'approches ascendantes et descendantes, comme le décodeur multi-sources par exemple, constitue une condition forte pour traiter les environnements acoustiques en reconnaissance de la parole. L'utilité d'une telle combinaison est discutée dans [Remez 94, Barker 99, Cooke 05]. Il est montré que l'utilisation d'approches ascendantes ne peut expliquer seule la capacité de notre système d'audition à gérer les environnements acoustiques.

3.4 Discussion

Nous avons proposé dans ce Aurora 2itre un état de l'art concernant l'estimation de masques de données manquantes. Nous ne nous sommes pas limités au seul cadre applicatif que constitue la reconnaissance de la parole. En effet, nous avons également prospecté des domaines connexes, comme la séparation aveugle de sources, l'analyse computationnelle de scène auditive ou encore la détection de parole utile. Nous avons montré que deux familles d'approches sont usuellement employées : les approches perceptives et les approches issues du traitement du signal. Celles-ci reposent sur un socle commun d'indices acoustiques. Cependant aucune étude, à notre connaissance, n'a montré quels indices étaient les mieux adaptés pour l'estimation de masques de données manquantes, ni comment ces indices devraient être combinés ou pondérés. Certes des travaux exploitant conjointement plusieurs de ces indices ont été proposés, mais ces travaux n'exploitent

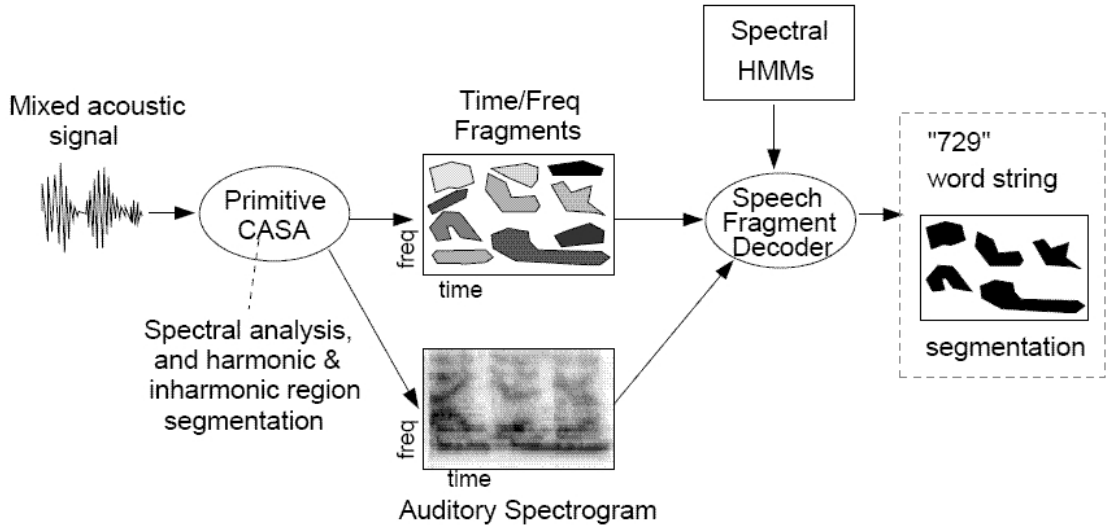


FIG. 3.7 – *Le décodeur multi-sources proposé par Barker combine explicitement l'utilisation de regroupement d'observations et une architecture reposant sur les modèles acoustiques (d'après [Barker 01a]).*

qu'une petite partie de ces indices et se résument très souvent à l'utilisation de l'harmonicité et d'une estimation du SNR. Les deux familles d'algorithmes que nous venons de décrire peuvent apparaître concurrentes. Il n'en est rien, des travaux récents [Barker 06] montrent qu'une combinaison de ceux-ci est très bénéfique.

La modalité d'évaluation de systèmes calculant et/ou utilisant des masques de données manquantes dépend du type d'application concernée. Dans le cadre de la reconnaissance automatique de la parole, une évaluation en terme de taux de reconnaissance est généralement privilégiée. Toutes les approches que nous venons de présenter, permettant d'inférer les masques de données manquantes, peuvent potentiellement être intégrées dans un système de reconnaissance avec données manquantes. Cependant, nombre de ces travaux n'ont pas réalisé cette intégration ou ne sont pas directement concernés par une évaluation à partir du taux de reconnaissance. Pour cette raison, les systèmes CASA sont souvent évalués qualitativement en terme de séparation de sources par l'intermédiaire d'un mesure traduisant l'amélioration du SNR [Cooke 93]. C'est le cas par exemple des travaux de Brown et Cooke [Brown 94] ainsi que ceux présentés dans [Wang 99] et [Hu 03]. Certains systèmes CASA sont évalués pour des tâches bien spécifiques. Par exemple le système de Godsmark et Brown [Godsmark 99] est évalué dans le cadre de la séparation de musique polyphonique en ses composantes mélodiques. Une alternative à l'amélioration du SNR est l'amélioration du TIR (Target to Interferer Ratio) [Yantorno 03] qui est identique au TMR [Barker 06] (Target to Masker Ratio). Des tests d'écoute sont également utilisés pour évaluer la qualité de la resynthèse du signal après avoir identifié les signaux de la

parole [Masuda-Katsuse 99, Kristjansson 04].

Très peu de résultats expérimentaux obtenus en combinant un système CASA avec un système de reconnaissance de la parole avec données manquantes ont été publiés jusqu'ici. La principale raison pourrait être que la reconnaissance n'est pas l'objectif principal du domaine CASA ou parce qu'il n'est pas évident de mettre en œuvre un tel système. De ce point de vue, nous pouvons citer l'article de Brown, Barker et Wang [Brown 01]. Les auteurs présentent un système combinant le réseau de neurones oscillant de Wang avec un moteur de reconnaissance de la parole avec données manquantes. Ce système est évalué sur une version bruitée de la base de donnée TiDigits. Le système combiné surpasse la soustraction spectrale et fournit un taux de reconnaissance de 47 % à 0 dB. Le décodeur multi-sources de Barker [Baker 05, Barker 06] est également évalué sur la version bruitée de TiDigits et fournit des résultats comparables à ceux mentionnés ci-dessus à 0 dB mais sans aucune dégradation en absence de bruit.
La plupart des systèmes de reconnaissance de la parole avec données manquantes sont également évalués en termes de taux de reconnaissance [Kim 05, Seltzer 00, Raj 00, Barker 00, Renevey 01a], mais une comparaison fine des différents résultats reportés n'est en général pas possible car les conditions expérimentales de ceux-ci diffèrent. Bien que l'utilisation de bases de données standardisées, telle qu'Aurora 2, offre l'opportunité de comparer différentes approches, l'évaluation en présence de bruit non stationnaire tel que de la parole concurrente ou de la musique n'est pas indiquée pour cette base.
Le réseau cortronique [Sagi 01] est lui aussi évalué en terme de taux de reconnaissance, mais les conditions expérimentales sont très spécifiques. Les signaux de parole de 1 à 20 locuteurs sont mixés avec des énergies à peu près identiques. Le vocabulaire est composé de 1024 mots. Le taux de reconnaissance dépasse les 98 % si l'on considère au plus 5 locuteurs, il chute à 70 % pour 10 locuteurs et 20 % pour 20 locuteurs. Ces résultats sont très encourageants surtout si l'on considère l'originalité de l'approche dans un contexte où les HMM occupent une place hégémonique.

Le domaine de la reconnaissance de la parole avec données manquantes souffre d'une définition de masque assez floue. En effet, différents types de masques peuvent être utilisés. Par exemple, les masques basés sur le SNR pourraient être utilisés de manière plus appropriée qu'ils ne le sont en marginalisation et imputation de données, le SNR_{local} fournissant d'importantes informations sur le signal. De plus, la définition d'un masque dans le domaine CASA n'est pas figée. Les approches CASA ne sont qu'un ensemble d'outils permettant de décrire une scène auditive, la définition d'un masque dépend alors de l'application envisagée [Wang 05].

Le but premier des systèmes CASA est de fournir une représentation du signal où la contribution de chaque source est clairement identifiée. Ceci se traduit par l'utilisation d'algorithmes complexes. On peut noter qu'une telle description détaillée du signal est inutile dans le cadre de la reconnaissance de la parole avec données manquantes, seule la contribution du signal d'intérêt est utile. Il n'est donc pas nécessaire de connaître la contribution de chaque signal interférant avec

celui d'intérêt. Néanmoins le domaine CASA est depuis longtemps mentionné par les chercheurs comme une source intéressante de concepts et de principes permettant d'estimer les masques de données manquantes. Cependant l'utilisation de ces principes pour la reconnaissance de la parole avec données manquantes commence seulement à émerger. Deux explications sont possibles. Premièrement, les taux de reconnaissance obtenus avec ces approches ne sont pas suffisamment satisfaisants, à l'exception de quelques travaux comme [Barker 06]. Une mise en œuvre efficace des principes de Bregman n'est pas aisée. Il s'agit de modéliser des principes physiologiques et cognitifs à partir de modèles et concepts mathématiques. Le choix de ces derniers constitue donc une étape cruciale. Deuxièmement, nous avons présenté des approches fondées sur la statistique ou le traitement du signal. Ces idées nouvelles constituent de sérieuses alternatives aux systèmes CASA. De ce point de vue, entraîner des modèles stochastiques de masques apparaît comme une voie intéressante. Les algorithmes de traitement du signal fondés sur une estimation du SNR montrent quelques difficultés à traiter les bruits non-stationnaires. Seltzer [Seltzer 00] et Raj Ramakrishnan [Raj 00] ont montré que des masques estimés à partir de modèles stochastiques sont bien moins pénalisés par la non stationnarité du bruit. Considérer le masquage comme un processus stochastique offre alors de nouvelles perspectives. Cependant très peu de travaux en ce sens ont été publiés jusqu'ici. Des nouvelles stratégies de modélisation de masque semblent nécessaires.

Chapitre 4

Deux nouvelles approches de modélisation des masques

« Ne dites pas que ce problème est difficile.
S'il n'était pas difficile, ce ne serait plus un problème. »

- Maréchal Foch -

Sommaire

4.1 Introduction

L'estimation de masque de données manquantes à partir de modèles stochastiques n'est pas nouvelle. Elle a été initiée à l'université de Carnegie Mellon par Seltzer [Seltzer 00] et Raj Ramakrishnan [Raj 00] sous la direction de Richard Stern et continue d'être développée par Kim [Kim 05, Kim 06].

Les travaux de Seltzer et Raj Ramakrishnan ont permis de montrer que cette approche permet d'estimer les masques de données manquantes aussi bien pour des bruits stationnaires, comme un bruit blanc gaussien, que pour des bruits non stationnaires comme de la musique. Les expérimentations reportées dans ces travaux ont montré que l'utilisation des masques estimés par le classifieur bayésien qu'ils proposent permet d'atteindre des taux de reconnaissance proches de ceux atteints avec des masques oracles en présence de bruit stationnaire tel le bruit blanc gaussien. De plus, les résultats reportés lorsque le signal interférant est de la musique montrent que les masques estimés à partir de ces modèles stochastiques permettent une bien meilleure reconstruction du spectre que des masques estimés à partir d'une estimation du SNR.

Une des difficultés liée à l'estimation de masques de données manquantes à partir de modèles stochastiques est l'indépendance des estimateurs vis-à-vis de l'environnement acoustique. Seltzer et Raj Ramakrishnan ont montré le fort potentiel de cette approche en présence de bruit non-stationnaire telle la musique. Cependant les meilleures taux de reconnaissance ont été obtenus lorsque le bruit de test était présent dans la base d'entraînement des modèles de masques. Ce problème d'indépendance vis-à-vis de l'environnement fut adressé par Kim [Kim 05]. Il a souligné le fait qu'entraîner ces modèles sur du bruit blanc, tel que l'ont fait Seltzer et Raj Ramakrishnan, n'est valide que si l'estimation du masque de chaque coefficient est indépendante des estimations des masques des coefficients adjacents. Ils proposent alors, comme alternative, de modéliser les variations spectrales induites par l'environnement acoustique sur une base de données d'entraînement en corrompant chacune des bandes de fréquences aléatoirement par du bruit coloré de durées variables. Les résultats reportés montrent que cette méthode d'apprentissage est encourageante. Cette approche fut ensuite améliorée en entraînant des modèles de masques indépendants de la bande de fréquence [Kim 06].

L'estimateur de masque de données manquantes développé par Seltzer et Raj Ramakrishnan repose sur l'utilisation d'un couple de GMM pour chaque bande de fréquence du spectrogramme. Chaque GMM modélise respectivement la distribution des vecteurs de paramètres ω dans l'espace des indices acoustiques Ω pour la classe des données manquantes et celle des données fiables. Une telle architecture peut être motivée par le fait que le bruit n'affecte pas le signal de la parole de la même manière dans chaque bande de fréquence. En effet, si l'on considère le masquage comme un seuillage du SNR local, les régions spectrales du signal de la parole les moins énergétiques sont plus vulnérables à l'effet du bruit, et inversement les régions hautement énergétiques du spectre

de la parole sont moins affectées par le bruit. Seltzer et Raj Ramakrishnan traitent alors chaque coefficient comme une entité indépendante. Cependant, l'énergie du signal de la parole est très localisée dans le plan temps-fréquence et par conséquent le processus de masquage ne peut être considéré comme un processus aléatoire. Il apparaît nécessaire d'introduire des corrélations entre les masques de chaque coefficient. Ce point fut abordé par Seltzer. Il proposa deux solutions. La première consiste à étendre la paramétrisation d'un coefficient par celles de ces neuf voisins dans le plan temps-fréquence. La deuxième consiste à appliquer un filtre médian sur les masques estimés de façon à éliminer des erreurs locales de masque. Cependant, ces solutions n'ont pas permis d'obtenir une amélioration significative des taux de reconnaissance.

Les masques de données manquantes fournissent de l'information additionnelle exploitée dans le but de reconstruire le signal (imputation de données) ou d'adapter le calcul des vraisemblances des observations (marginalisation de données). Plusieurs raffinements de ces deux approches ont été proposés. Ces raffinements sont dérivés en grande partie des propriétés du domaine Mel spectrale ainsi que du critère de fiabilité utilisé pour estimer les masques. Dans le premier cas, des algorithmes d'imputation et de marginalisation bornées ont montré qu'il est possible d'exploiter avantageusement le fait que les valeurs des coefficients spectraux sont des énergies et qu'elles sont par conséquent bornées inférieurement par zéro. De plus, en supposant l'additivité des énergies dans le spectre, les valeurs de ces coefficients ont une borne supérieure correspondante à l'énergie observée. Dans le deuxième cas, le critère de fiabilité permet de définir des intervalles de marginalisation spécifiques aux données manquantes et aux données fiables. Ceci a conduit à l'algorithme de marginalisation qualifiée d'uniforme-uniforme proposé par Morris [Morris 01a]. Ce schéma de marginalisation est fortement lié à la définition du masque de données manquantes reposant sur un seuillage du SNR local. L'évaluation de différents schémas de marginalisation que nous avons proposés (Fig. 2.2 page 47) montre qu'il est possible d'améliorer les taux de reconnaissance en réduisant les intervalles de marginalisation et qu'il est bénéfique de remettre en cause l'hypothèse de dominance en postulant que tout coefficient n'est jamais complètement manquant ou fiable.

Nous adressons, dans ce chapitre, le problème de la modélisation des masques de données manquantes. Nous proposons un nouveau modèle de masque permettant de modéliser les corrélations entre les masques des coefficients spectraux. Le masque de chaque coefficient spectral est alors considéré dans son contexte et non plus comme une entité indépendante. Deux types de corrélation sont considérés : les dépendances temporelles et les dépendances fréquentielles. Nous proposons ensuite une nouvelle caractérisation de masques dans le but d'améliorer leur prise en compte durant la phase de reconnaissance. Nous montrons qu'il est possible de réduire les intervalles de marginalisation à partir de cette nouvelle définition comparativement aux masques fondés sur le seuillage du SNR. Une évaluation des approches que nous proposons dans ce chapitre est présentée au chapitre 5 dans le cadre de la marginalisation de données.

4.2 Dépendances temporelles et fréquentielles sur les valeurs de masques

4.2.1 Introduction

Raj et Seltzer ont proposé un estimateur traitant chaque coefficient spectral comme une entité indépendante. Bien qu'erronée, cette hypothèse permet l'utilisation de modèles de masques spécifiques pour chaque bande de fréquence. Cependant le principal inconvénient d'une telle modélisation est l'incapacité à restituer convenablement la structure spectrale des masques.

L'objectif de la prise en compte de dépendances sur les valeurs des masques est d'améliorer la structure des masques estimés. En effet, les masques dépendent des caractéristiques du bruit mais aussi des caractéristiques du signal de la parole. Les régions spectrales où l'énergie du signal de la parole est faible sont plus vulnérables aux effets d'un bruit additif. La structure spectrale d'un masque reflète alors généralement la structure de l'enveloppe énergétique du signal de la parole d'intérêt. La similitude entre l'enveloppe énergétique du signal de la parole et la structure spectrale des masques est illustrée par la figure 4.1. La structure des régions spectrales fiables du signal bruité reflètent l'enveloppe énergétique du signal de parole. En condition très bruitée, -5 dB par exemple, les coefficients spectraux fiables sont les coefficients résidant aux fréquences harmoniques du signal car c'est à ces fréquences que l'énergie du signal de parole se concentre.

Nous proposons un nouveau modèle stochastique de masques permettant de prendre en compte les dépendances structurelles existantes entre les différentes valeurs de masques. Nous définissons deux types de dépendances : les dépendances temporelles et les dépendances fréquentielles. Nous décrirons dans un premier temps comment ces dépendances sont prises en compte dans le processus d'estimation de masque, puis nous proposerons quatre estimateurs de masques afin d'évaluer l'impact de chaque dépendance sur la qualité des masques et sur le taux de reconnaissance.

4.2.2 Dépendances fréquentielles

Une des caractéristiques de la paramétrisation spectrale est que les coefficients d'un même vecteur d'observations sont très corrélés. L'utilisation de matrices de covariances pleines au sein des modèles acoustiques permet de rendre compte des corrélations inter-coefficients et fournit de meilleures performances que des modèles s'appuyant sur des matrices de covariances diagonales sous l'hypothèse d'indépendance des coefficients. En assumant que la structure des masques de données manquantes est fortement liée à l'enveloppe spectrale du signal de la parole, les valeurs de masques devraient obéir aux mêmes corrélations que les coefficients spectraux.

Nous proposons alors de considérer le masque d'un vecteur d'observations comme une entité à part entière. Cette considération implique d'estimer un masque de données manquantes vectoriel plutôt que d'estimer le masque de chaque coefficient indépendamment. L'originalité de cette

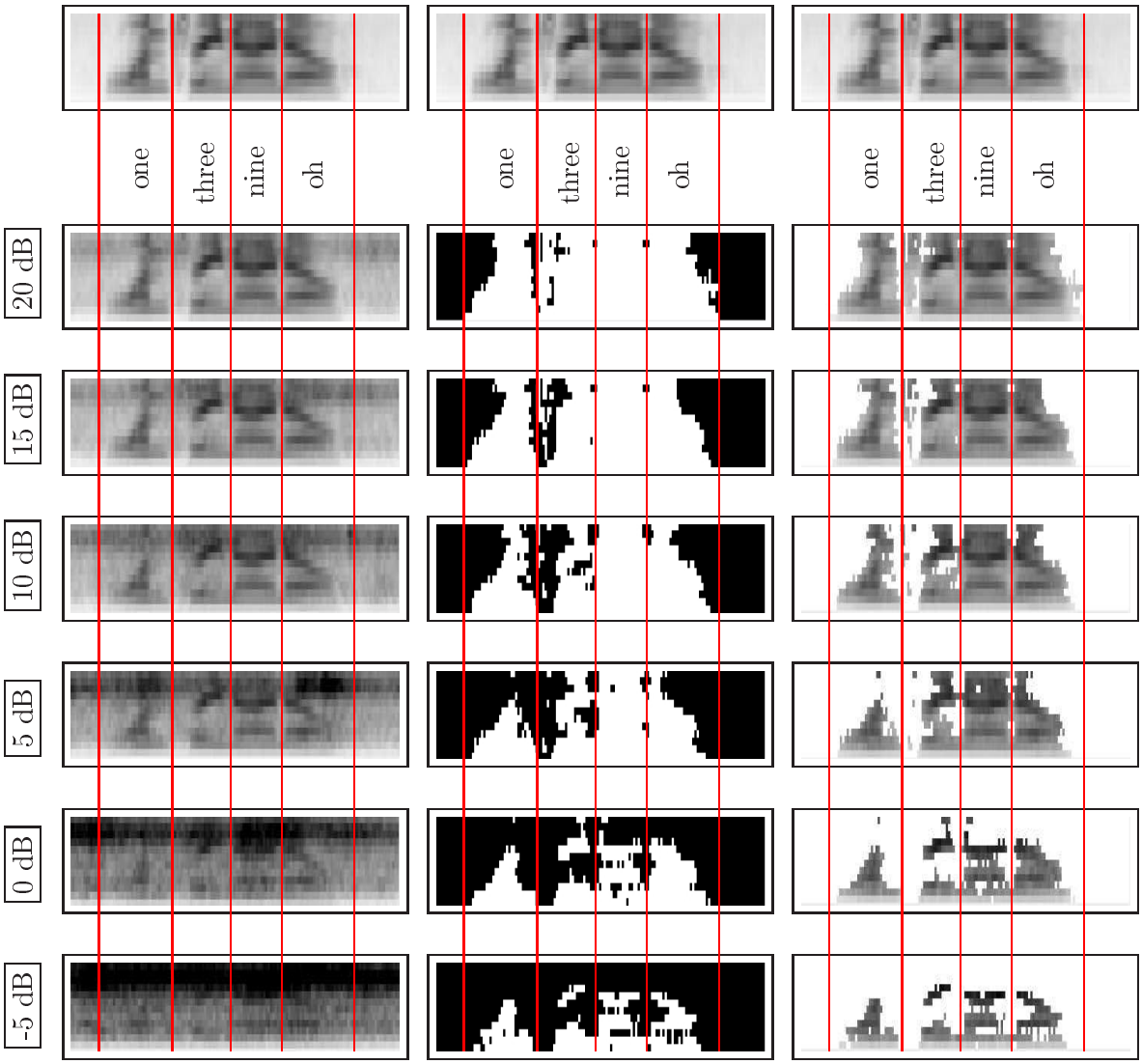


FIG. 4.1 – *Illustration de la structure spectrale des masques de données manquantes pour un signal de parole corrompu par le bruit du métro à divers SNR. Haut : Mel-spectrogrammes du signal de la parole seule. Gauche : Mel-spectrogrammes du signal corrompu par le bruit du métro. Centre : masques oracles associés aux signaux de parole corrompus. Droite : Mel-spectrogrammes des fichiers bruités. Seules les énergies des coefficients fiables sont représentées. Ces exemples mettent en évidence le lien entre la structure spectrale des masques et l'enveloppe énergétique du signal de la parole.*

approche est que nous ne distinguons plus les données manquantes des données fiables puisque le masque d'un vecteur acoustique inclut des données masquées et non masquées.

L'un des principaux inconvénients de cette vision plus large d'un masque est l'explosion de l'espace de recherche. En effet, soit un vecteur spectral y_t composé de D coefficients spectraux, il

existe alors 2^D masques vectoriels éligibles $m_t = [m_{1,t}, m_{2,t}, \ldots, m_{D,t}] \in 0,1^D$. Une approche naïve consiste alors à entraîner un modèle pour chacun d'eux. Cependant la distribution des vecteurs acoustiques du signal de parole dans le domaine spectral ne couvre pas la totalité de l'espace acoustique, elle est, au contraire, très localisée en des régions homogènes de cet espace. Puisque les portions du plan temps-fréquence correspondant à des régions hautement énergétique du signal de la parole sont moins sensibles aux effets du bruit additif, l'espace des masques vectoriels est probablement très localisé lui aussi.

Nous proposons d'identifier des masques vectoriels élémentaires. Nous considérons les masques vectoriels élémentaires comme les entités élémentaires à partir desquelles nous pouvons exprimer tout masque associé à un signal de parole bruitée. A partir des bases d'apprentissage propre et bruitée d'Aurora 2, nous calculons les masques oracles associés aux enregistrements de la base bruitée. Nous déterminons alors l'ensemble des masques vectoriels élémentaires comme les masques dont les occurrences cumulées expliquent α % des masques oracles.

La figure 4.2 fournit le nombre de masques vectoriels différents couvrant α % des masques oracles pour des paramétrisations spectrales comportant de 12 à 24 coefficients. D'une part, les masques oracles sont effectivement très localisés dans l'espace des masques. Par exemple, pour une paramétrisation à 12 coefficients, les masques oracles sont exprimés à partir de 3044 masques vectoriels différents sur les 4096 possibles ; 25 % des masques vectoriels éligibles ne sont donc jamais observés. De manière plus flagrante, moins de 1 % (120 000 sur 16 000 000) des masques éligibles pour une paramétrisation spectrale à 32 coefficients couvrent la totalité des masques oracles. D'autre part, la figure 4.2 illustre clairement l'explosion du nombre de masques vectoriels liée à l'augmentation de la dimension des vecteurs acoustiques. En effet, le nombre de masques vectoriels est exponentiel par rapport au nombre de coefficients. Cependant, pour une paramétrisation de faible dimension (12 coefficients par exemple), une grande proportion des masques oracles peut être expliquée par seulement quelques masques de trames. Le choix d'une paramétrisation de faible dimensionalité semble alors s'imposer.

L'identification des masques vectoriels élémentaires est une étape critique. Un simple seuillage du taux de couverture des masques oracles est bien évidemment insuffisant. Nous définissons l'ensemble des masques vectoriels élémentaires comme le plus petit ensemble de masques vectoriels expliquant au mieux les masques oracles, tout en minimisant la dégradation de taux de reconnaissance induite par la réduction de l'espace des masques. La figure 4.3 propose une évaluation des taux de reconnaissance obtenus sur la base d'apprentissage bruitée d'Aurora 2 à partir d'ensembles de masques vectoriels élémentaires correspondants à différents seuils de couverture des masques oracles. Cette expérimentation a été réalisée pour des paramétrisations à 12 et 24 coefficients spectraux.

Pour chaque seuil de couverture α ; un ensemble de masques vectoriels élémentaires est déter-

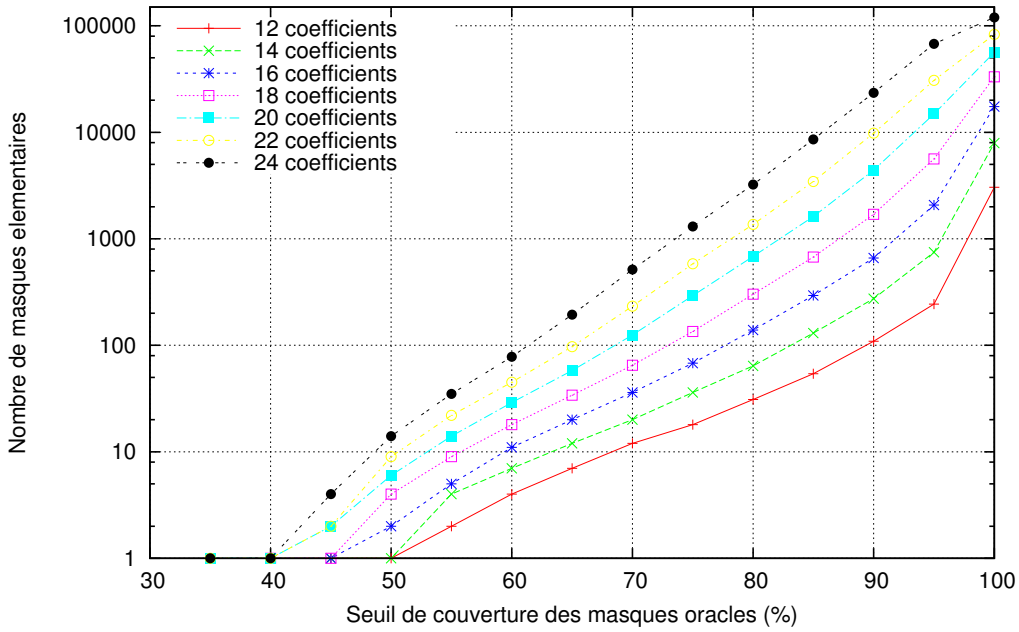


FIG. 4.2 – *Évaluation du nombre de masques élémentaires en fonction du seuil de couverture α des masques oracles et du nombre de coefficients spectraux sur Aurora 2.*

miné : les N masques les plus fréquents tels que ces N masques expliquent α % des masques oracles. Les masques oracles sont ensuite réexprimés à partir de ces seuls masques vectoriels élémentaires. Tout masque vectoriel des masques oracles ne faisant pas partie de l'ensemble des masques vectoriels élémentaires est substitué par le masque vectoriels élémentaire dont il est le plus proche en accord avec la distance euclidienne.

Nous identifions sur la figure 4.3, pour chacune des paramétrisations à 12 et 24 coefficients, des seuils de couverture intéressants (points de fonctionnement). 445 masques vectoriels élémentaires expliquent 70 % des masques oracles pour une paramétrisation à 24 coefficients et 31 masques expliquent 80 % des masques oracles pour une paramétrisation à 12 coefficients. La réduction de l'espace des masques pour ces points de fonctionnement n'entraîne qu'une faible dégradation des taux de reconnaissance de 95.36 % à 93.99 % pour 12 coefficients et de 96.22 % à 95.32 % pour 24 coefficients. Puisque la base d'apprentissage bruitée d'Aurora 2 comporte 4 bruits différents à divers SNR, nous supposons que les masques vectoriels élémentaires sont tributaires de l'enveloppe spectrale de la parole. Les autres masques vectoriels, les moins fréquents, ne contribuent que très peu à la reconnaissance. Nous supposons alors que ces masques sont tributaires des caractéristiques spectrales des différents bruits.

Il est tout à fait envisageable d'entraîner 445 modèles de masques vectoriels élémentaires, mais une grande partie d'entre eux ne se produit que rarement. Si l'on ordonne les masques vectoriels élémentaires par occurrence décroissante, alors le $445^{\text{ième}}$ masque est observé moins d'une fois

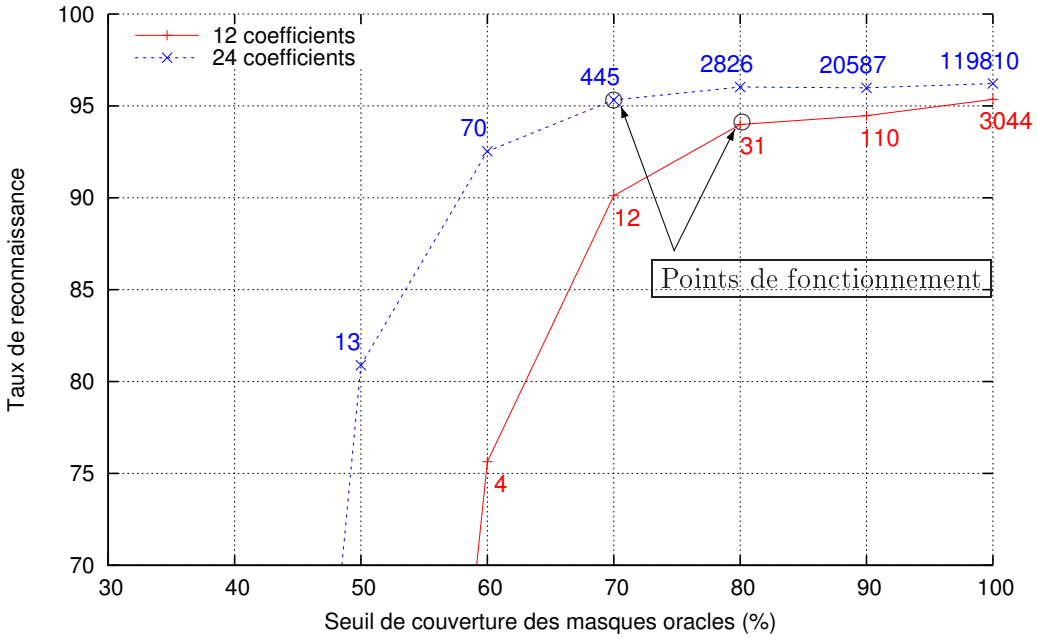


FIG. 4.3 – *Détermination du nombre de masques vectoriels élémentaires sur Aurora 2.*

sur 20000. Un problème d'apprentissage des modèles de masques élémentaires se pose alors. Nous ne disposons pas de données d'entraînement en quantité suffisante pour estimer correctement les modèles des masques élémentaires les moins fréquents. Cependant pour une paramétrisation à 12 coefficients, le $31^{\text{ième}}$ masque vectoriel élémentaire est observé 1 fois sur 300. Nous retiendrons donc une paramétrisation reposant sur des vecteurs acoustiques Mel spectraux à 12 coefficients statiques et un ensemble de 31 masques vectoriels élémentaires pour les expérimentations reportées au chapitre suivant.

4.2.3 Dépendances temporelles

Les dépendances temporelles sont représentées en modélisant les probabilités de transitions sur les modèles de masques. Cela revient à l'application d'une grammaire bigramme lors du processus d'estimation des masques. Ces probabilités de transitions sont calculées sur les masques oracles des enregistrements de la base d'apprentissage bruitée. En pratique, un HMM ergodique est construit. Chaque état correspond à un modèle de masque. Lorsque le masque est estimé pour chaque bande de fréquence, ce HMM est constitué de 2 états, le premier pour les données fiables et le second pour les données manquantes. Lorsque le masque est estimé pour un vecteur d'observations, le HMM est constitué de K états, chacun d'eux étant un modèle d'un des K masques vectoriels élémentaires.

Bien que ces transitions soient définies localement, leur influence est globale puisque le masque résultant est obtenu par l'algorithme de Viterbi, et par conséquent maximise la vraisemblance

des observations sur toute la phrase.

4.2.4 Estimateurs de masques

Dans le but d'évaluer la contribution de chaque dépendance, temporelle et fréquentielle, nous proposons quatre estimateurs de masques. Le premier estimateur est l'estimateur de référence : chaque masque de coefficient est estimé de manière indépendante, aucune dépendance n'est prise en compte. Le second exploite les dépendances temporelles seules, le troisième exploite les dépendances fréquentielles seules et la quatrième exploite de manière conjointe les dépendances temporelles et fréquentielles. Nous proposons par la suite une brève description de ceux-ci.

Estimateur AD : Aucune Dépendance (Fig. 4.4(a))

Pour chaque bande de fréquence i, 2 GMM modélisent respectivement les distributions $p(y_t|m_{i,t}=0)$ et $p(y_t|m_{i,t}=1)$ des vecteurs acoustiques pour lesquels le i^{me} coefficient est fiable ($m_{i,t}=0$) et manquante ($m_{i,t}=1$). Ce système est notre système de référence. Le masque $m_{i,t}$ associé à tout coefficient spectral bruité $y_{t,i}$ est déterminé comme suit :

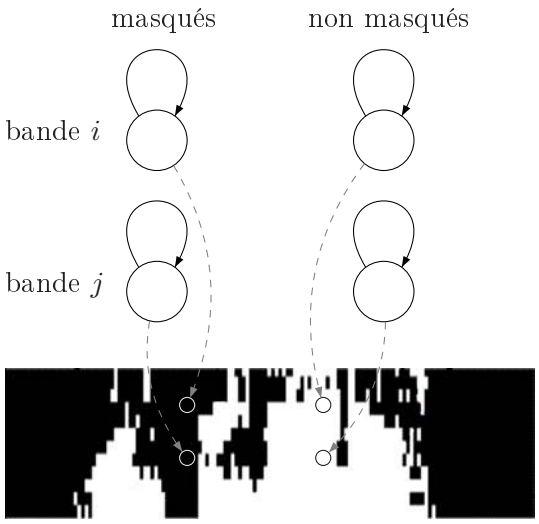
$$\begin{aligned} m_{i,t} &= 1 \quad \text{if} \quad p(y_t|m_{i,t}=1).p(m_{i,t}=1) > p(y_t|m_{i,t}=0.p(m_{i,t}=0) \\ &= 0 \quad \text{sinon} \end{aligned} \tag{4.1}$$

Cet estimateur de masques est identique à celui utilisé par Seltzer, Raj Ramakrishnan et Kim.

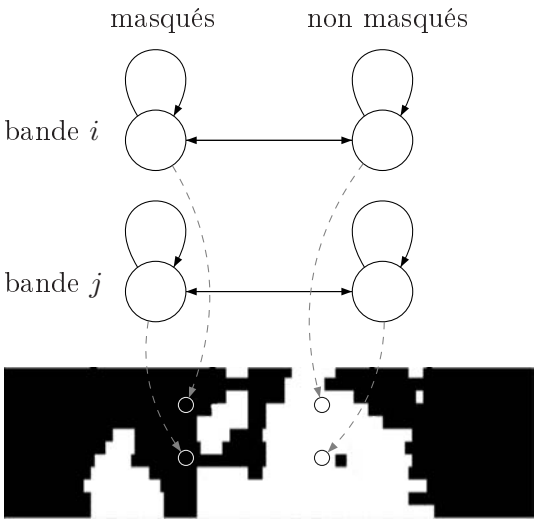
Estimateur DT : Dépendances Temporelles (Fig. 4.4(b))

Pour cet estimateur un modèle de masque est construit pour chaque bande de fréquence i. Ce modèle est un HMM ergodique à 2 états. Le premier état appelé q_o modélise la distribution $p(y_t|m_{i,t}=0)$ des vecteurs acoustiques pour lesquels le i^{me} coefficient est fiable. Le second état appelé q_1 modélise la distribution $p(y_t|m_{i,t}=1)$ des vecteurs acoustiques pour lesquels le i^{me} coefficient est masqué. Les probabilités de transition entre les états sont entraînées indépendamment des paramètres des GMM. Soit $s_i = (s_{i,1}, \cdots, s_{i,T})$ une séquence d'état associé à la bande de fréquences i, avec $s_{i,t} \in \{q_0, q_1\}$. Le masque de données manquantes associée à cette bande de fréquence est obtenu à partir de la meilleure séquence d'états $\hat{s}_i$ qui maximise la vraisemblance des observations.

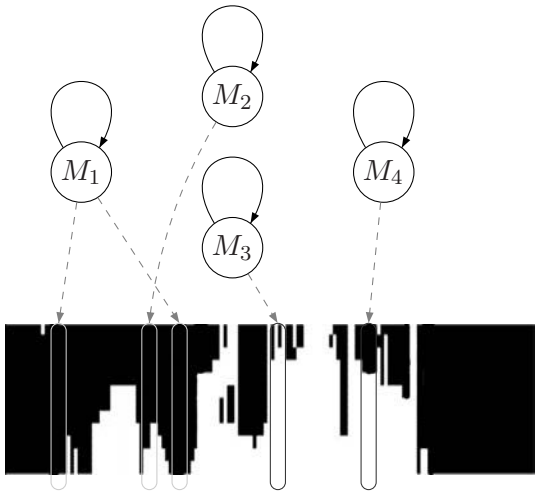
$$\begin{aligned} \hat{s}_i &= \arg\max_{s_i} \quad p(s_i|y) \\ &= \arg\max_{s_i} \quad \prod_{t=0}^{t=T} p(s_{i,t}|y_t).p(s_{i,t}|s_{i,t-1}) \end{aligned} \tag{4.2}$$



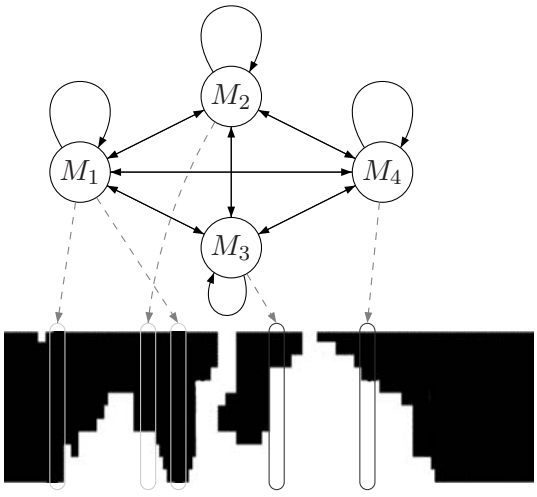
(a) **Estimateur AD** : Aucune Dépendance
Masques indépendants par bande et par trame



(b) **Estimateur DT** : Dépendances Temporelles
Masques par bandes avec transitions temporelles



(c) **Estimateur DF** : Dépendances Fréquentielles
Masques élémentaires indépendants par trame



(d) **Estimateur DTF** : Dépendances Temporelles et Fréquentielles
Masques élémentaires avec transitions temporelles

FIG. 4.4 – *Représentation des 4 estimateurs de masques dans le cadre de l'évaluation des dépendances sur les masques : (a) aucune dépendance, (b) dépendances temporelles seules, (c) dépendances fréquentielles seules et (d) dépendances temporelles et fréquentielles.*

Estimateur DF : Dépendances Fréquentielles (FIG. 4.4(c))

Soit $\mathcal{C}$ l'ensemble des K masques élémentaires $\tilde{m}^1, \ldots, \tilde{m}^K$. Pour chaque masque $\tilde{m}^i$, un GMM $p(y_t|m_t = \tilde{m}^i)$ est entraîné sur tous les vecteurs acoustiques ayant $\tilde{m}^i$ pour masque.

De cette manière, chaque GMM modélise un masque vectoriel et non plus le masque d'un seul coefficient. Durant l'estimation, le masque m_t associé à chaque vecteur acoustique y_t est calculé comme suit :

$$m_t = \arg\max_{\tilde{m}^i} \; p(\tilde{m}^i|y_t).p(\tilde{m}^i) \tag{4.3}$$

avec $\tilde{m}^i \in \mathcal{C}$.

Estimateur DTF : Dépendances Temporelles et Fréquentielles (FIG. 4.4(d))

Le dernier estimateur de masque exploite les dépendances temporelles et fréquentielles. Le modèle de masque est un HMM ergodique composé de K états $(q_1, \ldots, q_K)$, où chaque états contient respectivement un des K GMM définis ci-dessus. Soit $s = (s_1, \cdots, s_T)$ une séquence d'états, avec $s_t \in \{q_1, \ldots, q_K\}$. Le masque de données manquantes associé à un enregistrement bruité est obtenu à partir de la meilleure séquence d'états $\hat{s}$ qui maximise la vraisemblance des observations.

$$\begin{aligned} \hat{s} &= \arg\max_{s} \; p(s|y) \\ &= \arg\max_{s} \prod_{t=1}^{t=T} p(s_t|y_t).p(s_t|s_{t-1}) \end{aligned} \tag{4.4}$$

4.3 Une nouvelle caractérisation des masques de données manquantes

4.3.1 Introduction

Un système de reconnaissance automatique de la parole avec données manquantes s'articule autour de la définition même des données manquantes ainsi que sur des propriétés propres au domaine de paramétrisation. Dans le cadre de la marginalisation de données, la définition d'une donnée manquante joue un rôle de premier plan. En effet, nous avons montré que les bornes de marginalisation peuvent être affinées en prenant en compte le critère permettant d'inférer la fiabilité des paramètres acoustiques. Il est montré au chapitre 2 que reconnaître de la parole en présence de données manquantes revient à rechercher la séquence d'états acoustiques $Q^\star$ qui maximise l'espérance de la vraisemblance des T vecteurs d'observations bruitées Y de dimension D :

$$Q^\star = \arg\max_{Q} \; \left(\pi_0 \,.\, \prod_{t=1}^{T} a_{q_{t-1}q_t}\right) . \left(\prod_{t=1}^{T} \int_{\Lambda} b_{q_t}(x_t) \,.\, P(x_t|Y,\kappa) \; \mathrm{d}x_t\right) \tag{4.5}$$

Sous l'hypothèse d'indépendance des coefficients d'un même vecteur acoustique nous pouvons reformuler l'équation 4.5 :

$$Q^\star = \arg\max_{Q} \; \left(\pi_0 \,.\, \prod_{t=1}^{T} a_{q_{t-1}q_t}\right) . \left(\prod_{t=1}^{T} \prod_{i=1}^{D} \int_{\Lambda} b_{q_t}(x_{t,i}) \,.\, P(x_{t,i}|Y,\kappa) \; \mathrm{d}x_{t,i}\right) \tag{4.6}$$

L'incertitude associée à la valeur de chaque coefficient spectral bruité $y_{t,i}$ est modélisée par le terme $P(x_{t,i}|Y,\kappa)$ qui est une fonction de densité de probabilité de la variable aléatoire $x_{t,i}$ représentant l'énergie du signal de la parole seule au temps t et pour la $i^{\text{ème}}$ bande de fréquence. Cette fonction de densité de probabilité est conditionnée par les observations ainsi que divers critères comprenant entre autres le critère de fiabilité, ou de manière plus générale la caractérisation des différentes classes de données.

Un masque de données manquantes est généralement perçu comme un bi-partitionnement du spectrogramme, séparant les données fiables des données manquantes. Le critère de partitionnement, ou critère de fiabilité, est fondé sur le seuillage du SNR local. L'exploitation de ce critère a conduit Morris [Morris 01a] à proposer le schéma de marginalisation UUmarg. Cette marginalisation se distingue des autres par le fait que les deux types de données, manquantes et fiables, sont marginalisées sur des intervalles disjoints et dont l'union forme l'intervalle $[0, y_{t,i}]$ (FIG. 4.5).

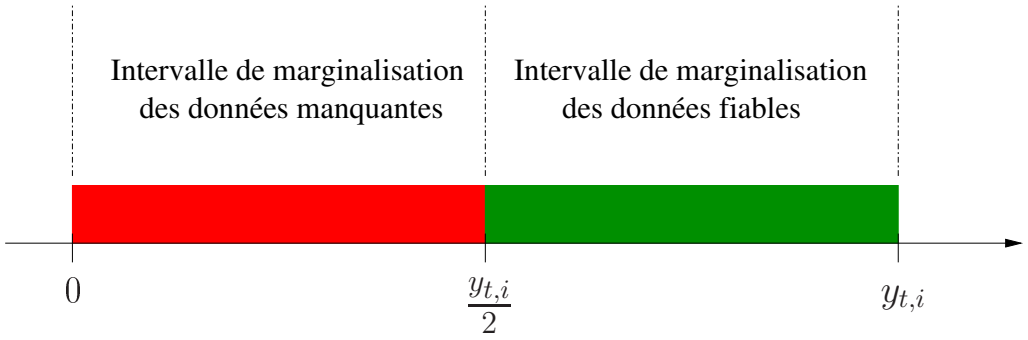


FIG. 4.5 – *Intervalles de marginalisation affinés pour des masques fondés sur le seuillage du SNR local à 0 dB.*

Les intervalles de marginalisation sont plus restreints et le taux de reconnaissance amélioré (Cf. paragraphe 2.2 page 47). Le masque idéal, dans l'optique de supprimer les effets néfastes du bruit, devrait permettre d'inférer des intervalles de marginalisation aussi petits que possible est centrés autour de l'énergie du signal de la parole seule :

$$\begin{aligned} \hat{p}(Y_i|\Theta,Q) &= \lim_{\xi\to 0}\int_{X_i-\xi}^{X_i+\xi} p(x_i|\Theta,Q).p(x_i|Y_i,m_i)dx_i \\ &= p(X_i|\Theta,Q) \end{aligned} \tag{4.7}$$

avec :

$$\int_{X_i-\xi}^{X_i+\xi} p(x_i|Y_i,m_i)dx_i = 1$$

Nous proposons une nouvelle définition de masque fondée sur la contribution du signal de la parole dans le signal observé. Cette nouvelle définition conduit à une nouvelle interprétation des masques permettant de réduire les intervalles de marginalisation. Nous développons cette approche dans le paragraphe suivant.

4.3.2 Masque de contribution

Nous décrivons ici une nouvelle caractérisation de masque permettant de réduire les intervalles de marginalisation comparativement aux intervalles dérivés des masques de données manquantes fondés sur le seuillage du SNR local.

Dans un premier temps, pour chaque trame du signal de la parole X et son homologue en condition bruitée Y, nous calculons le rapport $C_X^Y = X/Y$. Il résulte de ce calcul une représentation temps-fréquence où chaque coefficient $C_{X_i}^{Y_i}$ représente la contribution du signal de la parole X_i dans le signal bruité Y_i. Ce rapport peut être assimilé à une mesure de SNR local :

$$C_{X_i}^{Y_i} = \frac{1}{1 + 10^{-\frac{\text{SNR}_{\text{local}}(Y_i)}{20}}} \tag{4.8}$$

Dans un second temps, l'ensemble des vecteurs de contribution est partitionné en K classes $(M^k)_{k\in[1,K]}$. Chaque classe M^k est caractérisée par un vecteur moyen $\mu^k = (\mu_1^k, \mu_2^k, \ldots, \mu_i^k, \ldots, \mu_D^k)^T$ et une matrice de covariances diagonales $\Sigma^k = \text{diag}(\sigma_1^k, \sigma_2^k, \ldots, \sigma_i^k, \ldots, \sigma_D^k)$ où D est le nombre de coefficients spectraux. Nous proposons d'utiliser ces classes comme masques vectoriels de contribution. Les intervalles de marginalisation seront calculés à partir des observations bruitées et des paramètres de ces classes.

4.3.3 Masque de contribution et intervalle de marginalisation

Soient $Y = (Y_1, \ldots, Y_D)$ un vecteur d'observations de la base d'apprentissage bruitée et $X = (X_1, \ldots, X_D)$ la contribution du signal de la parole pour Y. Nous supposons que le masque vectoriel de contribution associé à Y est M^k.

Par définition :

$$p(\beta \leq C_{x_i}^{y_i} \leq \gamma | M^k) = \int_\beta^\gamma \mathcal{N}(x_i; \mu_i^k, \sigma_i^k) dx_i \tag{4.9}$$

Les masques vectoriels de contribution étant caractérisés par des distributions gaussiennes $\mathcal{N}(x_i; \mu_i^k, \sigma_i^k)$ pour chaque $C_{x_i}^{y_i}$:

$$p(\mu_i^k - 2.\sqrt{\sigma_i^k} \leq C_{x_i}^{y_i} \leq \mu_i^k + 2.\sqrt{\sigma_i^k} | M^k) = 0.95 \tag{4.10}$$

Les valeurs en dehors de cet intervalle ne représentant que 5 % des valeurs possibles de $C_{x_i}^{y_i}$, nous pouvons supposer :

$$\mu_i^k - 2.\sqrt{\sigma_i^k} \leq C_{x_i}^{y_i} \leq \mu_i^k + 2.\sqrt{\sigma_i^k} \tag{4.11}$$

Sous l'hypothèse que le masque M_k est aligné avec le vecteur d'observations Y_i, l'énergie X_i d'un coefficient bruité Y_i imputable au signal de la parole est bornée par $\mu_i^k - 2.\sigma_i^k$ et $\mu_i^k + 2.\sigma_i^k$.

Nous avons défini le terme $C_{x_i}^{y_i}$ comme étant le rapport des énergies de la parole sur l'énergie observée. Par conséquent :

$$X_i = C_{x_i}^{y_i}.Y_i \tag{4.12}$$

En combinant l'équation 4.12 avec l'inéquation 4.11, il est possible de déterminer des bornes de marginalisation pour tout coefficient Y_i du vecteur d'observations bruitées Y, en supposant que le masque associé à Y soit M^k :

$$Y_i.(\mu_i^k - 2.\sqrt{\sigma_i^k}) \leq X_i \leq Y_i.(\mu_i^k + 2.\sqrt{\sigma_i^k}) \tag{4.13}$$

En considérant un nombre raisonnable de masques vectoriels de contribution M^k, nous espérons obtenir des variances σ_i^k suffisamment faibles pour construire de petits intervalles de marginalisation et améliorer ainsi la prise en compte de l'information acoustique observée pendant le décodage.

4.3.4 Gestion des coefficients de vitesse

Le $i^{\text{ème}}$ coefficient de vitesse d'un vecteur acoustique bruité Y au temps t (noté $\Delta Y_i(t)$) est calculé sur une fenêtre temporelle comprenant $2*N+1$ coefficients statiques $Y_i(t+j)$ appartenant à la $i^{\text{ème}}$ bande de fréquences avec $j \in [-N, N]$:

$$\Delta Y_i(t) = \frac{\sum_{j=-N}^{j=N} j.Y_i(t+j)}{\sum_{j=-N}^{j=N} j^2} \tag{4.14}$$

Nous proposons dans le cadre des masques de contribution de ne pas les marginaliser durant la reconnaissance. Ceux-ci sont réestimés à partir des coefficients statiques $Y_i(t+j)$ observés et du masques M^k. Les nouveaux coefficients dynamiques $\Delta\hat{X}_i(t)$ sont alors calculés à partir des estimations $\hat{X}_i(t+j)$ faites de la contribution du signal de parole pour chacun des coefficients statiques :

$$\Delta\hat{X}_i(t) = \frac{\sum_{j=-N}^{j=N} j.\hat{X}_i(t+j)}{\sum_{j=-N}^{j=N} j^2} \tag{4.15}$$

avec :

$$\hat{X}_i(t+j) = \arg\max_x \; p(X_i(t+j) = x | M^k) \tag{4.16}$$

$$\hat{X}_i(t+j) = Y_i(t+j).\mu_i^k \tag{4.17}$$

4.4 Conclusion

L'énergie d'un signal de parole est très localisé dans le plan temps-fréquence. La présence d'un bruit additif n'affecte donc pas les différents coefficients spectraux de la même manière. Les portions hautement énergétiques du signal de parole sont moins perturbées que les régions de faible énergie. Ceci explique les similitudes observées entre l'enveloppe énergétique du signal de parole et la structure des masques dans le domaine spectral. Dans le but de restituer cette structure de masque, nous avons proposé une nouvelle architecture d'estimateur de masque. Nous proposons tout d'abord de considérer un masque au niveau du vecteur d'observations et non plus à l'échelle d'un coefficient spectrale. Nous avons montré qu'il est possible d'expliquer une grande partie des masques oracles à partir d'un petit ensemble de masques vectoriels que nous nommons masques vectoriels élémentaires. De plus, restreindre de cette manière l'espace de recherche des masques résulte en une baisse de performance acceptable en terme de taux de reconnaissance. Le nouveau modèle est donc un HMM ergodique dont chacun des états est un modèle paramétrique (plus particulièrement un GMM) d'un masque vectoriel particulier. Les probabilités de transition d'un état à un autre sont définies localement, mais leur influence est globale puisque la séquence de masques vectoriels est obtenue par l'algorithme de Viterbi. Le masque résultant est donc la séquence de masques vectoriels maximisant la vraisemblance des indices acoustiques.

Nous avons présenté au chapitre 2 différentes approches de prise en compte des masques pendant la reconnaissance. Ces algorithmes sont présentés en deux familles : les algorithmes d'imputation de données dont l'objectif est de reconstruire le signal incomplet, c'est-à-dire remplacer les coefficients dominés par le bruit (manquants) par leur estimée, et les algorithmes de marginalisation de données reposant sur une modification du calcul des vraisemblances des observations. Les différents raffinements de ces algorithmes proposés montrent que la définition du masque, et plus précisément le critère permettant d'évaluer la fiabilité des observations, jouent un rôle important. Dans le cadre de la marginalisation, nous avons proposé une nouvelle définition de masque basée sur la contribution de l'énergie du signal de parole dans l'énergie du signal observé. Nous avons montré que ces masques permettent de déterminer des intervalles de marginalisation plus fins que ceux obtenus à partir des masques fondés sur le seuillage du SNR classiquement utilisés. Réduire ces intervalles permet théoriquement d'améliorer la reconnaissance en réduisant la confusion des vraisemblances des observations.

Chapitre 5

Évaluations

« On dit souvent qu'il faut expérimenter sans idées préconçues.
Cela n'est pas possible ; non seulement ce serait rendre toute expérience stérile, mais on le voudrait qu'on ne le pourrait pas. »

- Henri Poincaré -

Sommaire

5.1 Introduction

Ce chapitre est consacré à l'évaluation des propositions faites au chapitre précédent. Nous commençons par décrire le cadre expérimental, puis nous étudions l'influence des modèles proposés sur l'estimation des masques et le taux de reconnaissance.

5.2 Cadre expérimental

5.2.1 Les bases de données

5.2.1.1 Hiwire

La base de données Hiwire contient des enregistrements de commandes aéronautiques prononcées en anglais. Ces commandes sont basées sur un vocabulaire de 133 mots. Chaque commande suit un formatage bien spécifique correspondant à des ordres ou des contrôles. La structure des commandes est donc régie par une grammaire déterministe dont la perplexité est 14.9. 331 commandes issues de cette grammaire sont utilisées pour les enregistrements.
Un total de 8099 enregistrements prononcés par 81 locuteurs et locutrices non natifs (dont la langue maternelle n'est pas l'anglais) composent cette base de données. Chaque locuteur et locutrice a prononcé 100 commandes. Le tableau 5.1 synthétise la répartition des 8099 enregistrements en fonction de la nationalité des locuteurs. Quatre sous bases sont proposées. Chacune

Langue maternelle	nombre de locuteurs	nombre d'enregistrements
Français	31	3100
Grec	20	2000
Italien	20	2000
Espagnol	10	999
Total	81	8099

Tab. 5.1 – Nombre d'enregistrements et de locuteurs par langue maternelle pour le corpus Hiwire.

d'elles est composée des 8099 enregistrements mentionnés ci-dessus corrompus par un bruit additif. L'unique bruit de cette base est enregistré dans le cockpit d'un Boeing 737 en vol. Les sous bases représentent respectivement des commandes prononcées en milieu exempt de bruit et en milieu bruité à 10 dB (LN : low noise), 5 dB (MD : mid noise) et -5 dB (HN : high noise). Nous avons créé pour nos besoins, à partir de cette base, une base d'apprentissage et une base de test. Chacune d'elles est constituée de 50 enregistrements de chaque locuteur et locutrice dans les quatre conditions acoustiques. Le tableau 5.2 présente la composition de la base Hiwire telle que nous l'utilisons.

	Hiwire							
	Apprentissage				Test			
	Français	Grec	Italien	Espagnol	Français	Grec	Italien	Espagnol
Propre	1550	1000	1000	499	1550	1000	1000	500
LN	1550	1000	1000	499	1550	1000	1000	500
MN	1550	1000	1000	499	1550	1000	1000	500
HN	1550	1000	1000	499	1550	1000	1000	500

TAB. 5.2 – Composition en nombre d'enregistrement de la base Hiwire utilisée pour nos expériences.

5.2.1.2 Aurora 2

La base Aurora 2 est construite à partir de la base de données TIDigits qui contient des enregistrements de chiffres isolés (7 chiffres au maximum par enregistrement) prononcés par des locuteurs et locutrices adultes américains. Les enregistrements originellement échantillonnés à 20 kHz sont rééchantillonnés à 8 kHz en utilisant un filtre passe-bas pour extraire le signal compris entre 0 et 4 kHz. Deux filtres additionnels sont utilisés pour simuler de manière réaliste les caractéristiques fréquentielles des équipements de télécommunication : G.172 et MIRS. Le filtre MIRS simule le comportement d'un système de télécommunication répondant aux normes GSM.

Les enregistrements rééchantillonnés et filtrés sont ensuite corrompus par différents bruits additifs à différents niveaux de SNR. Les bruits considérés sont ici :

- N1 : bruit dans le métro - *subway*
- N2 : murmures - *babble*
- N3 : bruit dans une voiture - *car*
- N4 : bruit dans un salon d'exhibition - *exhibition*
- N5 : bruit dans un restaurant - *restaurant*
- N6 : bruit dans la rue - *street*
- N7 : bruit dans un aéroport - *airport*
- N8 : bruit dans une gare - *train-station*

Bases d'apprentissage

Deux bases d'apprentissage sont disponibles : une base propre et une base bruitée. Seul le filtre G.172 est utilisé pour chacune des bases. La base propre contient 8440 enregistrements, issus de la base d'apprentissage de TIDigits, prononcés par 55 hommes et 55 femmes. Ces mêmes 8440 enregistrements sont utilisés pour la base bruitée. Ils sont partitionnés en 20 sous-ensembles de 422 enregistrements. Chaque sous-ensemble contient quelques enregistrements de chaque locuteur et locutrice. Les 20 sous-ensembles résultent de la combinaison de 4 types de bruits et 5 niveaux de SNR. Les 4 bruits sont le métro, les murmures, la voiture et la salle d'exhibition. Les niveaux de SNR sont 20 dB, 15 dB, 10 dB, 5dB et la parole seule.

Bases de test

Trois bases de test différentes, test A, test B et test C sont proposées. Les bases de test A et B sont fondées sur l'utilisation du filtre G.172 et la base de test C sur l'utilisation du filtre MIRS. 4004 enregistrements, issus de la base de test de TIDigits et prononcés par 52 femmes et 52, forment 4 sous-ensembles de 1001 enregistrements chacun. Tous les locuteurs et toutes les locutrices sont représentés dans chaque sous-ensemble. Chaque sous-ensemble est corrompu par un bruit additif à 7 niveaux de SNR : 20 dB, 15 dB, 10 dB, 5 dB, 0 dB, -5 dB et la parole seule.

Base de test A : Chaque sous-ensemble de 1001 enregistrements définis ci-dessus est corrompu par un des 4 bruits suivants : le métro, les murmures, la voiture et la salle d'exhibition et pour un des 7 niveaux de SNR définis précédemment. Ces bruits sont les mêmes que ceux utilisés pour la base d'apprentissage bruitée. Cette base est alors constituée de $4*7*1001 = 28028$ enregistrements.

Base de test B : Chaque sous-ensemble de 1001 enregistrements définis ci-dessus est corrompu par un des 4 bruits suivants : le restaurant, la rue, l'aéroport et la gare et pour un des 7 niveaux de SNR définis précédemment. Ces bruits sont différents de ceux utilisés pour la base d'apprentissage bruitée. Cette base est alors constituée de $4 * 7 * 1001 = 28028$ enregistrements.

Base de test C : Cette base contient deux sous-ensembles de 1001 enregistrements. Chacun d'eux est bruité par un des deux bruits suivants : métro et rue pour chaque niveau de SNR. 14014 enregistrements bruités composent alors cette base. Cette base de test est utilisée pour montrer l'influence de la modification des caractéristiques fréquentielles du signal sur la reconnaissance.

	Aurora 2															
	Apprentisage					Test										
	Propre (G.172)	Bruité (G.172)				Test A ((G.172)				Test B G.172)				Test C (MIRS)	
	-	N1	N2	N3	N4	N1	N2	N3	N4	N5	N6	N7	N8	N1	N6
Propre	8440	422	422	422	422	1001	1001	1001	1001	1001	1001	1001	1001	1001	1001
20 dB	-	422	422	422	422	1001	1001	1001	1001	1001	1001	1001	1001	1001	1001
15 dB	-	422	422	422	422	1001	1001	1001	1001	1001	1001	1001	1001	1001	1001
10 dB	-	422	422	422	422	1001	1001	1001	1001	1001	1001	1001	1001	1001	1001
5 dB	-	422	422	422	422	1001	1001	1001	1001	1001	1001	1001	1001	1001	1001
0 dB	-	-	-	-	-	1001	1001	1001	1001	1001	1001	1001	1001	1001	1001
-5 dB	-	-	-	-	-	1001	1001	1001	1001	1001	1001	1001	1001	1001	1001

TAB. 5.3 – Composition de la base de données Aurora 2. Le nombre d'enregistrements est fournit pour chaque condition (bruit - SNR) des bases d'apprentissage et de test.

5.2.1.3 Aurora 4

Aurora 4 est basée sur la base de données WSJ0 Wall Street Journal et s'appuie sur un vocabulaire d'environ 5000 mots. Les données de WSJ ont été enregistrées avec un microphone

Sennheiser ainsi qu'avec un second microphone en parallèle. Les différents enregistrements contiennent des extraits du Wall Street Journal lus. Les enregistrements effectués avec le second microphone sont employés pour permettre des expériences de reconnaissance avec des signaux présentant des caractéristiques fréquentielles différentes dues au canal de transmission. Pour évaluer la robustesse au bruit, 6 bruits sont artificiellement ajoutés aux enregistrements originaux. Ces 6 bruits sont les mêmes que ceux utilisés pour Aurora 2 :

- N1 : murmures - *babble*
- N2 : bruit dans une voiture - *car*
- N3 : bruit dans un restaurant - *restaurant*
- N4 : bruit dans la rue - *street*
- N5 : bruit dans un aéroport - *airport*
- N6 : bruit dans une gare - *train-station*

Une première base d'apprentissage contient 7138 phrases prononcées par 83 locuteurs et enregistrées avec le microphone Sennheiser et sans adjonction de bruit. Une seconde base d'apprentissage est également fournie. Celle-ci se compose de ces mêmes 7138 phrases, certaines exemptes de bruit d'autres bruitées par un des 6 bruits à de SNR compris entre 5 et 15 dB.

Deux bases de tests sont disponibles. Chacune d'elles contient 7 conditions d'enregistrement caractérisées par l'adjonction d'un des 6 bruits à différents SNR compris entre 5 et 15 dB, la dernière condition correspondant à des enregistrements exempts de bruit. La première base de test ne contient que des enregistrements effectués avec le microphone Sennheiser tandis que la seconde base contient des enregistrements effectués avec d'autres microphones que le Sennheiser. Pour chaque condition de bruit une base réduite de 166 enregistrements prononcés par 8 locuteurs différents est proposée.

5.2.2 Système de reconnaissance avec données manquantes ou incertaines

Toutes les expériences reportées dans ce chapitre ont été réalisées à partir de la boite à outils HTK 3.0 (HMM Toolkit). La procédure de décodage a été modifiée de manière à pouvoir marginaliser les vraisemblances des observations. La procédure de marginalisation implémentée est la Uniform-Uniform marginalisation, proposée par Morris [Morris 01a], décrite au paragraphe 2.3.3.2 du chapitre 2. Ce choix est motivé par les résultats présentés par la figure 2.2 page 47, illustrant une nette amélioration des taux de reconnaissance obtenus par cette procédure de marginalisation comparativement à la Uniform-Dirac marginalisation et à la Full marginalisation.

Nous considérerons tout coefficient spectral comme manquant si son SNR_{local} est inférieur à 0 dB. Certains travaux utilisent un seuil de SNR supérieur à 0 dB. Par exemple Seltzer et Raj

ont adopté un seuil de 2.5 dB dans le cadre de la soustraction spectrale. Augmenter ce seuil dans le cadre de la soustraction spectrale augmente la capacité à traiter des environnements non-stationnaires. Seuls les coefficients les plus énergétiques seront considérés comme fiables ce qui permet de biaiser le masque de manière à limiter la proportion de fausses acceptations. Nous définissons une fausse acceptation par le fait d'étiqueter une donnée manquante comme fiable. Par opposition, un faux rejet est le fait d'étiqueter une donnée fiable comme manquante. Cette conséquence prend encore plus d'importance dans le cadre de l'imputation de données. Puisque les valeurs des données manquantes sont reconstruites à partir des valeurs des données fiables, il est alors judicieux de biaiser l'estimateur de masque de manière à réduire le nombre de fausses acceptations même si cela implique une augmentation du nombre de faux rejets. Ceci se vérifie également dans le cas de la Uniform-Dirac marginalisation qui est la plus largement utilisée en marginalisation de données. Les vraisemblances des données fiables étant calculées classiquement il convient de réduire le nombre de fausse acceptation afin de limiter l'influence du bruit sur le décodage.
Ce choix est, a priori, moins crucial pour nous car nous marginalisons les données manquantes mais aussi les données fiables et donc il n'y a pas de raison de favoriser les fausses acceptations ou les faux rejets.

Les vraisemblances des coefficients statiques manquants et fiables sont marginalisées respectivement sur les intervalles $[0, y_{SNR-0}]$ et $[y_{SNR-0}, y]$, où y est la valeur du coefficient observé et y_{SNR-0} est une valeur telle que si $y = y_{SNR-0}$ alors le SNR_{local} du coefficient vaut 0 dB.

Un coefficient dynamique est considéré comme manquant si au moins un des coefficient statiques contribuant à son calcul est manquant. Les vraisemblances des coefficients dynamiques manquants sont marginalisées sur $[-\infty, +\infty]$, alors que les vraisemblances des coefficients dynamiques fiables sont calculés de manière classique.

5.2.3 Modèles acoustiques

Les modèles acoustiques sont tous entraînés sur les bases d'apprentissage de parole seule des corpus présentés précédemment. Les coefficients Mel-spectraux sont calculés puis compressés cubiquement. Enfin, les coefficients statiques sont étendus par leurs dérivées premières. La quasi-totalité des expériences proposées dans ce chapitre repose sur l'utilisation de 12 coefficients statiques étendus par leurs dérivées premières ce qui résulte en des vecteurs d'observations à 24 paramètres. L'utilisation d'un nombre de paramètres plus important sera précisée le cas échéant.

Une modélisation de mot est adoptée pour Aurora 2 compte tenu de la petite taille du vocabulaire considéré (12 mots). Chaque modèle de chiffre est un HMM gauche-droite à 16 états. Le modèle de silence est un HMM gauche-droite à 3 états et permettant un saut entre le premier et dernier état. Chaque état est modélisé par un mélange de 7 gaussiennes à matrice de covariance diagonale.

Une modélisation en phonème est utilisée pour les modèles Hiwire. Chaque phonème est représenté par un HMM gauche-droite à 3 états, dont chaque état est un mélange de 128 gaussiennes à matrice de covariance diagonale. Le corpus Hiwire ne fournissant pas de véritable base d'apprentissage, des modèles de phonèmes initiaux sont entraînés sur Aurora 4. Ces modèles sont ensuite réestimés sur une petite base d'apprentissage Hiwire que nous avons définie au paragraphe 5.2.1.1.

5.2.4 Paramétrisation pour les modèles de masques

De nombreux indices acoustiques permettant d'estimer les masques de données manquantes ont été présentés au chapitre 3. Cependant, aucun travail à notre connaissance n'a établi clairement quels indices étaient les plus significatifs, ni comment combiner ces différents indices. Les propositions que nous avons fait au chapitre précédent portent sur la modélisation des masques ainsi que sur la topologie de leurs modèles paramétriques. Nous proposons d'évaluer ces propositions dans les paragraphes suivants en comparant les différents estimateurs de masques que nous avons proposés à un estimateur de référence. Nous considérons donc, dans ce contexte, le choix des indices comme secondaire. L'unique contrainte que nous nous imposons est d'utiliser les mêmes indices acoustiques pour les différents estimateurs de masques.

Nous avons donc choisi d'entraîner les estimateurs de masques sur les seules observations acoustiques. Ces observations sont paramétrées dans le domaine cepstral. Certes, le domaine cepstral diffuse l'effet du bruit sur tous ces coefficients et par conséquent ne semble pas être une paramétrisation très adaptée pour modéliser les modèles de masques. Cependant, des tests préliminaires ont montré que la qualité des masques estimés est meilleure si les modèles de masques sont entraînés dans le domaine cepstral plutôt que spectral.

Les observations acoustiques pour les modèles de masques sont alors exprimées par des vecteurs composés de 13 coefficients Mel-cepstraux (dont le coefficient C_0 correspondant à l'énergie de l'observation) et 13 coefficients de vitesse. Pour chaque enregistrement ces vecteurs sont normalisés (CMN : Cepstral Mean Normalization) par le vecteur moyen calculé sur cet enregistrement.

5.3 Dépendances sur les valeurs de masque

Nous proposons d'évaluer l'influence des dépendances sur les valeurs des masques en terme de taux d'erreurs de classification ainsi qu'en terme de taux de reconnaissance en mots. Cette évaluation est effectuée sur la base de données Aurora 2. Les quatre estimateurs bayésiens de masques décrits au chapitre précédent (chapitre 4 paragraphe 4.2.4) sont entraînés sur la base d'apprentissage bruitée. Nous désignons ces estimateurs par :

1. Estimateur **AD** (Aucune Dépendance) : Cet estimateur représente notre système de référence. La fiabilité de chaque coefficient spectral est évaluée indépendamment de la fiabilité des

autres coefficients.

2. Estimateur **DT** (Dépendances Temporelles) : Des dépendances temporelles sont considérée entre les valeurs de masques des coefficients d'une même bande de fréquence, l'indépendance de valeurs de masques des coefficients résidant en des bandes de fréquences différentes est conservée.
3. Estimateur **DF** (Dépendances Fréquentielles) : Des dépendances fréquentielles sont prises en compte. La granularité de cet estimateur est moins fine que celle des estimateurs AD et DT puisque les masques sont estimés pour un vecteur d'observations entier et non plus pour un seul coefficient spectral. Les dépendances fréquentielles sont modélisées en réduisant l'espace de recherche.
4. Estimateur **DTF** (Dépendances Temporelles et Fréquentielles) : Comme l'estimateur DF, l'estimateur DTF estime des masques vectoriels. Il se distingue cependant de celui-ci par l'introduction de probabilité de transition entre les différents masques vectoriels.

Les dépendances temporelles sont des probabilités de transition calculées sur les masques oracles des enregistrements bruités de la base d'apprentissage. Les masques vectoriels élémentaires correspondant aux masques vectoriels les plus fréquents sont également déterminés à partir des masques oracles des enregistrements bruités de la base d'apprentissage. Nous avons retenu 31 masques vectoriels élémentaires couvrant 80 % des masques oracles.

5.3.1 Influence des dépendances sur les masques

Nous commençons par analyser qualitativement l'effet des dépendances sur les masques estimés sur la figure 5.1. Un signal de parole correspondant à la séquence de chiffre « one three nine oh » (Fig. 5.1(a)) est bruité à 0 dB par un signal représentant le bruit d'un métro. Le masque oracle (Fig. 5.1(e)) est calculé à partir des signaux de parole seule (Fig. 5.1(a)) et du signal de parole bruitée (Fig. 5.1(c)). La figure 5.1(g) représente ce même masque oracle mais exprimé à partir des seuls 31 masques vectoriels élémentaires.

Les masques des coefficients spectraux fournis par l'estimateur AD (Fig. 5.1(b)) semblent relativement inorganisés par rapport aux autres masques. En effet, la structure qui se dégage présente des discontinuités temporelles et fréquentielles importantes résultant de l'indépendance de ces masques vis-à-vis de leur voisinage.

Les discontinuités temporelles sont moins nombreuses pour le masque (Fig. 5.1(d)) estimé par l'estimateur DT. Les masques d'une même bande de fréquence présentent des intervalles temporels homogènes de données fiables et manquantes. Toutefois les discontinuités fréquentielles sont toujours présentes. Bien que les coefficients fiables forment des régions spectrales plus ho-

(a) - Mel spectrogramme de la séquence de chiffres : “one three nine oh”.

(b) - Pas de dépendance.

(c) - Mel spectrogramme de la séquence de chiffres : “one three nine oh” corrompue par le bruit du métro à 0 dB.

(d) - Dépendances temporelles.

(e) - Masque oracle.

(f) - Dépendances fréquentielles.

(g) - Masque oracle exprimé avec seulement les 31 masques primitifs.

(h) - Dépendances temporelles et fréquentielles.

FIG. 5.1 – *Impression visuelle de l'effet des dépendances temporelles et fréquentielles sur les masques : (a) Mel-spetrogramme d'un signal de parole représentant la suite de chiffre « one three nine oh », (c) Mel-spectrogramme de ce même signal de parole corrompu par le bruit du métro à 0 dB, (e) masque oracle calculé à partir de (a) et (c), (g) masque oracle exprimé à partir des 31 masques vectoriels élémentaires, (b) (d) (f) et (h) sont respectivement les masques estimés par les estimateurs AD, DT, DF et DTF.*

mogènes, ce masque ne reflète pas de manière fidèle l'enveloppe énergétique du signal de parole.

A l'inverse, le masque (FIG. 5.1(f)) estimé par l'estimateur DF présente une structure temporelle « saccadée », due à lindépendance des masques vectoriels, alors que les discontinuités fréquentielles ont disparu.

Enfin, le masque (FIG. 5.1(h)) obtenu par l'estimateur DTF donne la meilleure impression visuelle. Les coefficients spectraux sont partionnés en blocs homogènes et les discontinuités temporelles et fréquentielles ont totalement disparu.

5.3.2 Évaluation des masques

5.3.2.1 Erreurs de masque

Nous proposons, dans un premier temps, d'évaluer la qualité des masques estimés. Cette évaluation consiste à comparer les masques estimés avec les masques oracles qui constituent les masques de références (exempts d'erreurs). Nous définissons une acceptation comme le fait d'étiqueter un coefficient spectral comme fiable. Inversement, un rejet est l'étiquetage d'un coefficient spectral comme manquant.

Vraie acceptation : une vraie acceptation traduit le fait d'étiqueter un coefficient spectral dominé par le signal de parole comme fiable.

Fausse acceptation : une fausse acceptation est un coefficient spectral manquant étiqueté comme fiable.

Vrai rejet : un vrai rejet traduit le fait d'étiqueter un coefficient spectral dominé par le bruit comme manquant.

Faux rejet : un faux rejet est un coefficient spectral fiable étiqueté comme manquant.

5.3.2.2 Résultats

La figure 5.2 présente les taux de vraies acceptations, fausses acceptations, vrais rejets et faux rejets pour chacun des quatre estimateurs de masques sur l'ensemble des bases de test d'Aurora 2. Ces taux sont présentés pour chacune des trois bases de test.

Il est possible de déduire à partir de ces histogrammes le pourcentage de coefficients spectraux réellement dominés par le bruit en sommant les taux de vrais rejets et de fausses acceptations. En présence de bruit ce taux croît de façon linéaire proportionnellement au niveau de bruit. En moyenne, environ 50 % des coefficients sont masqués pour un niveau de bruit de 20 dB et environ 80 % pour un niveau de bruit de 0 dB. Notons toutefois qu'une partie importante (de l'ordre de 35 %) de ces coefficients réellement masqués correspond à des coefficients spectraux pour lesquels l'énergie de la parole est faible ou négligeable correspondant à des instants de silence par exemple. Ceci explique le fort taux de faux rejets pour les masques vectoriels (estimateurs DF et DTF) en absence de bruit car ces estimateurs masquent vraisemblablement les régions du spectre ne contenant pas ou très peu d'énergie de la parole même en absence de bruit.

Les erreurs de masque définies comme l'union des faux rejets et des fausses acceptations concernent 10 à 15 % des coefficients spectraux avec une légère augmentation proportionnellement au niveau du bruit. Les dépendances affectant profondément la structure des masques (Fig. 5.1), ces erreurs de masque diffèrent pour chacun des estimateurs par leur localisation dans le spectrogramme. Par conséquent, l'impact de ces erreurs sur la reconnaissance sera très probablement différent d'un estimateur de masque à un autre.

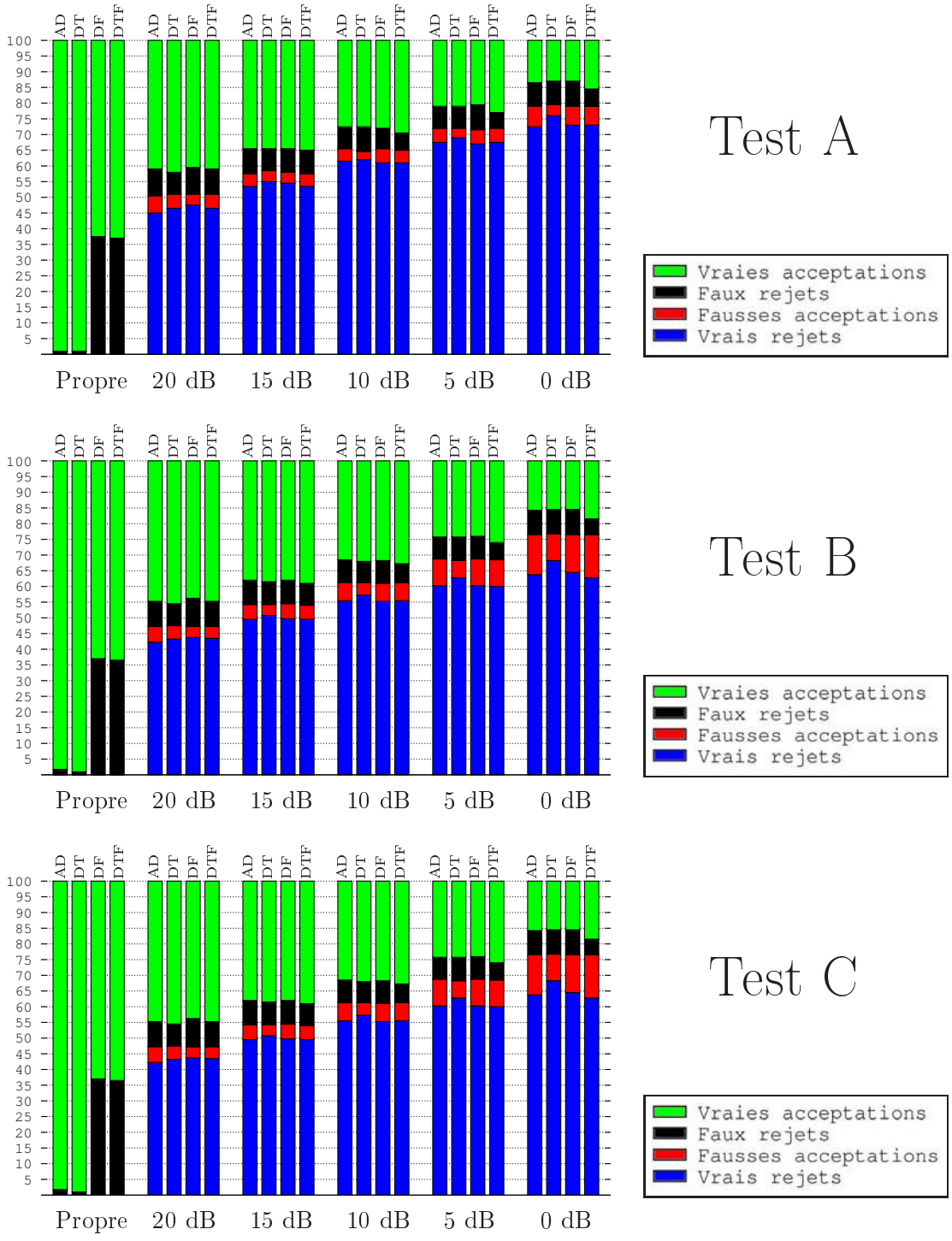


FIG. 5.2 – *Évaluation des dépendances sur les masques sur la base Aurora 2. Cette évaluation est présentée en terme de pourcentage de vraies acceptations (barres vertes), de vrais rejets (barres noires), de fausses acceptations (barres rouges) et de faux rejets (barres bleues).*

Une analyse plus fine montre que l'estimateur DTF, exploitant conjointement les dépendances temporelles et fréquentielles, fournit le meilleur taux de vraies acceptations. Une forte proportion de coefficients réellement fiables est correctement détectée. En contrepartie le taux de fausse acceptation est plus important que pour les autres estimateurs. Cet estimateur semble alors favoriser la détection des coefficients spectraux les moins bruités.

A l'inverse l'estimateur DT exploitant seulement les dépendances temporelles est l'estimateur qui fournit le taux de vraies acceptations le plus faible mais aussi celui qui minimise le taux de fausses acceptations. Cet estimateur favorise donc la détection des coefficients dominés par le bruit.

Les estimateurs de masques DT et DTF commettent de manière générale moins d'erreurs que le système de référence (estimateur AD). L'estimateur DF, quant à lui, ne se distingue que très peu de ce dernier.

Il est difficile de conclure sur la pertinence de la modélisation des dépendances sur les masques à partir de ces taux de classification. Nous proposons donc au paragraphe suivant une évaluation en terme de taux de reconnaissance.

5.3.3 Évaluation de la reconnaissance

L'objectif de la reconnaissance de la parole avec données manquantes étant d'améliorer la robustesse au bruit des systèmes de reconnaissance, nous présentons dans ce paragraphe une évaluation de l'influence des dépendances sur les masques en terme de taux de reconnaissance. Cette évaluation est également réalisée sur la base Aurora 2.

5.3.3.1 Le taux de reconnaissance

Les taux de reconnaissance sont calculés à partir de la transcription de référence (transcription exacte de ce qui a été réellement prononcé) et de l'hypothèse émise par le système de reconnaissance. Une comparaison de ces transcriptions est effectuée par un algorithme fondé sur la programmation dynamique. Cet algorithme recherche le meilleur alignement entre la transcription de référence et l'hypothèse de reconnaissance. Trois types d'erreurs sont considérés en reconnaissance, celles-ci sont illustrées par le figure Fig. 5.3.

Substitution : Une erreur de substitution correspond à un mot mal reconnu. Ceci se traduit par l'alignement de deux mots différents appartenant respectivement à la transcription de référence et à l'hypothèse de reconnaissance..

Délétion : Une délétion correspond à un mot de la transcription de référence qui n'est aligné avec aucun mot de l'hypothèse de reconnaissance.

Insertion : Une insertion est un mot de l'hypothèse de reconnaissance qui n'est aligné avec aucun mot de la transcription de référence.

Soient N, S, D, I les nombres de mots à reconnaître, de substitutions, de délétions et d'insertions, le taux de reconnaissance est défini par :

$$\text{Taux de reconnaissance} = \frac{N - S - D - I}{N} * 100$$

FIG. 5.3 – *Exemple d'erreur de substitution, de délétion et d'insertion en reconnaissance de la parole.*

5.3.3.2 Résultats

La figure 5.4 fournit les taux de reconnaissance obtenus avec chacun des quatre estimateurs de masques sur chacune des bases de test d'Aurora 2. Les taux de reconnaissance sont présentés pour différents niveaux de bruit allant de la parole seule à une corruption à 0 dB. Pour chacune des bases de test et chaque niveau de bruit, les taux de reconnaissances sont les taux moyens obtenus sur les différents bruits de la base de test.

Les dépendances fréquentielles dégradent légèrement les performances pour les conditions les moins bruitées (parole seule et 20 dB). Une particularité des estimateurs de masques vectoriels (estimateurs DF et DTF) est de masquer les régions spectrales correspondant aux silences ou pour lesquelles l'énergie du signal de parole est faible, même en absence de bruit. Ce phénomène se traduit par un taux de faux rejets important en reconnaissance de la parole seule. Il est alors très probable que certains coefficients spectraux pour lesquels l'énergie du signal de parole est forte le soient également. Ceci explique la légère baisse des taux de reconnaissance en conditions faiblement bruitées lorsque les masques de données manquantes sont estimés comme une suite de masques vectoriels.

Les dépendances temporelles améliorent significativement les taux de reconnaissance sur l'ensemble de la base de test comparativement au système de référence que constitue l'estimateur AD estimant le masque de chaque coefficient spectral indépendamment des masques des autres coefficients. Cette amélioration est d'autant plus significative en conditions fortement dégradées

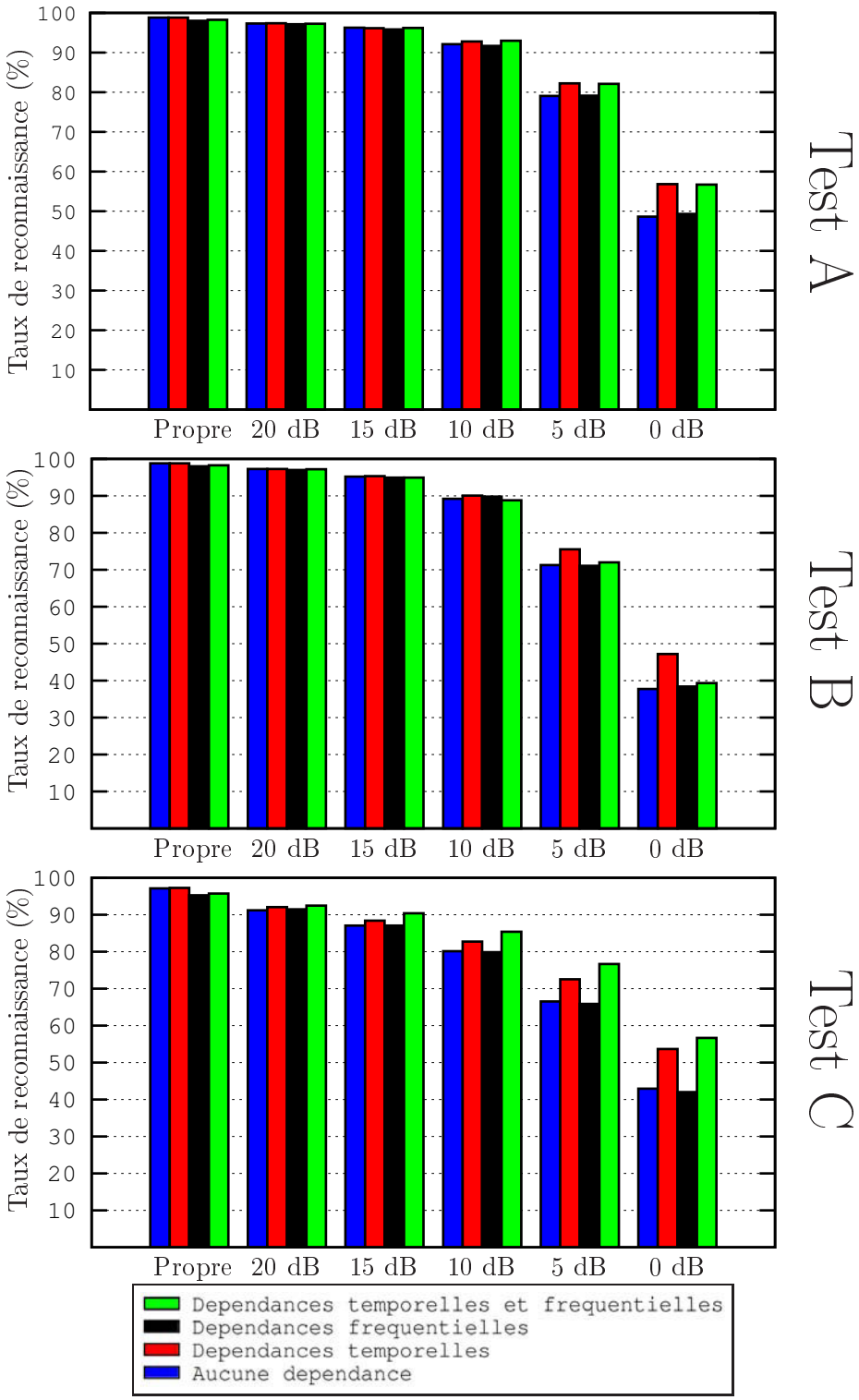


FIG. 5.4 – *Évaluation des dépendances sur les masques en taux de reconnaissance sur la base Aurora 2.*

(SNR $<$ 10 dB).

	Test A (Filtre : G.172)			
Estimateurs	Métro	Murmures	Voiture	Exhibition
AD	85.2	76.2	84.4	84.9
DT	88.1	79.9	85.6	86.8
DF	85.7	77.1	83.4	84.3
DTF	88.6	76.4	87.2	88.0

	Test B (Filtre : G.172)			
Estimateurs	Restaurant	Rue	Aéroport	Gare
AD	75.0	79.9	78.7	78.9
DT	77.4	83.0	82.4	81.5
DF	75.5	79.2	79.1	78.4
DTF	74.1	80.4	79.5	79.9

	Test C (Filtre : MIRS)	
Estimateurs	Métro	Rue
AD	74.2	73.0
DT	79.0	76.7
DF	74.7	71.8
DTF	82.1	78.5

TAB. 5.4 – *Évaluation de la contribution des dépendances sur les masques par les taux de reconnaissance obtenus sur Aurora 2. Les taux de reconnaissance sont les taux moyens obtenus pour chaque bruit des bases de test A, B et C. Les niveaux de bruit considérés sont : pas de bruit, 20 dB, 15 dB, 10 dB, 5dB et 0 dB.*

La prise en compte des seules dépendances fréquentielles ne permet pas d'améliorer la reconnaissance, au contraire, de légères baisses de performance sont observées pour certains bruits (TAB. 5.4). Cependant, une prise en compte conjointe des dépendances fréquentielles et temporelles peut s'avérer bénéfique. L'utilisation des masques estimés par l'estimateur DTF se traduit par une augmentation conséquente des taux de reconnaissance pour la base de test C. Les caractéristiques fréquentielles des enregistrements sonores de cette base de test sont différentes de celles des enregistrements de la base d'apprentissage des modèles acoustiques. Cette différence est due à l'utilisation d'un filtre MIRS pour la base de test C au lieu du filtre G.172 utilisé pour la base d'apprentissage. Les dépendances sur les masques permettent alors de modéliser correctement l'enveloppe spectrale du signal de parole. Puisque l'ensemble des masques élémentaires vectoriels sont déterminés sur le base d'apprentissage bruitée (filtre G.172) et que leurs modèles sont entraînés sur cette même base, la structure des masques estimés par l'estimateur DTF restitue

l'enveloppe spectrale d'un signal de parole dont les caractéristiques fréquentielles correspondent à celles du filtre G.172. Par conséquent, la différence entre les conditions de test (filtre MIRS) et d'apprentissage des modèles acoustiques (filtre G.172) est réduite. Ceci explique le gain considérable de performance pour la base de test C apporté par l'utilisation d'un estimateur de masque exploitant les dépendances temporelles et fréquentielles.

Les résultats obtenus sur la base de test B sont bien inférieurs à ceux obtenus sur la base de test A. À l'inverse, les bruits utilisés pour le test B sont différents de ceux rencontrés durant l'apprentissage. Nous voyons alors deux explications possibles justifiant cette baisse de performance. La première est que les masques vectoriels élémentaires ne peuvent peut-être pas se généraliser à tous les types de bruits. La seconde est que la robustesse des modèles de masques n'est pas suffisante pour estimer les masques en conditions acoustiques différentes des conditions d'entraînement de ces modèles.

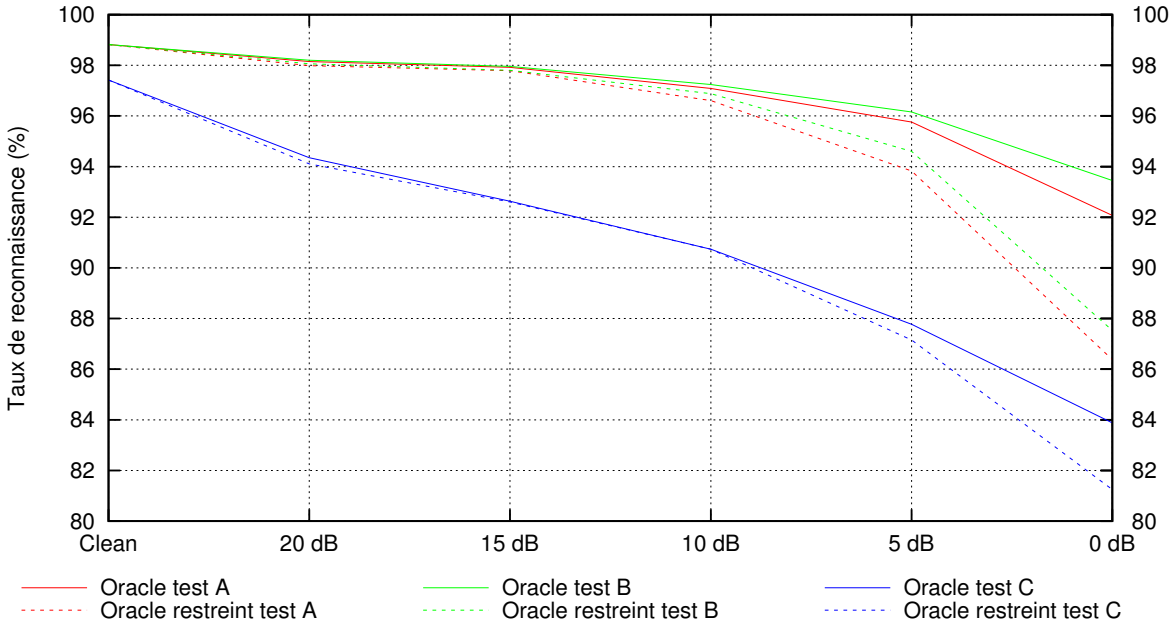


FIG. 5.5 – *Taux de reconnaissance obtenus à partir des masques oracles originaux et restreints sur Aurora 2.*

La figure 5.5 tend à rejeter la première hypothèse au profit de la seconde. Les taux de reconnaissance moyens obtenus à partir des masques oracles sur chacune des bases de test sont représentés par les courbes pleines. Les taux de reconnaissance obtenus à partir des masques oracles exprimés dans l'espace des masques restreints composé des seuls 31 masques vectoriels retenus, sont représentés par les courbes pointillées. Les couleurs de tracé correspondent aux taux de reconnaissance obtenus sur chacune des bases de test A, B et C. La réduction de l'espace des masques pénalise d'avantage la reconnaissance pour les conditions acoustiques les plus dégradées (SNR $<$ 10 dB), avec une baisse significative des performances à 0 dB. Ceci provient du fait que les masques vectoriels élémentaires sont extraits à partir de la base d'apprentissage bruitée. Les

niveaux de corruption de cette base ne vont pas en dessous de 5 dB. Par contre, la baisse de performance induite par la réduction de l'espace des masques n'est pas plus importante pour le test B que pour le test A. Les masques vectoriels retenus ne semblent pas être dépendants du bruit. Ceci renforce l'hypothèse selon laquelle les masques les plus fréquents modélisent l'enveloppe spectrale du signal de parole et ne reflètent pas ou peu l'enveloppe spectrale du bruit. Notons que la réduction de l'espace des masques affecte moins les taux de reconnaissance pour le test C comparativement aux taux obtenus avec les masques oracles. Cette expérience souligne donc une nouvelle fois la faculté des masques vectoriels à réduire les différences de caractéristiques fréquentielles entre les signaux de parole à reconnaître et ceux utilisés pour entraîner les modèles acoustiques.

La réduction de l'espace des masques ne semble pas dégrader la robustesse aux environnements acoustiques inconnus. La baisse de performance constatée sur le test B comparativement au test A (Fig. 5.4) semble alors provenir d'une mauvaise estimation des modèles de masques. Ces modèles sont apparemment très dépendants des conditions acoustiques rencontrées durant leur entraînement. Les bruits du test B sont certes différents des bruits du test A d'un point de vue perceptif, mais ceux-ci restent tout de même relativement proches. Nous proposons donc au paragraphe suivant d'évaluer l'indépendance des masques vectoriels élémentaires vis-à-vis du signal interférant dans le cadre du problème « cocktail party ».

5.3.3.3 Le problème « cocktail party »

Reconnaître un signal de parole parmi d'autres signaux de parole est une des tâches les plus ardues en reconnaissance automatique de la parole. Ce problème est connu sous la dénomination de « cocktail party ». Nous proposons dans ce paragraphe d'évaluer l'indépendance des masques vectoriels élémentaires vis-à-vis du bruit dans le cadre applicatif de la reconnaissance d'un signal de parole corrompu par un autre signal de parole. Les 31 masques vectoriels extraits de la base d'apprentissage bruitée d'Aurora 2 sont conservés.

Deux séries d'expérimentations sont présentées simultanément. La première consiste à reconnaître le signal de parole du locuteur d'intérêt à partir de masques estimés par l'estimateur DTF (dépendances temporelles et fréquentielles) en conservant les mêmes modèles de masques définis précédemment, entraînés sur la base d'apprentissage d'Aurora 2. La seconde consiste à réapprendre ces modèles de masques dans le cadre du problème « cocktail party ».

Une base de test est construite à partir des 1001 enregistrements de parole seule de la base de test A d'Aurora 2. Ces enregistrements sont d'abord triés par durée et en fonction du sexe du locuteur. 500 appariements d'enregistrements prononcés par des locuteurs de sexe différent et de durée similaires sont formés. Le $i^{\text{ème}}$ enregistrement prononcé par un locuteur est associé au $i^{\text{ème}}$ enregistrement prononcé par une locutrice. Ces derniers sont ensuite mixés dans le domaine

temporel. 10 bases de test sont ainsi construites formant deux sous-ensembles de test, 5 bases pour lesquelles le signal d'intérêt est le signal du locuteur et 5 autres pour lesquelles le signal d'intérêt est le signal de la locutrice. Le signal d'intérêt constitue le signal cible (target) et le signal interférant constitue le signal d'interférence (masker). Chacune des 5 bases propose des niveaux d'interférence TMR (Target to Masker Ratio) différents : 0, 5, 10, 15 et 20 dB. Une base d'apprentissage est également construite de la même manière à partir de la base d'apprentissage de parole seule d'Aurora 2.

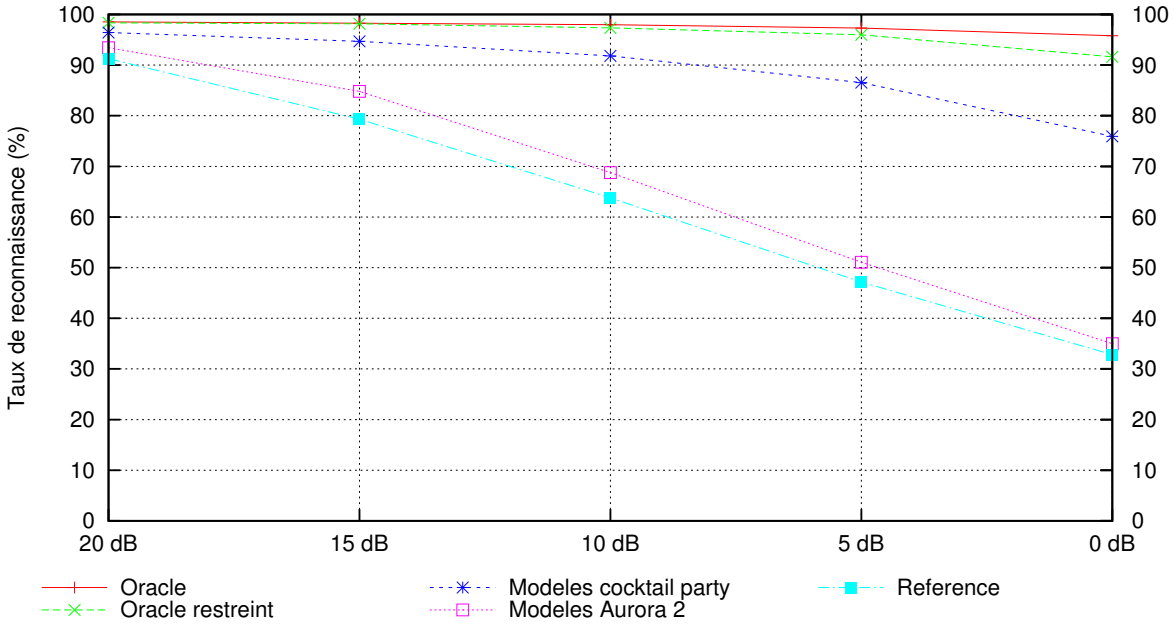


FIG. 5.6 – *Évaluation de la réduction de l'espace des masques dans le cadre du problème « cocktail party ».*

Les résultats présentés par la figure 5.6 sont les taux de reconnaissance moyens obtenus pour les différents TMR. Nous ne présentons pas les résultats en fonction du sexe du locuteur cible, ceux-ci étant similaires dans le deux cas. La faible différence entre les courbes « Oracle » et « Oracle restreint » montre que les masques oracles peuvent être approximés par les 31 masques vectoriels élémentaires extraits d'Aurora 2 sans pour autant observer une baisse très importante des taux de reconnaissance. Ceci démontre l'indépendance des masques vectoriels élémentaires vis-à-vis du signal interférant, le signal de parole étant très différent des bruits d'Aurora 2.

La courbe « Modèles Aurora 2 » représente les taux de reconnaissance obtenus à partir de masques estimés avec des modèles entraînés sur Aurora 2. Ces performances ne sont guère meilleures que les performances de « Référence ». Les taux de reconnaissance du système de référence sont ici les taux obtenus par une reconnaissance classique, c'est-à-dire hors du contexte des données manquantes. Les modèles de masques sont pris en défaut. En effet ceux-ci sont entraînés sur la base d'apprentissage bruitée d'Aurora 2 et permettent donc de discrimer un signal

de parole d'un signal interférant différent. Dans le cadre du problème « cocktail party » le bruit est également un signal de parole. L'estimateur de masque ne détecte alors aucun bruit. Par conséquent seules des petites portions du spectrogramme sont masquées. C'est pourquoi ces performances sont proches de celles du système de référence, la majorité des coefficients spectraux étant considérés comme fiables.

Par contre, les taux de reconnaissance sont bien meilleurs lorsque les modèles de masques sont entraînés sur une base de signaux de parole concurrents (« Modèles cocktail party »). Ces modèles sont spécifiques au sexe du locuteur cible et ont été réappris sur la base d'apprentissage de signaux de parole concurrents, construite à partir de la base d'apprentissage de parole seule d'Aurora 2. Nous constatons une amélioration importante des taux de reconnaissance traduisant une meilleure estimation des masques.

La baisse de performance constatée sur la base de test B d'Aurora 2 comparativement aux taux de reconnaissance obtenus sur le test A (Fig. 5.4 et Tab. 5.4) est alors imputable en grande partie au manque de robustesse de nos modèles de masques vis-à-vis d'environnements acoustiques inconnus. Nos modèles de masques ont été entraînés directement sur les observations bruitées sans aucun recours à des indices acoustiques plus évolués, tels que l'harmonicité ou encore une estimation du SNR. Ce choix est motivé par le fait que nous évaluons nos propositions comparativement à un système de référence. Le choix des indices acoustiques utilisés pour entraîner les modèles de masques est alors relayé au second plan, la seule contrainte étant d'exploiter les mêmes indices acoustiques pour les différents estimateurs de masques.

5.3.4 Conclusion

La prise en compte de dépendances temporelles et fréquentielles sur les masques des coefficients spectraux permet d'améliorer les taux de reconnaissance comparativement aux taux obtenus à partir de masques pour lesquels ces dépendances ne sont pas considérées. Les masques estimés ne sont pas vraiment meilleurs si l'on évalue la qualité de ceux-ci en terme de taux de vraies et fausses acceptations et de vrais et faux rejets. Les proportions d'erreurs de masques sont sensiblement équivalentes pour chacun des quatre estimateurs considérés. Cependant ces erreurs ne sont pas localisées identiquement dans le spectrogramme, ce qui explique les différences de performance observées en terme de taux de reconnaissance.

Les dépendances temporelles apportent un gain de performance significatif quelles que soient les conditions acoustiques. Par contre, aucune amélioration notable des taux de reconnaissance n'est observée lorsque les dépendances fréquentielles sont considérées. Toutefois, prendre en compte conjointement les deux types de dépendances permet d'améliorer la robustesse de la reconnaissance aux distorsions du signal induites par des canaux de transmission différents.

Les taux de reconnaissance résultant de l'estimation de masques vectoriels (dépendances fréquentielles) sont quelque peu décevants et inattendus. En effet, les masques exprimés comme une séquence de masques vectoriels semblent visuellement plus proches des masques oracles (Fig. 5.1). Cependant, les évaluations effectuées en terme de taux de reconnaissance vont à l'encontre de cette impression visuelle. La structure spectrale des masques exploitant les dépendances temporelles et fréquentielles semble restituer correctement l'enveloppe spectrale du signal de parole à reconnaître, cependant cette approximation reste trop grossière.

5.4 Réduction des intervalles de marginalisation

Les masques de contributions proposés au paragraphe 4.3.2 sont évalués dans la suite sur la base de données Hiwire, et sont comparés aux masques classiques basés sur un seuillage du SNR à 0 dB (masque SNR-0). Cette base ne contient qu'un seul environnement bruité. Le bruit, enregistré dans un cockpit d'avion de ligne, est quasi-stationnaire.

Nous avons construit 64 masques de contribution à partir de la base d'apprentissage bruitée Hiwire par la méthode décrite au paragraphe 4.3.2 page 83. L'estimateur de masque utilisé dans la suite est l'estimateur DTF, qui est implémenté sous la forme d'un HMM ergodique à 64 états. Chaque état est un GMM composé de 64 gaussiennes, qui modélise la distribution des vecteurs d'observations bruitées alignés avec un des 64 masques.

Ce système est comparé à l'estimateur DTF sans réduction des intervalles de marginalisation. 64 masques vectoriels élémentaires ont été construits à partir de la base d'apprentissage bruitée Hiwire.

Dans le cadre de l'évaluation des dépendances sur les masques décrite au paragraphe 5.3.2, nous avons proposé de mesurer la qualité des masques estimés en terme d'erreur de classification des coefficients spectraux. Une telle évaluation n'est pas possible ici, car nous ne faisons plus de distinction entre coefficient masqué et coefficient fiable. Nous associons à tous les coefficients un intervalle de marginalisation qui dépend de la contribution de l'énergie du signal de parole dans l'énergie du signal observé. Nous proposons une mesure pour quantifier la réduction des intervalles : l'erreur marginale aux moindres carrés.

5.4.1 L'erreur marginale aux moindres carrée : MaMSE

En imputation de données, le spectrogramme reconstruit est souvent évalué en terme d'erreur moyenne au sens des moindres carrés (MSE) comparativement au spectrogramme du signal de parole seule. Soient un spectrogramme reconstruit $\hat{X}$ pour lequel chaque coefficient est noté $\hat{x}(t,f)$ et le spectrogramme de référence X (celui de la parole seule) pour lequel chaque chaque coefficient est noté $x(t,f)$. La meilleure reconstruction $X^\star$ est celle qui minimise la déviation

moyenne des valeurs reconstruites par rapport aux valeurs de référence :

$$X^{\star} = \arg\min_{\hat{X}} \frac{1}{T.F} \sum_{t=1}^{T} \sum_{f=1}^{F} \big(x(t,f) - \hat{x}(t,f)\big)^2 \tag{5.1}$$

avec T le nombre de vecteurs d'observations et F le nombre de coefficients par vecteur.

En marginalisation de données, nous modélisons la part d'énergie de chaque coefficient spectral bruité $Y(t,f)$ pouvant être attribuée au signal de parole par une variable aléatoire $\hat{X}(t,f)$. L'intervalle de marginalisation associé à la variable aléatoire $\hat{X}(t,f)$ est donc l'ensemble de ses réalisations $\hat{x}(t,f)$ possibles. De manière générale, un intervalle de marginalisation est caractérisé par une borne inférieure $b_i(t,f)$ et une borne supérieure $b_s(t,f)$ ainsi que par une fonction de densité de probabilité $\gamma(.)$ définie sur $[b_i(t,f), b_s(t,f)]$ fournissant les probabilités *a priori* de chaque réalisation $\hat{x}(t,f) \in [b_i(t,f), b_s(t,f)]$. Nous définissons l'erreur marginale aux moindres carrés comme l'espérance de l'erreur aux moindres carrés :

$$MaMSE = \frac{1}{T.F} \sum_{t=1}^{T} \sum_{f=1}^{F} \frac{\int_{b_i(t,f)}^{b_s(t,f)} \big(x(t,f) - \hat{x}(t,f)\big)^2 \, \gamma(\hat{x}(t,f)) \, \mathrm{d}\hat{x}(t,f)}{\int_{b_i(t,f)}^{b_s(t,f)} \gamma(\hat{x}(t,f)) \, \mathrm{d}\hat{x}(t,f)} \tag{5.2}$$

Nous avons montré comment définir les bornes de marginalisation $b_i(t,f)$ et $b_s(t,f)$ à partir des masques de contribution (équation 4.13 page 84). Nous modélisons $\gamma(.)$ par une loi de probabilité uniforme $u(b_i(t,f), b_s(t,f))$ sur $[b_i(t,f), b_s(t,f)]$. De ce fait :

$$\gamma(\hat{x}(t,f)) = \frac{1}{b_s(t,f) - b_i(t,f)} \quad \forall \quad \hat{x}(t,f) \in [b_i(t,f), b_s(t,f)]$$

et

$$\int_{b_i(t,f)}^{b_s(t,f)} \gamma(\hat{x}(t,f)) \, \mathrm{d}\hat{x}(t,f) = 1$$

L'équation 5.2 peut alors être reformulée comme suit :

$$MaMSE = \frac{1}{T.F} \sum_{t=1}^{T} \sum_{f=1}^{F} \frac{\int_{b_i(t,f)}^{b_s(t,f)} \big(x(t,f) - \hat{x}(t,f)\big)^2 \, \mathrm{d}\hat{x}(t,f)}{b_s(t,f) - b_i(t,f)} \tag{5.3}$$

Nous proposons donc d'évaluer quantitativement la réduction des intervalles de marginalisation résultant de l'approche proposée au paragraphe 4.3.2 par cette mesure MaMSE. Toutefois, et afin d'évaluer l'importance que l'on peut accorder à cette mesure, nous commençons par étudier dans la suite les limites de cette mesure quant à ce qu'elle permet de conclure.

5.4.2 Interprétation de la mesure MaMSE

La figure 5.7 illustre le comportement de la mesure MaMSE sur un exemple fictif. Supposons que l'on souhaite classifier une observation bruitée Y parmi un ensemble de classe $\mathcal{C} = \{C_1, \ldots, C_k\}$. Nous entendons par valeur de référence, la valeur X qui aurait due être

observée en l'absence du bruit N : $Y = X + N$. Nous supposons $X = 0$ dans cet exemple. Considérons maintenant que nous disposons d'un intervalle de marginalisation fournissant l'ensemble des valeurs pouvant être attribuées à X. L'intervalle de marginalisation est caractérisé par sa longueur (borne supérieure - borne inférieure) et par la distance de sa valeur médiane à la valeur de référence, que nous appelons décalage.

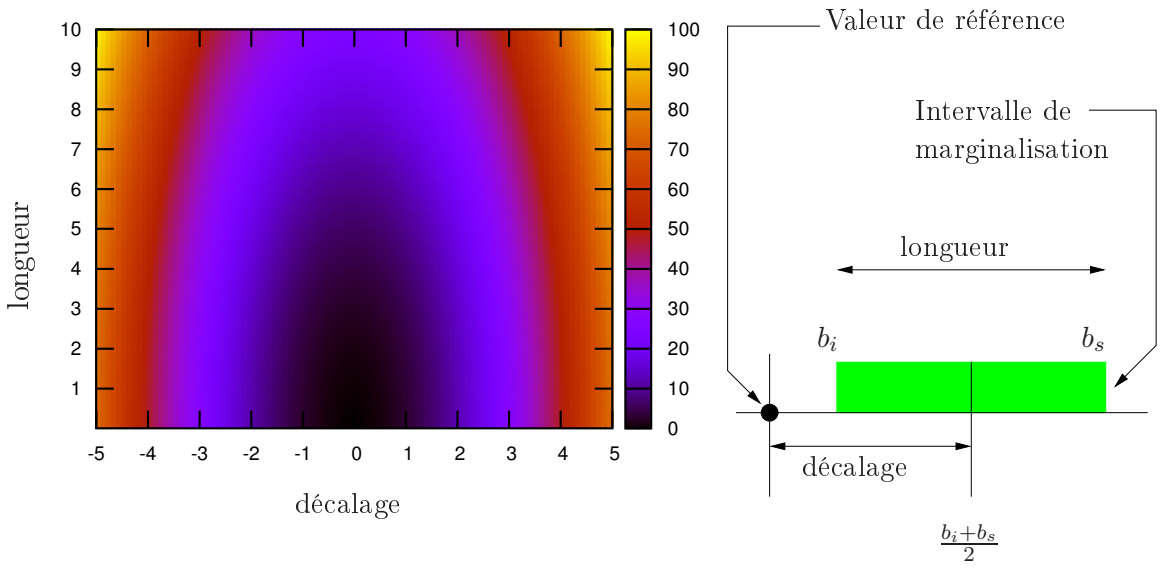


FIG. 5.7 – *Évaluation de la mesure MaMSE sur un exemple fictif. La figure de gauche fournit les scores MaMSE pour une longueur d'intervalle de marginalisation (longueur) et une distance entre la valeur de référence et le centre de l'intervalle (décalage). La figure de gauche illustre les notions de longueur et décalage.*

La figure 5.7 montre que la mesure MaMSE est notamment représentative de la longueur de l'intervalle de marginalisation. Ainsi, pour un décalage d fixé, le score de la mesure MaMSE est toujours plus important pour un intervalle de longueur l_1 que pour un intervalle de longueur $l_2 < l_1$.

$$l_1 > l_2 \rightarrow MaMSE\Big(Intervalle(d, l_1)\Big) > MaMSE\Big(Intervalle(d, l_2)\Big) \tag{5.4}$$

où $Intervalle(d, l)$ est un intervalle de marginalisation de longueur l dont la valeur médiane est à une distance de d par rapport à la valeur de référence.

La mesure MaMSE est également dépendante du décalage, et elle favorise ainsi un intervalle contenant la valeur de référence par rapport à un intervalle de même taille ne contenant pas cette valeur. Cependant, cette mesure réduit la notion de coût d'un intervalle proposé à l'erreur moyenne par rapport à la valeur de référence, ce qui n'est pas forcément représentatif du véritable coût tel que nous l'imaginons, et qui accorde par exemple plus d'importance au fait de contenir la vrai valeur qu'à la largeur de l'intervalle. Des effets de bord non désirables peuvent

donc apparaître, comme par exemple de préférer un petit intervalle décentré par rapport à un intervalle plus grand mais contenant la valeur non bruitée.

Finalement, la mesure MaMSE permet effectivement d'avoir une idée assez précise de l'erreur moyenne produite par les différents intervalles de marginalisation, mais il ne faut pas s'appuyer uniquement sur cette mesure pour évaluer les techniques proposées, c'est pourquoi nous complétons cette étude par une comparaison des taux de reconnaissance.

5.4.3 Résultats

5.4.3.1 Score MaMSE

Le tableau 5.5 présente le score MaMSE moyen pour chaque condition de test proposée par la base Hiwire ainsi que le taux d'erreur de masque, qui est calculé simplement en divisant le nombre de masques ne contenant pas la valeur de référence par le nombre de masques total.

		Masques SNR-0		Masques de contribution	
		MaMSE	Taux d'erreur (%)	MaMSE	Taux d'erreur (%)
Oracles	Clean	10	0	0.02	0
	LN	64	1	11	3
	MN	101	1	15	3
	HN	269	1	29	5
Estimés	Clean	46	1	3	1
	LN	75	6	18	19
	MN	113	7	28	23
	HN	288	5	75	27

TAB. 5.5 – *Évaluation de la réduction des intervalles de marginalisation sur la base Hiwire par la mesure MaMSE.*

Les scores MaMSE calculés sur les masques de contribution sont bien inférieurs à ceux calculés sur les masques SNR-0, confirmant que les intervalles de marginalisation, qui sont définis à partir des masques, sont effectivement réduits. Cette réduction est observée non seulement pour les masques oracles, mais également pour les masques estimés.

En condition de test, les masques de contribution permettent de réduire le score MaMSE d'environ 75 % par rapport au score MaMSE calculés sur les masques SNR-0. Cette baisse du score MaMSE confirme la réduction des intervalles. Nous illustrons sur les figures 5.8 et 5.9 les intervalles de marginalisation dans chacune des conditions de test de la base Hiwire.

Les figures 5.8 et 5.9 présentent différentes enveloppes spectrales d'un même signal. Ce signal

correspond à la prononciation du mot « standby » dans chacune des quatre conditions de test (Clean : parole seule, LN : low noise, MN : middle noise et HN : high noise) proposées par la base Hiwire. Les enveloppes spectrales illustrent l'évolution de l'énergie du signal pour la $6^{\text{ième}}$ bande de fréquence sur 12 considérés. La courbe bleue représente le signal de parole seule (les valeurs de référence), la courbe rouge représente le signal observé et l'aire verte représente les intervalles de marginalisation. Les intervalles représentés sur la figure 5.8 sont déterminés à partir des masques de contribution tandis que les intervalles représentés sur la figure 5.9 sont déterminés à partir des masques SNR-0. Dans les deux cas, le masque est estimé, ce qui explique les différentes erreurs de classification conduisant à des intervalles de marginalisation ne contenant pas la valeur de référence.

Les intervalles de marginalisation dérivés des masques de contribution sont effectivement beaucoup moins larges que les intervalles dérivés des masques SNR-0. Ce phénomène est très prononcé pour les coefficients les plus bruités. Ces coefficients correspondent le plus souvent aux temps de silence caractérisés par une faible contribution énergétique du signal de parole.

Les intervalles de marginalisation associés aux coefficients pour lesquels l'énergie de la parole est forte ne profitent pas, ou dans une moindre mesure, de cette amélioration. Ces intervalles peuvent même être plus importants que ceux dérivés des masques SNR-0, comme nous pouvons le constater sur les tracés correspondant aux conditions de test les plus bruitées (HN).

De manière générale, les masques de contribution permettent une meilleure approximation de l'enveloppe énergétique du signal de parole, notamment pour les portions du spectre les moins énergétiques.

Toutefois, les taux d'erreurs reportés dans le tableau 5.5 montrent que les masques ne contenant pas la véritable contribution de la parole sont bien plus fréquents lorsqu'ils sont réduits que lorsqu'ils ne le sont pas. Cet effet peut contrebalancer le gain obtenu en réduisant les masques, et il est donc difficile de conclure dès à présent quant aux performances de cette approche, performances qui sont donc évaluées par la suite en terme de taux de reconnaissance.

5.4.3.2 Taux de reconnaissance

La figure 5.10 fournit les taux de reconnaissance obtenus sur la base Hiwire à partir des masques oracles. Les résultats reportés confirment que les taux de reconnaissance sont accrus en réduisant les intervalles de marginalisation. Les taux obtenus à partir des masques de contribution sont très nettement supérieurs aux taux obtenus à partir des masques SNR-0. Ce n'est cependant pas pour la parole seule. Une explication probable est qu'en absence de bruit les intervalles de marginalisation obtenus à partir des masques de contribution sont trop fins, au point qu'ils sont difficilement discernables sur la figure 5.8. Une particularité de la base Hiwire par rapport à une

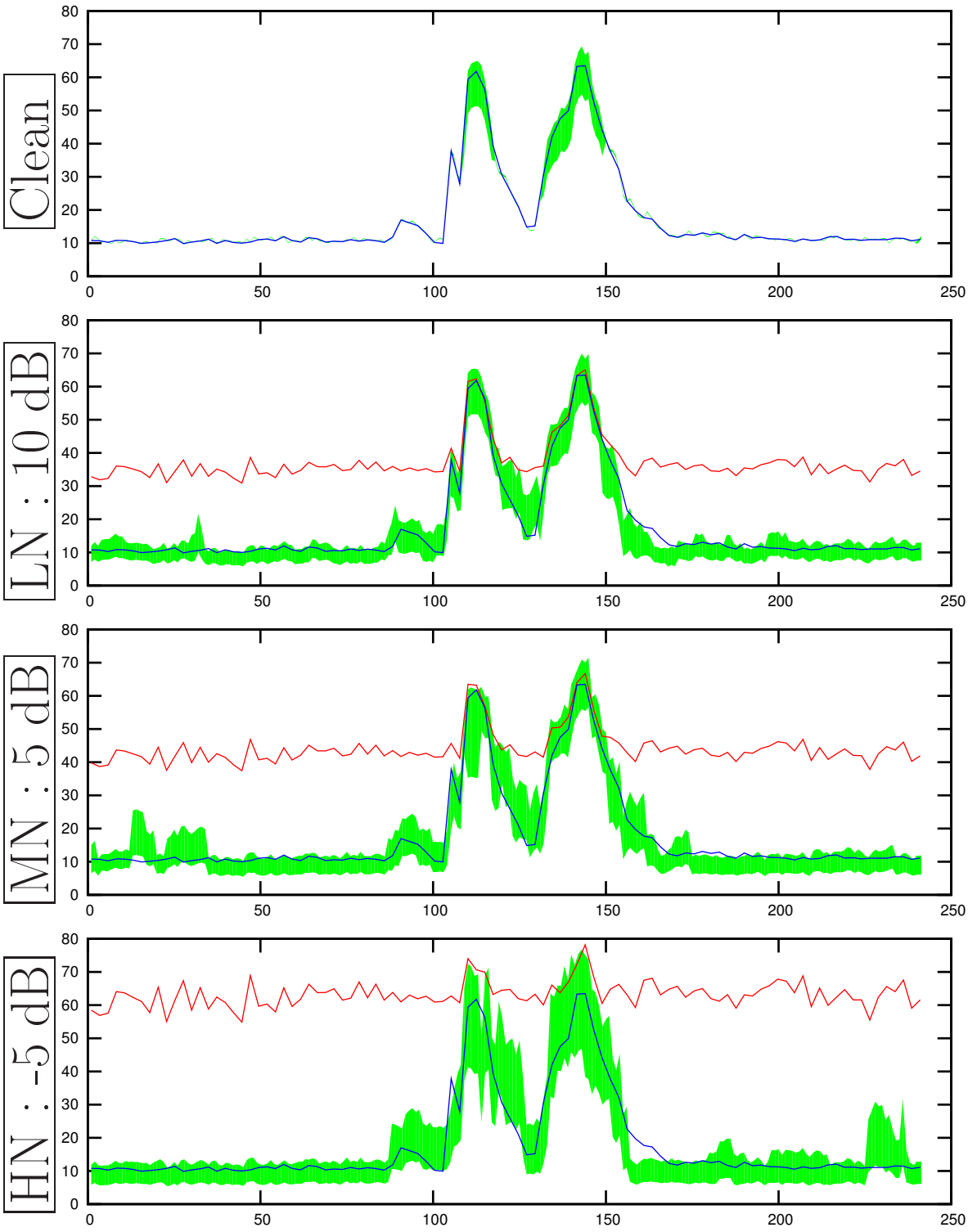


FIG. 5.8 – *Intervalles de marginalisation dérivés des masques de contributions. Les courbes bleue et rouge représentent respectivement l'évolution énergétique des signaux de parole seule et de parole bruitée pour la* $6^{ième}$ *bande de fréquence. Les intervalles de marginalisation sont représentés par les aires vertes.*

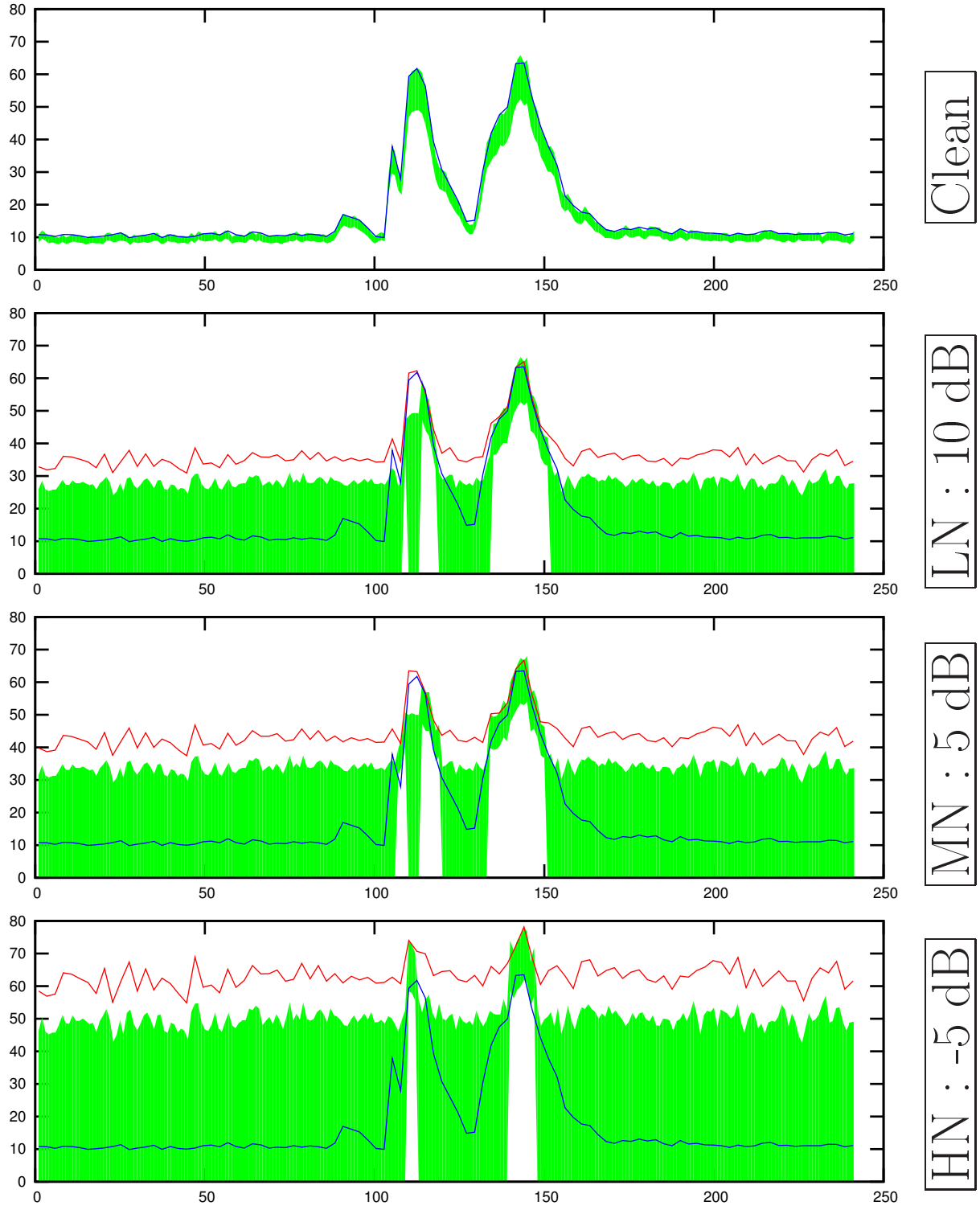


FIG. 5.9 – *Intervalles de marginalisation dérivés des masques SNR-0. Les courbes bleue et rouge représentent respectivement l'évolution énergétique des signaux de parole seule et de parole bruitée pour la* $6^{ième}$ *bande de fréquence. Les intervalles de marginalisation sont représentés par les aires vertes.*

base telle Aurora 2, est que les locuteurs ne sont pas des anglophones natifs. Il existe donc une très grande variabilité inter-locuteur du signal. Des intervalles de marginalisation plus importants semblent alors réduire l'influence de cette variabilité.

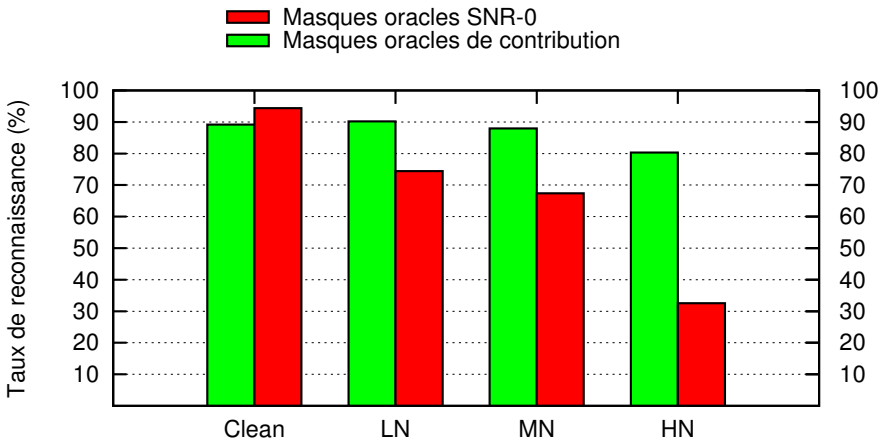


FIG. 5.10 – *Comparaison des taux de reconnaissance obtenus à partir des masques oracles de contribution et SNR-0 sur la base de test Hiwire.*

Les taux de reconnaissance obtenus à partir de masques estimés sont meilleurs pour les masques de contribution dans toutes les conditions (FIG. 5.11). Cependant nous observons une baisse significative de ceux-ci par rapport aux taux de référence obtenus avec les masques oracles, alors que cette baisse est moins importante pour les masques SNR-0.

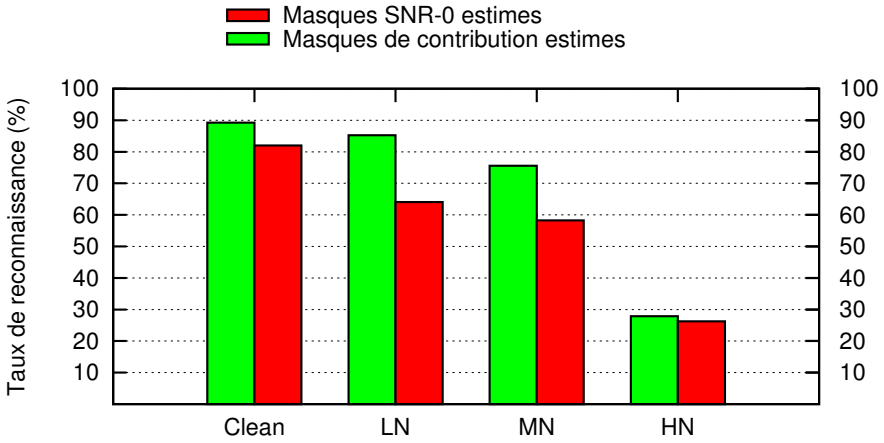


FIG. 5.11 – *Comparaison des taux de reconnaissance obtenus à partir des masques de contribution et SNR-0 estimés sur la base de test Hiwire.*

Cette baisse est très probablement une conséquence directe des taux d'erreurs importants reportés au tableau 5.5. Les intervalles de marginalisation sont certes réduits, mais une forte proportion de ceux-ci n'englobe pas les valeurs de référence. Cette baisse de performance illustre la difficulté d'estimer les masques de contribution comparativement aux masques SNR-0. Ces masques permettent d'améliorer significativement les taux de reconnaissance, cependant ce gain n'est possible qu'au prix d'une meilleure estimation des masques.

5.4.4 Conclusion

La définition même d'un masque joue un rôle important en reconnaissance de la parole avec données manquantes. Nous avons montré dans le cadre de la marginalisation de donnée, que l'algorithme de reconnaissance sur des observations partielles dépend de la manière dont la fiabilité des observations est évaluée. Les systèmes actuels, pour la plupart, sont fondés sur un seuillage du SNR. Dans cette optique nous avons considéré un coefficient spectral comme manquant si son SNR local est inférieur à 0 dB. Les travaux de Morris ont montré qu'il est possible de calculer à partir de tels masques des intervalles de marginalisation pour chaque coefficient spectral, que ce dernier soit peu ou très affecté par le bruit. Nous avons proposé une nouvelle définition de masque de données manquantes fondée sur la contribution de l'énergie du signal de parole dans l'énergie du signal observé. De tels masques permettent de calculer des intervalles de marginalisation plus fins. Cette approche est validée sur la base Hiwire caractérisée par un bruit de cockpit d'avion de ligne dont les caractéristiques sont proches d'un bruit blanc gaussien. Les expérimentations que nous avons reportées montrent que les intervalles de marginalisation sont significativement réduits conduisant à de meilleurs taux de reconnaissance. Cependant le fort taux d'erreur de masque révèle que ces masques sont plus difficiles à estimer que les masques SNR-0. Un estimateur de masque plus robuste que celui que nous avons utilisé semble alors nécessaire pour exploiter pleinement ce nouveau potentiel qu'offrent les masques de contribution.

Chapitre 6

Conclusion générale

« L'expérience de chacun est le trésor de tous. »

- Gérard de Nerval -

Sommaire

6.1 Cadre de notre étude

La parole est la modalité de communication la plus naturelle et la plus riche pour l'Homme. La reconnaissance automatique de la parole constitue donc aujourd'hui un acteur privilégié dans le développement des interfaces homme-machine de dernière génération. Certes cette technologie est de plus en plus utilisée dans des systèmes d'assistance (téléphonie mobile, assistant de navigation, systèmes d'exploitation, etc) mais son déploiement à grande échelle reste tout de même limité. Le manque de robustesse au bruit de ces systèmes constitue un frein important à ce déploiement. A cet égard, beaucoup de travaux concernent la robustesse au bruit additif. C'est en effet le type de perturbation le plus courant mais qui malheureusement se révèle être extrêmement gênant pour les systèmes automatiques.

Les systèmes de reconnaissance actuels reposent, pour la plupart, sur des modèles statistiques et plus particulièrement sur les modèles de Markov cachés. Ces modèles sont entraînés sur des bases d'apprentissage de parole, enregistrées en condition maîtrisée, c'est-à-dire qu'aucune autre source sonore ne vient corrompre le signal de parole. Par conséquent, ils ne sont plus adaptés pour décoder un signal acoustique complexe résultant de l'interférence de signaux issus de différentes sources. Un autre mode d'apprentissage est possible. Il consiste à entraîner les modèles acoustiques sur des bases bruitées. Cependant l'immense variabilité des bruits possibles ne permet pas de construire *a priori* des modèles adaptés à tous ces bruits sans augmenter excessivement la confusion des modèles. Il existe un grand nombre d'algorithmes permettant de traiter le problème de la robustesse au bruit additif, intervenant à différents étages du système de reconnaissance. Certains sont dédiés au débruitage du signal, d'autres à l'extraction de paramètres robustes ou encore à l'adaptation des modèles acoustiques aux conditions d'utilisation. Nous avons proposé au premier chapitre une brève introduction à la reconnaissance robuste de la parole.

La théorie des données manquantes constitue une sérieuse alternative à ces algorithmes. Elle s'appuie sur des travaux montrant que, lorsque l'environnement acoustique est bruité, certaines zones du spectre de puissance du signal à décoder sont effectivement dominées par le signal de parole, et notamment les régions très énergétiques comme les formants, mais que de nombreuses autres zones sont dominées par le bruit. Un coefficient spectral dominé par le bruit est considéré comme un coefficient manquant car il représente une information erronée ne traduisant pas les caractéristiques du signal de parole. Une fois les masques calculés, le moteur de reconnaissance peut alors mettre en œuvre des stratégies de décodage adaptées à la fiabilité des observations. Le formalisme mathématique de la reconnaissance de la parole avec données manquantes a été détaillé au second chapitre. Nous avons présenté deux familles d'algorithmes permettant de prendre en compte les masques. Les techniques de marginalisation qui, lors du calcul de la vraisemblance que le signal observé ait été généré par un modèle acoustique donné, marginalisent cette vraisemblance pour les paramètres masqués. Les techniques d'imputation estiment la contribution énergétique de la parole seule pour les parmètres masqués. Lorsque cette contribution est estimée

uniquement à partir du signal acoustique et de ses caractéristiques, on parle d'imputation de données. Lorsque l'alignement avec les modèles de parole est également considéré on parle alors d'imputation conditionnée par les classes acoustiques. Il est possible d'affiner ces approches en limitant le domaine de marginalisation ou en contraignant la reconstruction du signal par les contributions maximales et minimales des paramètres de parole masqués. Les différentes variations de ces algorithmes montrent que la définition d'un masque et son interprétation conditionnent la stratégie de décodage mise en œuvre.

Nous avons effectué une évaluation de trois algorithmes de marginalisation. Les résultats ont mis en évidence le rôle de premier plan que jouent les masques dans ce formalisme. En effet, de très bonnes performances sont atteintes lorsque les masques sont connus *a priori* (masques oracles), mais celles-ci se dégradent fortement si ces masques sont entachés d'erreurs. L'estimation de masque est donc un problème délicat faisant l'objet de cette thèse. Les nombreux travaux, que nous avons présentés au chapitre 3, concernent cet axe de recherche. Nos contributions se placent dans le contexte de l'estimation bayésienne des masques et plus précisément sur leur modélisation dans ce formalisme. Celles-ci ont été présentées au chapitre 4 et évaluées au chapitre 5. Nous les résumons dans les paragraphes suivants.

6.2 Contributions

6.2.1 Dépendances sur les masques

Après avoir montré la similitude entre la structure des masques de données manquantes et l'enveloppe énergétique de la parole dans le domaine spectral, nous avons proposé de nouveaux modèles de masques dans le but de restituer cette structure. Nous avons considéré pour cela des dépendances temporelles et fréquentielles entre les valeurs des masques des coefficients spectraux. Les dépendances temporelles sont modélisées sous la forme de probabilités de transition entre les différentes valeurs de masques. Les dépendances fréquentielles sont introduites en considérant le masquage non plus à l'échelle du coefficient mais au niveau du vecteur d'observations. Dans cette optique, nous avons défini des masques vectoriels élémentaires comme un ensemble restreint de masques vectoriels couvrant au mieux les masques oracles tout en minimisant la dégradation du taux de reconnaissance induite par la réduction de l'espace des masques. De cette façon nous avons montré qu'il est possible de réduire considérablement l'espace de recherche des masques vectoriels tout en pénalisant que très légèrement les performances. Cette réduction de l'espace de recherche est nécessaire puisque le nombre de masques vectoriels éligibles durant l'estimation est exponentiel par rapport à la dimension des vecteurs d'observations.

Nous avons proposé trois estimateurs de masques : le premier exploite les dépendances temporelles, le second les dépendances fréquentielles et le troisième les dépendances temporelles et fréquentielles. Ces estimateurs ont été évalués, dans le cadre de la marginalisation, comparative-

ment à un estimateur de référence, identique à celui proposé par Raj Ramakrishnan et Seltzer qui estime le masque de chaque coefficient spectral indépendamment des masques des coefficients de son voisinage. Cette analyse est effectuée sur la base de données Aurora 2.

Il ressort de cette analyse que les dépendances influent fortement sur la structure des masques. Visuellement, les masques exploitant les deux types de dépendance semblent les plus proches des masques oracles. Une analyse plus fine montre que les dépendances temporelles permettent de réduire le taux de fausses acceptations (proportion de coefficients faussement évalués comme fiables) et que la combinaison des deux types de dépendance permet d'améliorer le taux de vraies acceptations (proportion de coefficients correctement identifiés comme fiables) comparativement à l'estimateur de référence. De manière générale les dépendances permettent de réduire les erreurs de masque à l'exception cependant des dépendances fréquentielles seules.

Du point de vue de la reconnaissance, les meilleurs résultats sont obtenus à partir des seules dépendances temporelles. En moyenne, une amélioration relative de 15 % est observée par rapport aux taux de reconnaissance obtenus à partir de l'estimateur de référence sur l'ensemble des bases de test d'Aurora 2. Ceci souligne l'importance de prendre en compte le contexte temporel d'un masque. A cet égard, le HMM constitue un modèle de masque bien adapté de part sa capacité à modéliser des processus à évolution temporelle. Considérées individuellement, les dépendances fréquentielles n'ont pas permis d'améliorer les taux de reconnaissance comparativement au système de référence. Une explication possible de ce résultat quelque peu décevant est que le gain que nous espérions d'une telle modélisation de masque n'est pas suffisant pour compenser la baisse des taux de reconnaissance induite par la réduction de l'espace des masques vectoriels. Les masques vectoriels ne sont pas pour autant dénués d'intérêt. D'une part, la combinaison des dépendances temporelles et fréquentielles améliore considérablement la robustesse aux distorsions du signal induites par le canal de transmission, comme en attestent les bons résultats obtenus sur la base de test C dédiée à cette tâche. D'autre part, les masques vectoriels peuvent prendre place avantageusement dans un système de reconnaissance tel le décodeur multi-sources de Barker. Nous développons ce point dans les perspectives.

6.2.2 Une nouvelle définition de masques

Les différentes variantes des techniques d'imputation et de marginalisation proposées dans la littérature ont montré que la définition d'un masque et son interprétation conditionnent la stratégie de décodage mise en œuvre. Plus particulièrement, dans le cadre de la marginalisation, ces variantes se traduisent par une réduction du domaine de marginalisation de la vraisemblance des coefficients spectraux. La technique de marginalisation la plus utilisée, marginalisation Uniform-Dirac, marginalise la vraisemblance d'un coefficient spectral masqué sur $[0, y]$, où y représente l'énergie observée de ce coefficient. La vraisemblance d'un coefficient fiable est calculée de manière classique sous l'hypothèse de dominance postulant qu'un coefficient spectral est

dominé soit par l'énergie de la parole, soit par l'énergie du bruit. Morris a montré qu'une autre interprétation des masques de données manquantes fondés sur le seuillage du SNR est possible. Il propose la technique de marginalisation Uniform-Uniform. Cette technique se différencie des autres par le fait que toutes les vraisemblances sont marginalisées, que ce soit pour les données masquées ou les données fiables. Les intervalles de marginalisation diffèrent cependant au regard de la fiabilité des coefficients. Si l'on considère des masques fondés sur un seuillage du SNR à 0 dB, la vraisemblance d'un coefficient masqué est marginalisée sur l'intervalle $[0, y/2]$, et la vraisemblance d'un coefficient fiable l'est sur $[y/2, y]$. Dans un certain sens cette technique remet en cause l'hypothèse de dominance. On peut considérer qu'un coefficient spectral bruité n'est ni totalement manquant ni totalement fiable. L'évaluation comparative de trois techniques de marginalisation (Full-marginalisation, Uniform-Dirac marginalisation et Uniform-Uniform marginalisation) montre l'intérêt de réduire les domaines de marginalisation. De plus, elle souligne l'importance de la prise en compte de la définition du masque lors de la mise en œuvre de l'algorithme de décodage.

Nous avons proposé une nouvelle définition de masques de données manquantes. Ces masques sont fondés sur la contribution de l'énergie du signal de parole dans l'énergie du signal observé. Cette nouvelle définition conduit à une nouvelle interprétation des masques permettant de réduire les intervalles de marginalisation. Les masques de contributions sont construits à partir d'une représentation temps-fréquence dont chaque valeur est la contribution de l'énergie du signal de parole dans l'énergie du signal observé un instant t pour une bande de fréquences centrée sur f. L'espace des vecteurs de contributions est ensuite partitionné en K classes $(M^k)_{k\in[1,K]}$. Chacune d'elles est caractérisée par un vecteur moyen μ_k ainsi qu'une matrice de covariance diagonale Σ_k. Nous considérons alors chaque classe comme un masque de contribution possible. Nous avons montré que les intervalles de marginalisation peuvent être déterminés à partir des paramètres μ_k et Σ_k de ces masques.

Nous avons proposé un estimateur de masques de contributions. Par définition ces masques sont vectoriels. Nous avons alors utilisé la même architecture que l'estimateur DTF exploitant les dépendances temporelles entre les masques vectoriels successifs. Nous avons évalué comparativement les masques de contribution par rapport aux masques fondés sur le seuillage du SNR classiquement utilisés sur la base de données Hiwire. Dans ce contexte nous évaluons la qualité des masques par la réduction des intervalles de marginalisation mais aussi par les erreurs de masque. Nous entendons par erreur de masque le fait de déterminer un intervalle de marginalisation ne contenant pas l'énergie du signal de parole (valeur de référence). Nous avons proposé dans cette optique une nouvelle mesure : l'erreur marginale aux moindres carrés (MaMSE). Le score MaMSE est d'autant plus faible que les domaines de marginalisation sont centrés sur la valeur de référence et de petite taille. Cette évaluation montre que les intervalles de marginalisation déduits des masques de contribution sont fortement réduits. Des effets de bord peuvent cependant apparaître, comme par exemple de préférer un petit intervalle décentré par rapport à un intervalle plus grand mais contenant la valeur de référence. Nous avons donc complété cette

évaluation par une comparaison des taux de reconnaissance. Les masques oracles de contribution permettent d'obtenir un taux de reconnaissance moyen, sur l'ensemble de la base de test, de 87% alors qu'un taux de 68% est obtenu avec les masques oracles SNR-0. Ces taux constituent les seuils de performance atteignables par ces deux définitions de masques. Ils attestent donc le fort potentiel des masques de contribution. Lorsque les masques sont estimés, l'écart de performance diminue. Les taux de reconnaissance observés sont de 70% pour les masques de contribution et de 57% pour les masques SNR-0 avec des performances à peu près équivalentes pour les conditions les plus bruitées. Il ressort de cette évaluation que les erreurs de masque sont beaucoup plus fréquentes pour les masques de contributions ce qui pénalise fortement la reconnaissance. Il apparaît donc que les masques de contribution permettent effectivement d'améliorer les taux de reconnaissance, cependant une estimation de masque plus robuste est nécessaire pour exploiter pleinement leur potentiel.

Cette évaluation a été réalisée sur la base de données Hiwire. Celle-ci ne comporte qu'un seul bruit dont les caractéristiques sont proches de celles d'un bruit blanc gaussien. Des expériences complémentaires sont alors nécessaires afin d'évaluer ces nouveaux masques dans des conditions plus pénalisantes qu'un bruit blanc gaussien.

6.3 Perspectives

Les modèles de masques que nous avons évalués dans cette thèse sont entraînés directement sur les observations bruitées. Nous avons relégué le problème de la paramétrisation du signal pour les modèles de masques au second plan puisque ces modèles sont évalués comparativement à un modèle de référence. Par conséquent, la seule contrainte que nous nous sommes imposée est que cette paramétrisation soit la même pour tous les modèles de façon à pouvoir comparer les résultats. Une telle approche est certes simpliste mais elle permet une mise en œuvre simple et rapide des expérimentations. Il est à présent souhaitable d'utiliser des indices plus pertinents, tels que ceux mentionnés au chapitre 3. Nous y voyons un avantage double. D'une part les résultats obtenus à partir de tels indices pourront être comparés aux résultats reportés dans différents travaux du domaine. D'autre part, des indices acoustiques bien choisis peuvent en un certain sens permettre de réduire la dépendance des modèles de masques aux environnements acoustiques rencontrés durant leur apprentissage. Dans cette optique les paramètres proposés par Seltzer [Seltzer 00] constituent une base intéressante.

Une condition forte à la mise en œuvre des masques vectoriels SNR-0 concerne la dimension des vecteurs d'observations. En effet, nous avons montré que le nombre de masques vectoriels éligibles est exponentiel par rapport au nombre de coefficients des vecteurs acoustiques. L'usage d'une paramétrisation de faible dimension et donc nécessaire. Cette restriction n'est pas très pénalisante dans le cas de petits vocabulaires de l'ordre de quelques dizaines de mots, cependant, les systèmes grand vocabulaire (plusieurs dizaines de milliers de mots) nécessitent une paramétri-

sation plus fine du signal acoustique. Il semble alors intéressant de travailler au développement de nouvelles techniques pour prendre en compte les dépendances fréquentielles sur des vecteurs d'observations de plus grande dimension.

Le décodeur multi-sources développé par Jon Barker [Baker 05] est un système de reconnaissance vocale avec données manquantes. Celui-ci s'appuie sur un prétraitement du signal consistant à extraire des fragments cohérents du spectre de puissance du signal à reconnaître. Ces fragments correspondent à des regroupements de coefficients spectraux identifiés comme étant issus d'une même source sonore. L'algorithme de décodage qu'il propose recherche alors la séquence d'états acoustiques ayant engendré les observations au travers de tous les groupements possibles de fragments. Le produit de cette reconnaissance est par conséquent la transcription du signal vocal mais aussi le masque de données manquantes résultant de l'hypothèse de regroupement de fragments ayant engendré cette transcription. L'utilisation de la synchronicité d'attaque permet ici de réduire considérablement le nombre de fragments considérés à chaque instant. De cette façon, environ huit fragments sont présents en moyenne à chaque instant, ce qui représente 64 ($= 2^8$) hypothèses de regroupement.

Les masques vectoriels que nous avons proposés peuvent à notre avis prendre avantageusement place dans un tel système. Il est possible de remplacer les hypothèses de regroupement de fragments à chaque instant par un ensemble de masques vectoriels élémentaires. Cela permettrait de ne plus avoir recours à l'identification des fragments durant une phase de prétraitement et par conséquent de permettre un décodage « à la volée » du signal requis par la plupart des systèmes de reconnaissance vocale embarqués. Un second avantage est que nous n'aurions plus besoin d'apprendre des modèles paramétriques de ces masques.

Nous avons essayé de montrer, par nos travaux sur les masques de contribution, qu'il est possible de définir d'autres masques de données manquantes que ceux classiquement fondés sur un seuillage du SNR. Certes ces derniers sont intuitifs, et nous pouvons facilement les interpréter. Cependant ce critère de fiabilité reposant sur le SNR n'est interprétable que dans le domaine spectral. Nous pensons que ce domaine de paramétrisation constitue un frein à cette approche de la reconnaissance. Les très bons résultats reportés à partir de masques oracles soulignent le fort potentiel de la reconnaissance avec données manquante dans le domaine spectral, néanmoins ces performances se dégradent fortement lorsque les masques sont entachés d'erreurs. Cette forte sensibilité aux erreurs de masque est certainement une conséquence directe du manque de robustesse de cette paramétrisation.

Des travaux traitant de la reconnaissance avec données manquantes dans le domaine cepstral ont été proposés [Cerisara 03, van Hamme 04b, Srinivasan 06]. Cependant les masques cepstraux sont obtenus par une transformation non linéaire des masques spectraux. Il serait alors peut être judicieux de travailler au développement de nouveaux masques, interprétables directement

dans un domaine de paramétrisation plus robuste que le spectre. A cet égard, la paramétrisation PROSPECT développée par Van Hamme [van Hamme 04a] est intéressante. Cette paramétrisation dans le cadre de la reconnaissance avec données manquantes a permis d'obtenir de très bons résultats sur la base Aurora 4 [van Sebroeck 07]. Définir de nouveaux masques et par conséquent de nouvelles stratégies de décodage permettant l'utilisation de paramétrisations robustes du signal acoustique constitue à notre avis une orientation de recherche intéressante.

Annexe A

Rappel des concepts probabilistes pour la classification bayésienne

Dans cette annexe, nous rappelons différents concepts et terminologies utilisés dans le cadre de la classification bayésienne. Notons $\mathcal{C} = \{C_1, \ldots, C_k, \ldots, C_K\}$ l'ensemble des classes et $\mathcal{X} = \{\ldots, x, \ldots\}$ l'ensemble des observations (ou descripteurs) caractérisant tout objet à classer. Précisons que $\mathcal{C}$ est toujours fini et discret, et que $\mathcal{X}$ peut être fini ou non, continue ou discret. Chaque descripteur $x \in \mathcal{X}$ peut être mono-varié ou multi-varié.

A.1 Aspect probabiliste

A.1.1 Probabilité jointe

Soit (X, C) un couple de variables aléatoires (v.a.) à valeurs dans $\mathcal{X} \times \mathcal{C}$. Le modèle bayésien postule l'existence d'une fonction Π qui pour toutes co-réalisation $X = x$ et $C = C_k$, $(x, C_k) \in \mathcal{X} \times \mathcal{C}$, associe la probabilité $p(x, C_k)$, de cette co-réalisation. Cette fonction Π est appelée loi de probabilité jointe :

$$\Pi : \ X = x \times C = C_k \ \rightarrow \ p(x, C_k) \tag{A.1}$$

Dans le cas où $\mathcal{X}$ est *fini*, $p(x, C_k)$ est la probabilité jointe des évènements élémentaires $X = x$ et $C = C_k$:

$$p(x, C_k) \ = \ \Pi(X = x, C = C_k) \tag{A.2}$$

Dans le cas où $\mathcal{X}$ est *infini*, $p(x, C_k)$ est la densité de probabilité caractérisée par le fait que pour tout sous-ensemble $A \subseteq \mathcal{X}$:

$$\begin{aligned} p(x \in A, C_k) \ &= \ \int_A \Pi(X = x, C = C_k) \ \mathrm{d}x \\ &= \ \int_A p(x, C_k) \ \mathrm{d}x \end{aligned} \tag{A.3}$$

A.1.2 Loi marginale

La marginale de Π sur $\mathcal{C}$, ou loi marginale de C, notée Π_C, donne les probabilités *a priori* $p(C_k)$ qu'un objet non observé appartienne à chaque classe C_k :

$$\Pi_{\mathcal{C}} : C_k \rightarrow p(C_k) \quad \forall C_k \in \mathcal{C}$$

Si $\mathcal{X}$ est fini :

$$p(C_k) \;=\; \sum_{x_i \in \mathcal{X}} p(x_i, C_k) \tag{A.4}$$

Si $\mathcal{X}$ est infini :

$$p(C_k) \;=\; \int_{\mathcal{X}} p(x, C_k)\, \mathrm{d}x \tag{A.5}$$

La marginale de Π sur $\mathcal{X}$, ou loi marginale de X, notée Π_X, donne la probabilité *a priori* qu'un objet de classe inconnue soit décrit par $x \in \mathcal{C}$:

$$\Pi_X : \; x \rightarrow p(x)$$

$$\begin{aligned} p(x) \;&=\; \sum_{C_k \in \mathcal{C}} \Pi(X = x, C = C_k) \\ &=\; \sum_{C_k \in \mathcal{C}} p(x, C_k) \end{aligned} \tag{A.6}$$

La figure A.1 illustre le lien entre les notions de probabilité jointe et de loi marginale.

A.1.3 Loi conditionnelle

La loi conditionnelle de X sachant que $C = C_k$ donne la probabilité, notée $p(x|C_k)$, d'observer un objet décrit par x sachant que cet objet appartient à la classe C_k.
Si $\mathcal{X}$ est fini, $p(x|C_k)$ est la probabilité *a posteriori* de x sachant C_k.
Si $\mathcal{X}$ est infini, $p(x|C_k)$ est la vraisemblance de x. Notons que la vraisemblance de x n'a de sens qu'au regard de la classe considérée.
Dans les deux cas $p(x|C_k)$ est le rapport de la probabilité jointe $p(x, C_k)$ sur la probabilité *a priori* $p(C_k)$ de la classes C_k :

$$p(x|C_k) \;=\; \frac{p(x, C_k)}{p(C_k)} \tag{A.7}$$

La loi conditionnelle de C sachant que $X = x$ donne la probabilité, notée $p(C_k|x)$, qu'un objet observé décrit par x appartienne à la classe C_k.
Elle est définie par le rapport de la probabilité jointe $p(x, C_k)$ sur la probabilité *a priori* $p(x)$:

$$p(C_k|x) \;=\; \frac{p(x, C_k)}{p(x)} \tag{A.8}$$

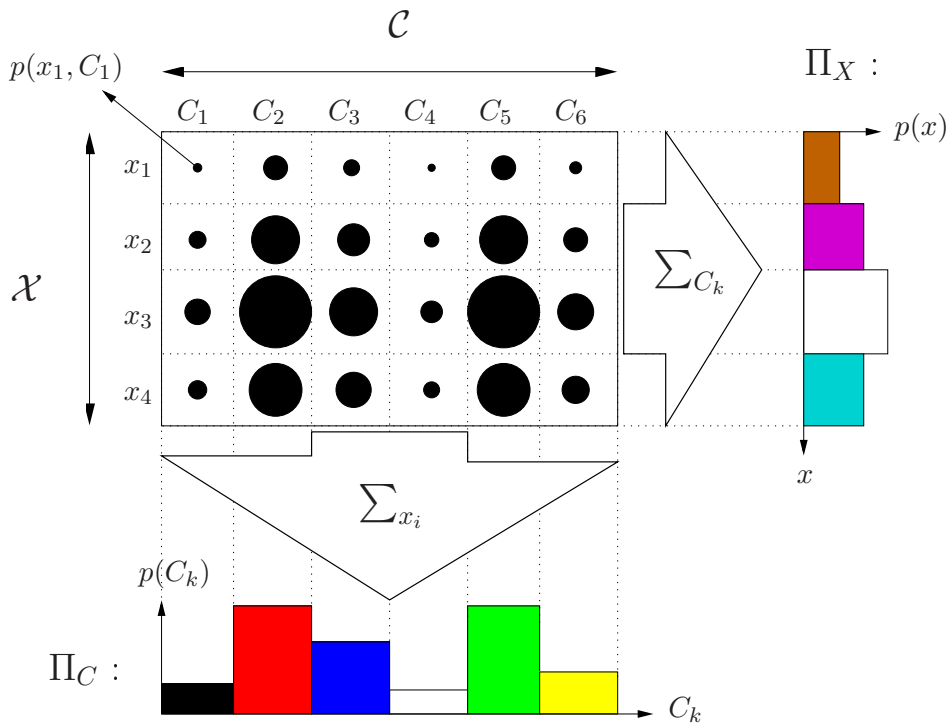


FIG. A.1 – *Représentation de la loi de probabilité jointe* $\Pi(X = x, C = C_k)$ *de deux variables aléatoires* X *et* C *à valeurs dans* $\mathcal{X} \times \mathcal{C}$ *et de ses marginales sur* $\mathcal{X}$ *et sur* $\mathcal{C}$ *:* Π_X *et* Π_C*.* $\mathcal{X}$ *est dans cet exemple un espace fini et discret. Chaque probabilité jointe* $p(x_i, C_k)$ *est proportionnelle au rayon du cercle qui la schématise.*

A.1.4 Règle de Bayes

Les équation A.7 et A.8 peuvent être réécrites :

$$p(x, C_k) = p(x|C_k).p(C_k) \tag{A.9}$$

$$p(x, C_k) = p(C_k|x).p(x) \tag{A.10}$$

En combinant ces deux équations, il vient :

$$p(C_k|x) = \frac{p(x|C_k).p(C_k)}{p(x)} \tag{A.11}$$

Cette équation est connue sous le nom de règle de Bayes.

A.2 Aspect décisionnel

A.2.1 Décisions et règles de décision

Classifier, c'est au vue de l'objet (= ses descripteurs) lui attribuer une classe C_k, ce qui s'exprime, en notant d_k la décision "ranger l'objet dans la classe C_k", comme choisir une décision d

dans l'ensemble de décisions $\mathcal{D} = \{d_1, \ldots, d_k, \ldots, d_K\}$.

Une règle de décision spécifie dans quelle classe doit être rangé tout objet susceptible d'être observé. Elle est donc choisie avant que l'observation ait lieu, c'est à dire *a priori*. Formellement, une règle de décision est une application $\delta : \mathcal{X} \to \mathcal{D}$; à tout objet x elle associe une décision $d = \delta(x)$. On notera Δ l'ensemble des règles de décision possibles.

A.2.2 Fonctions de coût et de risque

Prendre la décision d_k de ranger un objet dans une classe C_k qui se révélera ne pas être la bonne, l'objet appartenant en fait à une autre classe C_j, aura un coût - occasionnera une perte - que l'on note $w(d_k, C_j)$.

Si l'on est obligé de prendre une décision de classification sans avoir pu examiner l'objet, le choix, qui se réduit alors à celui d'une simple décision dans $\mathcal{D}$, ne peut être fait que sur la base :

1. des probabilités *a priori* que l'objet appartienne aux différentes classes.
2. des coûts d'erreur encourus.

Le critère de l'espérance de perte dit alors qu'il faut chercher à minimiser la fonction de risque *a priori* $W : \mathcal{D} \to \mathbb{R}$,

$$d_k \to W(d_k) = \sum_{j=1}^{K} p(C_j) w(d_k, C_j) \tag{A.12}$$

La meilleure décision $d_{k^\star}$ est alors celle qui minimise le risque *a priori* avec

$$k^\star = \arg\min_k W(d_k) \tag{A.13}$$

Si au contraire on peut examiner l'objet avant de le ranger, c'est une règle de décision δ qu'il faut choisir. L'évaluation *a priori* de l'espérance de perte à minimiser est alors, si $\mathcal{X}$ est fini,

$$r(\delta) = \sum_{x \in \mathcal{X}} \sum_{j=1}^{K} p(x, C_j) w(\delta(x), C_j) \tag{A.14}$$

et si $\mathcal{X}$ est infini,

$$r(\delta) = \int_{x \in \mathcal{X}} \sum_{j=1}^{K} p(x, C_j) w(\delta(x), C_j)\, dx \tag{A.15}$$

L'application $r : \Delta \to \mathbb{R}$ est appelé fonction de risque bayésien. La meilleure règle de décision $\delta^\star$ est alors celle qui minimise le risque bayésien :

$$\delta^\star = \arg\min_{\delta \in \Delta} r(\delta) \tag{A.16}$$

Le risque bayésien peut être réexprimé de façon intéressante en faisant intervenir les probabilités conditionnelles (conditionnées par x). Utilisant la relation

$$p(x, C_j) = p(x).p(C_j|x) \tag{A.17}$$

$r(\delta)$ se ré-écrit

$$r(\delta) \quad = \quad \sum_{x \in \mathcal{X}} p(x) \sum_{j=1}^{K} p(C_j|x).w(\delta(x), C_j) \quad pour\ \mathcal{X}\ fini. \tag{A.18}$$

$$ou \qquad \int_{x \in \mathcal{X}} p(x) \sum_{j=1}^{K} p(C_j|x).w(\delta(x), C_j)dx \quad pour\ \mathcal{X}\ infini. \tag{A.19}$$

On en déduit que l'on minimise le risque bayésien $r(\delta)$ en minimisant indépendamment, pour chaque $x \in \mathcal{X}$, la quantité

$$W(\delta|x) \ = \ \sum_{j=1}^{K} p(C_j|x).w(\delta(x), C_j) \tag{A.20}$$

appelée risque *a posteriori* sachant x.

Ce résultat a une conséquence pratique extrêmement importante : dans un problème donné, il est inutile de déterminer à l'avance, avant observation, la règle de décision optimale ; *il suffit d'attendre de connaître l'objet x à ranger et de minimiser alors le risque a posteriori sachant x* : au lieu d'une décision pour chaque objet x possible, on doit seulement déterminer une décision pour l'objet observé. Ce fait simplifie considérablement la mise en œuvre de la classification bayésienne.

A.2.3 Le classifieur du taux d'erreur minimum

Plaçons nous dans le cas où le coût d'erreur est constant (et où il n'y a aucun coût s'il n'y a pas d'erreur) :

$$w(d_k, C_j) = \begin{cases} 1 & si \quad k \neq j \\ 0 & si \quad k = j \end{cases}$$

Le risque *a posteriori* sachant x est alors

$$W(d_k|x) \ = \ \sum_{j \neq k} p(C_j|x) \ = \ 1 \ - \ p(C_k|x) \tag{A.21}$$

Il revient donc au même de minimiser $W(d_k|x)$ et de maximiser $\pi(C_k|x)$. Il faut donc ranger l'objet décrit par x dans la classe dont la probabilité a posteriori sachant x est la plus élevée. On dit que cette classification utilise le critère du Maximum A Posteriori (MAP).

Annexe B

Liste des publications

Revues internationales

Christophe Cerisara, Sébastien Demange and Jean-Paul Haton,
"**On noise masking for automatic speech recognition with missing data : A survey and discussion**",
Computer Speech and Language, Volume 21, issue 3, pages 443-457, Juillet 2007.

Sébastien Demange, Christophe Cerisara and Jean-Paul Haton,
"**Missing data mask estimation with frequency and temporal dependencies**",
Computer Speech and Language, En révision.

Conferences internationales

Sébastien Demange, Christophe Cerisara and Jean-Paul Haton,
"**Accurate marginalization range for missing data recognition**",
INTERSPEECH, Anvers - BELGIQUE, Août 2007.

Sébastien Demange, Christophe Cerisara and Jean-Paul Haton,
"**Missing data mask models with global frequency and temporal constraints**",
INTERSPEECH, Pittsburgh, Pennsylvanie - USA, Septembre 2006.

Sébastien Demange, Christophe Cerisara and Jean-Paul Haton,
"**Mask estimation for missing data recognition using background noise sniffing**",
ICASSP, Speech and Signal Processing, Toulouse - FRANCE, Mai 2006.

Glossaire

Acronymes	Définitions
AMS	Amplitude Modulation Spectrogramme
APPC	Adjacent Pitch Period Comparison
AR	Auto-Regressive coding
ASA	Auditory Scene Analysis
CASA	Computational Auditory Scene Analysis
EM	Expectation Maximisation
EMAP	Extended Maximum A Posteriori
ETSI	European Telecommunications Standards Institue
ETSI AFE	ETSI Advanced Front-End
FFT	Fast Fourier Transform
Fmarg	Full marginalisation
Gmarg	Gaussian marginalisation
GMM	Gaussian Mixture Model
HMM	Hidden Markov Model
HTK	HMM ToolKit
ICA	Independent Component Analysis
LPC	Linear Predictive Coding

M-SVM	Multiclass Support Vector Machine
MAP	Maximum A Posteriori
MAPLR	Maximum A Posteriori Linear Regression
MDR	Missing Data Recognition
MFCC	Mel Frequency Cepstral Coefficient
MLLR	Maximum Likelihood Linear Regression
MVS	Machine à Vecteur Support
pdf	probability density function
PLP	Perceptually-based Linear Prediction analysis
PMC	Prallel Model Combination
RAP	Reconnaissance Automatique de la Parole
RSB	Rapport signal sur bruit
SAPVR-Residuelle	Spectral Autocorrelation Peak Valley Ratio-Residual
SMAP	Structural Maximum a Posteriori
SNR	Signal to Noise Ratio
SVM	Support Vector Machine
TDNN	Time Delay Neural Network
TIR	Target to inferenced Ratio
TMR	Target to Masker Ratio
UDmarg	Uniform-Delta marginalisation
UUmarg	Uniform-Uniform marginalisation
VTS	Vector Taylor Series
WER	Word Error Rate
WI008	Paramétrisation robuste de la parole

Bibliographie

[Ahmad 93] S. Ahmad & V. Tresp. Some solutions to the missing feature problem in vision, chapitre 393-400. 1993.

[Arrowood 03] J. A. Arrowood. *Using observation uncertainty for robust speech recognition.* PhD thesis, Georgia Institute of Technology, 2003.

[Aye 02] Text, speech and dialogue : 5 th international conference, chapitre Keyword spotting Using Support Vector Machines, pages 1–6. Springer Berlin / Heidelberg, 2002.

[Bach 05] F. Bach & M. Jordan. *Blind one-microphone speech separation : A spectral learning approach.* In Advances in Neural Information Processing Systems (NIPS), volume 17, pages 65–72, 2005.

[Baker 75a] J. K. Baker. *The Dragon system - an overview.* IEEE Transaction Acoustic, Speech and Signal Processing, pages 24–29, 1975.

[Baker 75b] J. K. Baker. *Stochastic modeling as a means of automatic speech recognition.* PhD thesis, Carnegie-Mellon University, 1975.

[Baker 05] Jon Baker, Martin Cooke & Daniel P.W. Ellis. *Decoding speech in the presence of other sources.* Speech Communication, vol. 45, no. 1, pages 5–25, January 2005.

[Barker 99] J. Barker & M. Cooke. *Is the sine-wave speech cocktail party worth attending ?* Speech Communication, vol. 27, pages 159–174, 1999.

[Barker 00] J. Barker, L. Josifovski, M. Cooke & P. Green. *Soft decisions in missing data techniques for robust automatic speech recognition.* In Proc. ICSLP, Beijing, China, 2000.

[Barker 01a] J. Barker, M. Cooke & D.P.W. Ellis. *Integrating bottom-up and top-down constraints to achieve robust ASR : The multisource decoder.* In CRAC workshop, pages 63–66, Aalborg, Denmark, September 2001.

[Barker 01b] J. Barker, P. Green & M. Cooke. *Linking auditory scene analysis and robust ASR by missing data techniques.* In Proc. WISP, Stratford-upon-Avon, England, 2001.

[Barker 06] J. Barker, A. Coy, N. Ma & M. Cooke. *Recent advances in speech fragment decoding techniques.* In Proc. INTERSPEECH, Pittsburg, 2006.

[Baum 70] L. E. Baum, T. Petrie, G. Soules & N. Weiss. *A maximization technique occuring in the statistical analysis of probabilistic functions of Markov chains.* The Annals of Mathematical Statistics, vol. 41, no. 1, pages 164–171, 1970.

[Baum 72] L. E. Baum. *An inequality and associated maximization technique in statistical estimation for probabilistic functions of a Markov process.* Inequalities, pages 1–8, 1972.

[Benaroya 03] L. Benaroya. *Séparation de plusieurs sources sonores avec un capteur.* PhD thesis, Université de Rennes 1, 2003.

[Bernal-Chaves 05] J. Bernal-Chaves, C. Peláez-Moreno, A. Gallardo-Antolin & F. Díaz de Maria. *Multiclass SVM-based isolated-digit recognition using a HMM-guided segmentation.* In International Workshop on Non-Linear Spech Processing, pages 137–144, Barcelona, Spain, 2005.

[Boite 00] R. Boite, H. Bourlard, T. Dutoit, J. Hancq & H. Leich. Traitement de la parole. 2000.

[Boulard 96] H. Boulard & S. Dupont. *A new ASR approach based on independent processing and recombination of partial frequency bands.* In Proc. ICSLP, pages 422–425, Philadelphia, 1996.

[Bregman 90] A. S. Bregman. Auditory scene analysis. MIT Press, 1990.

[Breiman 84] V. Breiman, J.H. reidman & R.A. Olshen C.J. Stone. Classification and regression trees. Wadsworth, 1984.

[Bronkhorst 00] A. Bronkhorst. *The cocktail party phenomenon : a review of research on speech intelligibility in multiple-talker condition.* Acustica, vol. 86, pages 117–128, 2000.

[Brown 94] G. J. Brown & M. Cooke. *Computational auditory scene analysis.* Computer Speech and Language, vol. 8, no. 4, pages 297–336, 1994.

[Brown 01] G. J. Brown, J. Barker & D. L. Wang. *A neural Oscillator Sound Separator for Missing Data Speech Recognition.* In Proc. IJCNN-01, volume 4, pages 2907–2912, Washington, DC, USA, 2001.

[Brown 05] G. J. Brown & D. L. Wang. Speech enhancement, chapitre Separation of speech by Computational Auditory Scene Analysis, pages 371–402. Springer, New York, j. benesty and s. makino and j. chen edition, 2005.

[Cerisara 99] C. Cerisara. *Contributions de l'approche Multi-bandes à la reconnaissance automatique de la parole.* PhD thesis, INPL, 1999.

[Cerisara 03] C. Cerisara. *Towards missing data recognition with cepstral features.* In Proc. EUROSPEECH, pages 3057–3060, Geneva, Switzerland, 2003.

[Chandra 02] N. Chandra & R. E. Yantorno. *Usable speech detection using the modified spectral autocorrelation peak to valley ratio using hte LPC residual.* In 4th IASTED International conf. signal and image processing, pages 146–149, 2002.

[Christensen 07] H. Christensen, N. Ma, S. Wrigley & J. Barker. *Integrating pitch and localisation cues at a speech fragment level.* In Proc. INTERSPEECH, Antwerp, 2007.

[Cooke 93] M. Cooke. Modelling auditory processing and organization. Cambridge University Press, 1993.

[Cooke 96] M. Cooke, A. Morris & P. Green. *Recognising Occluded Speech.* In Workshop on the Auditory Basis of Speech Perception, pages 297–300, Keele university, UK, 1996.

[Cooke 97] M. Cooke, A. Morris & P. Green. *Missing data techniques for robust speech recognition.* In Proc. ICASSP, pages 863–866, 1997.

[Cooke 01a] M. Cooke & D. P. W. Ellis. *The auditory organization of speech and other sources in listeners and computational models.* Speech Communication, vol. 35, pages 141–177, 2001.

[Cooke 01b] M. Cooke, P.Green, L. Josifovski & A. Vizinho. *Robust automatic speech recognition with missing and unreliable acoustic data.* Speech communication, vol. 34, 2001.

[Cooke 03a] M. Cooke. *A glimpsing model of speech perception.* In Proc. ICPHS, pages 1425–1428, Barcelona, 2003.

[Cooke 03b] M. Cooke. *Glimpsing Speech.* Journal of Phonetics, vol. 31, pages 579–584, 2003.

[Cooke 05] M. Cooke. *A glimpsing model of speech perception in noise.* JASA, 2005. to appear.

[Cox 95] S. Cox. *Predictive speaker adaptation in speech recognition.* Computer Speech and language, vol. 9, pages 1–17, 1995.

[d'allessandro 92] C. d'allessandro & C. Demars. *Représentations temps-fréquence du signal de la parole.* Traitement du sgnal, vol. 9, no. 2, 1992.

[de Cheveigne 93] A. de Cheveigne. *Separation of concurrent harmonic sound : Fundamental frequency estimation and a time-domain cancellation model of auditory processing.* JASA, vol. 93, no. 6, pages 3271–3290, 1993.

[de Cheveigne 95] A. de Cheveigne, S. McAdams, J. Laroche & M. Rosenberg. *Identification of concurrent harmonic and inharmonic vowels : A test of the theory of harmonic cancellation and enhancement.* JASA, vol. 97, no. 6, pages 3736–3748, 1995.

[de Cheveigne 00] A. de Cheveigne. *Analyse de scène auditive sur la parole.* Aussois, France, 2000.

[Dempster 77] A. Dempster, N. Laird & D. Rubin. *Maximum likelihood from incomplete data via the EM algorithm.* Journal of the Royal Statistical Society, pages 1–38, 1977.

[Deng 05] L. Deng, J. Droppo & A. Acero. *Dynamic compensation of HMM variances using the feature enhancement uncertainty computed from a parametric model of speech distortion.* IEEE Trans. Speech and Audio Processing, vol. 13, no. 3, May 2005.

[Dupont 01] S. Dupont & C. Ris. *Assessing Local Noise Level Estimation Methods : application to noise robust ASR.* Speech Communication, vol. 34, no. 1-2, pages 141–158, april 2001.

[Ellis 96] D. P. W. Ellis. *Prediction-driven computational auditory scene analysis.* PhD thesis, MIT, June 1996.

[ETSI ES 202 050, 03] ETSI ES 202 050,. *Speech processing, Transmission and Quality aspects (STQ) ; distributed speech recognition ; advanced front-end feature extraction algorithm ; compression algorithms.* http ://www.etsi.org, 2003.

[Fletcher 37] H. Fletcher & W. A. Munson. *Relation between loudness and masking.* The Journal of Acoustical Socity of America, vol. 9, no. 1, page 78, 1937.

[Gales 93a] M. Gales & S. J. Young. *HMM recognition in noise using parallel model combibation.* In EUROSPEECH, pages 837–840, Berlin, 1993.

[Gales 93b] M. Gales & S. J. Young. *Segmental HMMs for speech recognition.* In EUROSPEECH, pages 837–840, Berlin, 1993.

[Ganapathiraju 00] A. Ganapathiraju, J. Hamaker & J. Picone. *Hybrid SVM/HMM architectures for speech recognition.* In Neural Information Processing Systems, NIPS, New York, 2000.

[Gauvain 94] J.-L. Gauvain & C.H. Lee. *Maximum a posteriori estimationj for multivariate Gaussian mixture observations of Markov chains.* IEEE Trans. Speech and Audio Processing, vol. 2, pages 291–298, 1994.

[Glotin 01] Herve Glotin. *Elaboration et comparaison de systèmes adaptatifs multiflux de reconnaisance robuste de la parole : incorporation des indices de voisement et de localisation.* PhD thesis, Institut National Polytechnique de Grenoble, 2001.

[Godsmark 99] D. Godsmark & G. J. Brown. *A blackboard architecture for computational auditory scene analysis.* Speech Communication, vol. 27, pages 351–366, 1999.

[Gong 94] Y. Gong, J.-P. Haton & F.-P. Mari. Progress and prospects of speech research and technology, chapitre Issues in acoustic modeling of speech for automatic speech recognition, pages 34–44. 1994.

[Green 95] P. Green, M. Cooke & M. Crawford. *Auditory scene analysis and HMM recognition of speech in noise.* In Proc. ICASSP, pages 401–404, 1995.

[Guermeur 05] Y. Guermeur, A. Eliseef & D. Zelus. *A comparative study of multi-class support vector machines in the unifying framework of large margin classifiers.* Applied Stochastic Model in Business and Industry 21, 2005.

[Guo 93] J. Guo & H.C. Hui. *A multilayer perceptron postprocessor to hidden Markov modeleling for speech recognition.* In ICASSP, 1993.

[Hall 92] D. L. Hall. Mathematical techniques in multisensor data fusion. Artech House, Boston, MA, 1992.

[Harris 79] D. M. Harris & P. Dallos. *Forward masking of auditory nerve response.* Journal of Neurophysiologie, 1979.

[Haton 06] J.P. Haton, C. Cerisara, D.F. Fohr, , Y. Laprie & K. Smaïli. Reconnaissance automatique de la parole : du signal à son interprétation. 2006.

[Haykin 05] S. Haykin & Z. Chen. *The Cocktail Party Problem.* Neural Computation, vol. 17, pages 1875–1902, 2005.

[Hermansky 96] H. Hermansky, S. Tibrewela & M. Pavel. *Towards ASR on partially corrupted speech.* In Proc. ICSLP, pages 462–465, Philadelphia, 1996.

[Hu 03] G. Hu & D. L. Wang. *Separation of stop consonants.* In Proc. ICASSP, volume 2, pages 749–752, 2003.

[Hu 04] G. Hu & D. L. Wang. *Monaural speech segregation based on pitch tracking and amplitude modulation.* IEEE Trans. on Neural Networks, vol. 15, no. 5, pages 1135–1150, 2004.

[Karjalainen 99] M. Karjalainen & T. Tolonen. *Separation of speech signals using iterative multi-pitch analysis and prediction.* In Proc. EUROSPEECH, volume 5, pages 2187–2190, Budapest, 1999.

[Kashino 97] K. Kashino & H. Murase. *A music stream segregation system based on adaptative multi-agents.* In Proc. of International Joint Conference on Artificial Intelligence, volume 2, pages 1126–1131, 1997.

[Keshet 07] J. Keshet, D. Grangier & S. Bengio. *Discriminative keyword spotting.* In International Workshop on Non-Linear Spech Processing, Paris, 2007.

[Khanwalkar 05] S.S. Khanwalkar, B.Y. Smolenski, R.E. Yantorno & S.J. Wenndt. *Enhancement of seaker identification using SID-usable speech.* In Eusipco, 2005.

[Kim 05] W. Kim, R. M. Stern & H. Ko. *Environment-Independent Mask Estimation for Missing-Feature Reconstruction.* In Proc. INTERSPEECH, Lisbon, Portugal, 2005.

[Kim 06] Wooil Kim & Richard M. Stern. *Band-independent mask estimation for missing-feature reconstruction in the presence of unknown background noise.* In Proc. ICASSP, volume 1, pages 305–308, Toulouse, France, May 2006.

[Kristjansson 04] T. Kristjansson, H. Attias & J. Hershey. *Single microphone source separation using high resolution signal reconstruction.* In Proc. ICASSP, pages 817–820, 2004.

[Köhler 29] W. Köhler. Gestalt psychology. New-York : Horace Liveright, 1929.

[Lazli 02] L. Lazli & M. Sellami. *Proposition d'une architecture d'un système hybride HMM-PMC pour la reconnaissance de la parole arabe.* In 7th Magrebian Conf. On Computer Sciences, volume 1, pages 101–109, Annaba, 2002.

[Lazri 84] M. J. Lazri & R. M. Stern. *A posteriori estimation of correlated jointly gaussian mean vectors.* IEEE Trans. PAMI, vol. 6, pages 530–535, 1984.

[Lee 91] C. H. Lee, C. H Lin & B.H Juang. *A study on speaker adaptation of the parameters of continuous density hidden Markov models.* IEEE Trans. Signal Processing, vol. 39, pages 806–814, 1991.

[Leggetteur 95] C. J. Leggetteur & P. C. Woodland. *Flexible speaker adaptation using maximum likelihood linear regression.* In EUROSPEECH, pages 1155–1158, Madrid, 1995.

[Levinson 86] S. Levinson. *Continuously variable duration hidden Markov models for automatic speeh recognition.* Computer Speech and Language, vol. 1, pages 29–45, 1986.

[Lippmann 97] R. P. Lippmann & B. A. Carlson. *Using missing feature theory to activately select features for robust speech recognition with interruptions, filtering and noise.* In Proc. EUROSPEECH, pages 37–40, 1997.

[Mariani 02] J. Mariani. reconnaissance de la parole : traitement automatique du langage parlé 2. 2002.

[Masuda-Katsuse 99] I. Masuda-Katsuse & H. Kawahara. *Dynamic sound stream formation based on continuity of spectral change.* Speech Communication, vol. 27, pages 253–259, 1999.

[Moore 82] B. C. J. Moore. An introduction to the psychology of hearing. Academic Press, 24/28 Oval Road, London NW1, 1982.

[Morgan 98] N. Morgan, H. Bourlard & H. Hermansky. *Automatic speech recognition : an auditory perspective.* Rapport technique, IDIAP-RR, 1998.

[Morris 98] A. C. Morris, M. Cooke & P. Green. *Some solutions to the missing feature problem in data classification, with application to noise robust ASR.* In Proc. ICASSP, pages 737–740, 1998.

[Morris 01a] A. Morris. *Data utility modelling for mismatch reduction.* In Proc. CRAC (workshop on Consistent & Reliable Acoustic Cues for sound analysis), Aalborg, Denmark, 2001.

[Morris 01b] A. Morris, J. Barker & H. Bourlard. *From missing data to maybe useful data : soft data modelling for noise robust ASR.* In Proc. WISP-01, Stratford-upon-Avon, England, 2001.

[Nakatani 02] T. Nakatani. *Computational Auditory Scene Analysis based on residue-driven architecture and its application to mixed speech recognition.* PhD thesis, Kyoto University, February 2002.

[Ofoegbu 05] U. Ofoegbu. Structure-based voiced/usable speech detection using state space embedding. Master's thesis, Temple University Graduate Board, 2005.

[O'Shaugnessy 00] D. O'Shaugnessy. Speech communication, human and machine. IEEE Press, 2000.

[Parsons 76] T. W. Parsons. *Separation of speech from interfering speech by means of harmonic selection.* JASA, vol. 60, no. 4, page 911, 1976.

[Pearce 00] D. Pearce & H.-G. Hirsch. *The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions.* In Proc. ICSLP, Beijing, China, 2000.

[Potamitis 00a] I. Potamitis, N. Fakotakis & G. Kokkinakis. *Impulsive Noise Removal using Neural Networks.* In Proc. ICASSP, volume 3, pages 1871–1874, Istanbul, Turkey, May 2000.

[Potamitis 00b] I. Potamitis, N. Fakotakis & G. Kokkinakis. *Reliable ASR based on unreliable features.* In Workshop ISCA ITRW ASR2000, Automatic Speech Recognition : Challenges for the new Millennium, volume 1, pages 53–57, 2000.

[Potamitis 01] I. Potamitis, N. Fakotakis & G. Kokkinakis. *Bayesian Independent Component Analysis as Applied to One-Channel Speech Enhancement.* In Proc. Intl. Conf. Artif. Neural Networks, volume 2130, pages 593–600, 2001.

[Quatieri 90] T. F. Quatieri & R. G. Danisewicz. *An approach to co-channel talker interference suppression using a sinusoidal model for speech.* IEEE Trans. Acoustics, Speech and Signal Processing, vol. 38, pages 56–69, 1990.

[Raj 00] B. Raj. *Reconstruction of incomplete spectrograms for robust speech recognition.* PhD thesis, Carnegie Mellon University, 2000.

[Reddy 80] D.R. Reddy, L.D. Erman, F. Hayes-Roth & V.R. Lesser. *The HEARSAY II speech-understanding system : integrating knowledge to resolve uncertainty.* ACM Computing Surveys, vol. 12, no. 2, pages 213–253, 1980.

[Remez 94] R. E. Remez, P. E. Rubin, S. M. Berns, J. S. Pardo & J. M. Lang. *On the perceptual organization of speech.* Psychological Review, vol. 101, pages 129–156, 1994.

[Renevey 01a] Philippe Renevey. *Speech recognition in noisy conditions using missing feature approach.* PhD thesis, Ecole Polytechnique Fédérale de Lausanne, 2001.

[Renevey 01b] Philippe Renevey & Andrzej Drygajlo. *Detection of reliable features for speech recognition in noisy conditions using a statistical criterion.* In Proc. CRAC-01, Aalborg, Denmark, September 2001.

[Rosenblatt 62] F. Rosenblatt. Principles of neurodynamics. Spartan Books, 1962.

[Roweis 03] Sam T. Roweis. *Factorial Models and Refiltering for Speech Separation and Denoising.* In Proc. EUROSPEECH, pages 1009–1012, Geneva, Switzerland, 2003.

[Russel 85] M. Russel & P. Moore. *Explicit modeling of state occupancy in hidden Markov models for automatic speech recognition.* In Proc. ICASSP, pages 2376–2379, Tampa, 1985.

[Russell 93] M. Russell. *A segmental HMM for speech pattern matching.* In ICASSP, pages 499–502, Mineapolis, 1993.

[Safayani 07] M. Safayani, H. Sameti, B. Babaali & M.T. Mansuri Shalmani. *An efficient multi-band spectral subtraction method for robust speech recognition.* In 20^{th} IEEE Symposium Signal Processing, ISSP, Sharjah, 2007.

[Sagi 01] S. Sagi, S. C. Nemat-Nasser, R. Kerr, R. Hayek, C. Downing & R. Hecht-Nielsen. *A biologically motivated solution to the cocktail party problem.* Neural Computation, vol. 13, pages 1575–1602, 2001.

[Sakoe 71] H. Sakoe & S. Chiba. *A dynamic programming approach to continuous speech recognition.* In 7th Int. Congress on Acoustics, Budapest, 1971.

[Sakoe 79] H. Sakoe. *Two-level DP matching-a dynamic programming based pattern matching algorithm for connected word recognition.* In ICASSP, pages 588–595, Washington DC, 1979.

[Scharenborg 06] O. Scharenborg, V. Wan & R. Moore. *Capturing fine-phonetic variation in speech through automatic classification of articulatory features.* In ITRW Workshop on Speech Recognition and Intrinsis Variation, Toulouse, France, 2006.

[Seltzer 00] M. L. Seltzer. Automatic detection of corrupt spectrographic features for robust speech recognition. Master's thesis, Departement of Electrical and Computer Engineering, Carnegie Mellon University, 2000.

[Seltzer 04] Michael L. Seltzer, Bhiksha Raj & Richard M. Stern. *A bayesian classifier for spectrographic mask estimation for missing feature speech recognition.* Speech Communication, vol. 43, pages 379–393, 2004.

[Sha 49] The mathematical theory of information. University of Illinois Press, 1949.

[Shinoda 97] K. Shinoda & C. H. Lee. *Structural MAP speaker adaptation using hierarchical priors.* In Proc. IEEE Workshop on Automatic Speech Recognition and Understanding Processes, pages 381–388, Santa Barbara, 1997.

[Shinoda 01] K. Shinoda & C. H. Lee. *A structural bayes approach to speaker adaptation.* IEEE Trans. Speech and Audio Processing, vol. 9, pages 276–287, 2001.

[Siohan 00] O. Siohan, T.A. Myrvoll & C. H. Lee. *Structural maximum a posteriori linear regression for fast HMM adaptation.* In Workshop on Automatic Speech Recognition : Challenges for the new Millenium, pages 120–127, Paris, 2000.

[Srinivasan 06] S. Srinivasan & D. Wang. *A supervised learning approach to uncertainty decoding for robust speech recognition.* In Proc. ICASSP, volume 1, pages 297–300, Toulouse, France, May 2006.

[Tchorz 01] Jurgen Tchorz, Michael Kleinschmidt & Birger Kollmeier. *Noise Suppression Based on Neurophysiologically-motivated SNR Estimation for Robust Speech Recognition.* In Advances in Neural Information Processing Systems, pages 821–827. MIT Press, 2001.

[Tchorz 02] Jurgen Tchorz & Birger Kollmeier. *Estimation of the signal-to-noise ratio with amplitude modulation spectrograms.* Speech Communication, vol. 38, no. 1, pages 1–17, 2002.

[Teukolsky 92] S.A. Teukolsky & B.P.Flannery W.T. etterling. Numerical recipes in c. Cambridge University Press, 1992.

[Tolonen 00] T. Tolonen & M. Karjalainen. *A computationally efficient multi-pitch analysis model.* IEEE Trans. Speech and Audio Processing, vol. 8, no. 6, pages 708–716, November 2000.

[van der Kouwe 01] A. J. W. van der Kouwe, D. L. Wang & G. J. Brown. *A comparison of auditory and blind separation techniques for speech segregation.* IEEE Trans. Speech and Audio Processing, vol. 9, pages 189–195, 2001.

[van Hamme 04a] H. van Hamme. *PROSPECT Features and their Application to Missing*

Data Techniques for Robust Speech Recognition. In Proc. ICSLP, pages 101–104, Jeju Island, Korea, 2004.

[van Hamme 04b] Hugo van Hamme. *Robust speech recognition using cepstral domain missing data techniques and noisy masks.* In Proc. ICASSP, volume 1, pages 213–216, Montreal, Quebec, Canada, 2004.

[van Sebroeck 07] M. van Sebroeck & H. Van hamme. *Proc. INTERSPEECH.* pages 910–913, Antwerp, Belgium, 2007.

[Vapnik 82] V.N. Vapnik. Estimation of dependences based on empirical data. Springer-Verlag, N.Y, 1982.

[Vapnik 98] V. Vapnik. Statistical learning theory. Wiley, 1998.

[Varga 90] A. P. Varga & R. Moore. *Hidden Markov model decomposition of speech and noise.* In ICASSP, pages 845–848, Albuquerque, 1990.

[Vintsyuk 68] T. K. Vintsyuk. *Speech dicrimination by dynamic programming.* Kybernetica 4, pages 81–88, 1968.

[Virtanen 02] T. Virtanen & A. Klapuri. *Separation of harmonic sounds using linear models for the overtone series.* In Proc. ICASSP, pages 1757–1760, Orlando, Florida, 2002.

[Vizinho 99] A. Vizinho, P. Green, M. Cooke & L. Josifovski. *Missing data theory, spectral subtraction and signal-to-noise estimation for robust ASR : An integrated study.* In Proc. EUROSPEECH, pages 2407–2410, 1999.

[Wachter 07] M. De Wachter. *Example based continuous speech recognition.* PhD thesis, Catholic university of Leuven, 2007.

[Wan 05a] V. Wan & J. Carmichael. *Polynomial dynamic time warping kernel support vector machines for dysarthric speech recognition with sparse training data.* In Proc. INTERSPEECH, Lisbon, Portugal, 2005.

[Wan 05b] V. Wan & S. Renals. *Speaker verification using sequence discriminant suppot vector machines.* IEEE Trans. Speech and Audio Processing, vol. 13, no. 2, pages 203–210, 2005.

[Wan 07] Kernel methods in bioengineering, signal and image processing, chapitre Building sequence kernels for speaker verification and word recognition. Idea Group Publishing, 2007.

[Wang 99] D. L. Wang & G. J. Brown. *Separation of speech from interfering sounds based on oscillatory correlation.* IEEE Trans. Neural Networks, vol. 10, pages 684–697, 1999.

[Wang 05] D. Wang. Speech separation by humans and machines, chapitre On ideal binary mask as the computational goal of auditory scene analysis, pages 181–197. Kluwer Academic, 2005.

[Weintraub 85] M. Weintraub. *A theory and computational model of auditory monaural sound separation.* PhD thesis, EE dept., Standford, 1985.

[Yantorno 03] R. E. Yantorno, B. Y. Smolenski & N. Chandra. *Usable speech measures and their fusion.* In Proc. ISCAS, 2003.

[Yen 99] K. Yen & Y. Zhao. *Adaptive co-channel speech separation and recognition.* IEEE Trans. Speech and Audio Processing, vol. 7, no. 2, pages 138–151, 1999.

Résumé

Ce mémoire propose, dans un premier temps, une introduction détaillée de la reconnaissance automatique de la parole avec données manquantes appuyée par de nombreuses références bibliographiques. Il est montré que l'estimation de masques constitue une étape cruciale. En effet, la qualité des masques estimés conditionne les performances du système de reconnaissance. L'amélioration de la fiabilité des masques constitue donc un enjeu important. Dans un second temps, les travaux menés dans le cadre de l'estimation bayésienne des masques de données manquantes sont présentés. D'une part je propose de nouveaux modèles de masques permettant de modéliser les dépendances entre les masques de différents coefficients d'un signal. Ces modèles sont évalués comparativement à un modèle de référence. Les résultats sont présentés en termes d'erreur de masques ainsi qu'en taux de reconnaissance. Les résultats montrent que ces dépendances contribuent à améliorer les taux de reconnaissance et soulignent l'importance du contexte temporel d'un masque. Je présente, dans un second temps, une nouvelle définition de masque : les masques de contribution. Ces nouveaux masques sont évalués comparativement aux masques usuellement utilisés, fondés sur le seuillage du SNR. Je montre que cette nouvelle définition permet d'améliorer l'algorithme de décodage en affinant les intervalles de marginalisation. L'évaluation, dans le cadre de la marginalisation de données et en présence d'un bruit stationnaire, montrent que les intervalles sont considérablement réduits entraînant une nette amélioration des taux de reconnaissance.

Mots clefs : reconnaissance de la parole, robustesse, données manquantes, masques

Abstract

This thesis dissertation proposes, as a first step, a detailed introduction to the automatic speech recognition with missing data supported by many bibliographic references. It is shown that the estimation of masks is a crucial step. Indeed, the quality of the estimated masks determines the performance of the recognition system. Improving the reliability of masks is thus an important issue. In a second step, new investigations in the field of Bayesian missing data mask estimation are presented. I propose first new mask models to model dependencies between the masks of different coefficients of a signal. These models are evaluated and compared to a reference model. The results are presented in terms of error of masks, as well as recognition rate. The results show that these dependencies contribute to improving the recognition rate and stress the importance of the temporal context of a mask. Second, I introduce a new missing data mask definition : the masks of contribution. These new masks are evaluated compared to masks commonly used, based on the SNR thresholding. I show how the decoding algorithm can be improved with such a mask definition by refining the likelihood marginalization intervals. The assessment, in the context of data marginalization and in the presence of a stationary noise, shows that the intervals are considerably reduced resulting in a significant improvement of the recognition rate.

Keywords : speech recognition, noise robustness, missing data, mask