UNIVERSITÉ
PARIS DESCARTES

**THÈSE DE DOCTORAT DE**
**L'UNIVERSITÉ PARIS DESCARTES**
**SORBONNE PARIS CITÉ**

Spécialité Informatique

ED130
Ecole doctorale EDITE

Présentée par

M. Matthieu Camus

Pour obtenir le grade de

**DOCTEUR de l'UNIVERSITÉ PARIS DESCARTES**

# Identification audio pour la reconnaissance de la parole.

Soutenue publiquement le 30 novembre 2011

Devant le jury composé de :

| | | |
|---|---|---|
| Mme. Marie-José Caraty | Université Paris Descartes | Directeur de thèse |
| Mme. Régine André-Obrecht | Université Toulouse 3 | Rapporteur |
| M. Paul Deléglise | Université du Maine | Rapporteur |
| M. Claude Montacié | Université Paris Sorbonne | Examinateur |
| M. Patrice Collen | Orange Labs Rennes | Examinateur |
| M. Jean-Bernard Rault | Orange Labs Rennes | Examinateur |

Université Paris Descartes - Paris 5
Sorbonne Paris Cité
Guichet Unique des Doctorants
Galerie Pasteur, bureau A4
12 rue de l'école de médecine
75006-Paris

# Remerciements

Tout d’abord, je tiens à remercier Patrice Collen et Jean-Bernard Rault pour m’avoir permis d’effectuer ce travail de recherche au sein des Orange Labs. J’étais pleinement intégré aux équipes R&D qui m’ont fourni un soutien précieux pour le bon déroulement de ma thèse de doctorat au quotidien. Un grand merci à Benoît Basset, Alexandre Delteil, Christophe Garcia, Romain Laroche, Olivier Le Blouch, Thierry Moudenc, Sébastien Onis, Pierrick Philippe, Henri Sanson.

Ce travail est le fruit d’une forte collaboration avec Marie-José Caraty, ma directrice de thèse et Claude Montacié, tous deux professeurs chercheurs universitaires de l’équipe Diadex au LIPADE. Merci à eux pour leur disponibilité, leurs conseils judicieux et leur aide constante dans l’orientation de ma recherche. Merci aussi aux autres membres de cette équipe David Janiszek et Julie Mauclair pour leur avis sincère et leur regard objectif.

Un grand merci également aux membres du jury de soutenance pour leur précieuse attention à mes travaux. Je suis particulièrement reconnaissant à mes rapporteurs Régine André-Obrecht et Paul Deléglise pour m’avoir consacré le temps nécessaire à la lecture attentive et critique de mon manuscrit.

Enfin, merci à tous ceux qui m’ont soutenu durant ce long parcours : Marie-Astrid, mes amis, ma famille.

# Résumé

Cette thèse de doctorat se place dans le cadre de la reconnaissance de la parole dans des documents audio. Le but de ce travail est d'adapter les principes de l'identification audio pour la reconnaissance de la parole ainsi que concevoir et développer des techniques d'identification robustes. Les systèmes d'identification audio par empreinte (audio fingerprinting) sont conçus pour l'indexation d'extraits de musique mais ne traitent pas des spécificités du signal de parole. Dans un premier temps, différentes méthodes d'identification audio par empreinte sont étudiées ainsi qu'un premier travail d'adaptation à la reconnaissance de la parole. Ce travail est poursuivi par le développement d'un système d'identification audio par empreinte dédié à la tâche de décodage acoustico-phonétique. De nouveaux types de sous-empreinte basés sur des paramètres usuels de la parole sont alors proposés. Dans un second temps, les différents types de variabilité du signal de parole sont décrits ainsi que les principaux paramètres de représentation acoustique du signal de parole. La robustesse de différents types de sous-empreinte à la variabilité extrinsèque et à la variabilité intrinsèque est évaluée. En présence de perturbations liées à l'environnement et aux conditions de transmission du signal de parole (CTIMIT), un type de sous-empreinte issu de l'identification audio s'avère alors le plus robuste.

# Acronymes

| | |
|---|---|
| **AM** | Amplitude Modulation |
| **ANN** | Artificial Neural Network |
| **AP** | Adaptation de la méthode de Philips |
| **API** | Alphabet Phonétique International |
| **AS** | Adaptation de la méthode de Shazam |
| **ATWV** | Actual Term-Weighted Value |
| **AUC** | Area Under Curve |
| **BER** | Bit Error Rate |
| **CLEF** | Cross Language Evaluation Forum |
| **DAP** | Décodage Acoustico-Phonétique |
| **DCT** | Discret Cosinus Transform |
| **DTW** | Dynamic Time Warping |
| **EM** | Expectation-Maximization |
| **ESTER** | Evaluation des Systèmes de Transcription enrichie d'Emissions Radiophoniques |
| **FFT** | Fast Fourier Transform |
| **FM** | Frequency Modulation |
| **FOM** | Figure Of Merit |
| **GMM** | Gaussian Mixture Model |
| **HDA** | Heteroscedastic Discriminant Analysis |
| **HMM** | Hidden Markov Model |
| **IFPI** | International Federation of the Phonographic Industry |
| **KWS** | KeyWord Spotting |
| **LDA** | Linear Discriminant Analysis |
| **LFCC** | Linear Frequency Cepstral Coefficient |
| **LPC** | Linear Predictive Coding |
| **LPCC** | Linear Predictive Cepstral Coding |
| **LSF** | Line Spectral Frequencies |
| **MCE** | Minimum Classification Error |
| **MFCC** | Mel Frequency Cepstral Coefficient |
| **MLP** | Multi-Layer Perceptron |
| **NIST** | National Institute of Standards and Technology |
| **NLDA** | NonLinear Discriminant Analysis |
| **NPC** | Neural Predictive Coding |
| **OCC** | OCCurrence-weighted value |

| | |
|---|---|
| **PCA** | Principal Component Analysis |
| **PDF** | Probability Density Function |
| **PE** | Précision Extrinsèque |
| **PLP** | Perceptual Linear Predictive |
| **PM** | Précision Moyenne |
| **QS** | Quantification Scalaire |
| **QV** | Quantification Vectorielle |
| **RAP** | Reconnaissance Automatique de la Parole |
| **RASTA PLP** | RelAtive SpecTrAl Perceptual Linear Predictive |
| **RIAA** | Recording Industry Association of America |
| **ROC** | Receiver Operating Characteristic |
| **SDR** | Spoken Document Retrieval |
| **SNR** | Signal to Noise Ratio |
| **STD** | Spoken Term Detection |
| **SVI** | Serveur Vocal Interactif |
| **SVM** | Support Vector Machine |
| **TDNN** | Time Delay Neural Network |
| **TRAP** | TempoRAl Pattern |
| **TREC** | Text REtrieval Conference |
| **VTLN** | Vocal Tract Length Normalization |
| **ZCR** | Zero Crossing Rate |

# Table des matières

# Table des figures

# Table des tableaux

# Introduction

L'identification de documents audio se place dans le cadre de la recherche par requête audio [Cano, 2007]. Dans ce cadre d'application, les systèmes d'identification audio par empreinte (*audio fingerprinting*) sont robustes à la tâche d'indexation d'extraits de musique [Haitsma et al., 2002; Chiu et al., 2010]. En effet, l'empreinte audio a pour objectif de caractériser de manière unique, compacte et robuste un extrait sonore [Haitsma et al., 2001]. Elle est idéalement un code d'identification unique du segment audio, la plus compacte possible pour limiter le volume de stockage et permettre une recherche rapide [Kurth, 2002]. Enfin, elle est conçue pour être robuste aux altérations et transformations des documents originaux telles les compressions dues à la transmission par un canal radiophonique et les bruits de l'environnement lors d'un enregistrement [Brück et al., 2004]. Il est alors possible dans de telles méthodes de faire varier le degré de similarité recherché selon le type d'application choisie [Cano et al., 2005]. Les méthodes d'identification audio par empreinte sont donc utilisées dans un cadre d'application plus large que la détection exacte d'un extrait audio original en acceptant l'identification d'un signal audio ayant subi certaines déformations [Lebossé, 2008]. Les systèmes d'identification audio sont adaptés à l'identification d'une reproduction dégradée d'un signal audio (e.g. musique de qualité compact disque audio et sa représentation compressée au format MP3) [Haitsma et al., 2002]. Cependant, ces systèmes sont limités lors de l'identification d'une nouvelle production d'un signal audio (e.g. musique enregistrée en studio et sa version rejouée en concert). Les spécificités du signal de parole ne sont pas alors traitées [Ogle et al., 2007]. Pourtant, dans les documents audio, la parole véhicule un message constituant grande partie de l'information recherchée [Allauzen, 2003].

Dans le domaine de la reconnaissance automatique de la parole (RAP), l'approche générale consiste en la transcription automatique du signal de parole en mots, telle une dictée vocale reconnaissant l'intégralité des mots prononcés. D'autres approches de reconnaissance sur un flux de parole continue existent comme la détection de mots-clés où seuls certains mots-clés sont reconnus [Juang et al., 2005]. Les applications de RAP pour le filtrage de

contenu, la recherche d'information ou la détection de mots-clés reposent alors sur le choix de l'approche mise en œuvre. La mise en place de telles applications est complexe en regard à un système d'identification audio par empreintes, quel que le type de paramétrisation acoustique choisie (vecteurs acoustiques, quantification vectorielle) [Gish et al., 2009; Leblouch, 2009; Aronowitz, 2010]. Pourtant, ces applications doivent répondre aux critères de robustesse et d'efficacité, quel que soit le type de document audio traité (durée, niveau de bruit, type de parole, superposition d'autres signaux).

Le but de ce travail de thèse est d'adapter les principes de l'identification audio pour la reconnaissance de la parole ainsi que concevoir et développer des techniques d'identification robustes. Il s'agit en particulier d'étudier la robustesse de nouvelles représentations du signal de parole issues des recherches en identification audio par empreinte [Cano et al., 2002]. La robustesse de ces nouvelles représentations est notamment évaluée en reconnaissance de mots isolés puis en reconnaissance de parole continue sur des applications de décodage acoustico-phonétique et d'identification acoustico-phonétique. Afin de mesurer les performances du système développé, on s'intéressera aux deux principaux paradigmes d'évaluation en RAP. Le premier paradigme consiste à évaluer les résultats de la robustesse des paramètres acoustiques de l'identification audio dans une tâche spécifique à la RAP sur des bases de données de signal de parole transcrit de référence internationale. Dans un second paradigme, les résultats de l'utilisation de ces paramètres acoustiques peuvent être comparés à ceux obtenus lors de l'exécution d'un système de référence proche de l'état de l'art. Ce document s'articule autour de deux parties principales avant d'aborder les perspectives en dernière partie :

- la première partie du document présente un état de l'art général sur les systèmes d'identification audio par empreinte utilisés en indexation d'extraits de musique. Dans un premier temps, le principe général de fonctionnement de ces systèmes d'identification y sera décrit ainsi que leurs caractéristiques. En particulier, les techniques de représentation du signal acoustique utilisées ainsi que les algorithmes de décodage et d'identification seront détaillés. Dans un second temps, un premier travail d'adaptation d'un tel système à la reconnaissance de la parole [Vasiloglou et al., 2004] y sera également discuté. Puis dans un troisième temps, un système d'identification audio par empreinte sera développé dans le cadre dédié de la tâche de décodage acoustico-phonétique. Ce système d'identification audio possèdera alors des spécificités adaptées à la parole. La seconde partie exposera les principaux axes sur lesquels nos recherches se sont orientées par la suite.

- la seconde partie du document présente une étude de la robustesse de paramètres acoustiques choisis face aux problématiques des variabilités extrinsèque et intrinsèque

du signal de parole. Dans un premier temps, les différents types de variabilité au sein d'un signal de parole seront discutés ainsi que les principaux paramètres acoustiques utilisés en RAP et leurs caractéristiques. Dans un second temps, deux paradigmes expérimentaux seront définis afin d'évaluer la robustesse des sous-empreintes à la variabilité extrinsèque puis aux variabilités extrinsèque et intrinsèque combinées. Un premier ensemble de sous-empreintes sera défini à partir de représentations acoustiques usuelles de la parole. Un second ensemble de sous-empreintes sera défini à partir de méthodes d'identification audio par empreinte. La proposition de telles sous-empreintes est originale par rapport aux paramètres utilisés dans les méthodes actuelles de RAP. S'appuyant sur les techniques de création d'empreinte audio, les travaux montreront la possibilité de l'usage de telles techniques à l'application d'identification acoustico-phonétique. La dernière partie proposera de poursuivre cette évaluation de robustesse dans une application d'indexation audio pour la parole.

- la dernière partie de ce document présente en perspective la possibilité de poursuivre l'évaluation de la robustesse des paramètres définis au cours de ce travail de recherche au sein d'un système de détection de mots-clés. Un système de détection de mot-clé de référence sera présenté. Les stratégies développées par le système de référence pour une meilleure détection de mots-clés seront également décrites. La robustesse de paramètres acoustiques usuels en RAP sera ensuite évaluée au sein de ce système de référence en présence de variabilités extrinsèque et intrinsèque. Nous discuterons alors des possibilités d'intégration de ces stratégies au sein d'un système d'identification audio par empreinte.

Notre démarche s'articulera autour de modules logiciels intégrant les variations de ce travail d'adaptation du système d'identification et de l'empreinte audio dans différents cadres de la RAP. Finalement, nous présenterons nos conclusions sur ce travail de recherche et rappellerons les principales perspectives relatives à l'ensemble de nos travaux.

# Première partie

Dans cette première partie, différentes méthodes d'identification audio par empreinte sont présentées. Dans un premier temps, un état de l'art non exhaustif décrira les principales méthodes d'identification par empreinte appliquées à la tâche de détection de morceaux de musique. Dans un second temps, nous discuterons d'un travail de recherche qui est la première tentative d'adaptation de l'identification audio par empreinte à une tâche de reconnaissance automatique de la parole (RAP) en la reconnaissance d'enregistrements bruités de mots isolés. En se basant sur ces méthodes d'identification audio par empreinte, nous développerons alors notre propre système de décodage acoustico-phonétique (DAP).

L'identification de documents audio se place dans le cadre de la recherche par requête audio [Cano, 2007]. Dans ce cadre, à partir d'un extrait sonore, une telle application d'identification permet de retourner de l'information textuelle, sonore ou audiovisuelle [Haitsma et al., 2001]. Par exemple, les techniques d'identification audio sont utiles au renvoi de données contenant des informations complémentaires au signal audio [Cook et al., 2006] ou à la recherche des moments d'apparition de contenus audio similaires [Lebossé, 2008]. Les catégories d'applications possibles sont variées. Elles couvrent tout autant la recherche de documents identiques que de documents similaires ayant subi des variations mineures ou de documents différents mais répondant à des caractéristiques de classification propres [Cano et al., 2005]. Ainsi les catégories d'applications rencontrées dans la littérature à travers les différentes études et travaux de recherche peuvent être :

- la recherche d'informations audio pour retourner des données associées. Il s'agit par exemple de retrouver des informations complémentaires à un extrait de musique non-identifié telles le titre du morceau, le nom de l'artiste et toute information sur la production de ce titre de musique [Cano et al., 2002]. Cette recherche peut également être utile pour identifier et localiser les éléments audio similaires dans une base de données [Betser, 2008].

- la recherche d'occurrences audio pour localiser la présence de ces occurrences. Il s'agit par exemple de localiser les jingles et points de rupture dans un continuum audio afin d'analyser le contenu d'émissions radiophoniques ou télévisuelles [Baluja et al., 2008]. Cette recherche peut également permettre de parcourir le flux audio afin de localiser certains éléments comme les applaudissements [Pinquier, 2004].

- la recherche d'occurrences audio pour contrôler la diffusion. Il s'agit par exemple de filtrer les documents audio pour ne conserver que ceux possédant un droit de diffusion adéquat [Haitsma et al., 2002]. Cette recherche peut également retourner une analyse statistique pour compter le nombre de diffusions d'un extrait de musique [Lebossé, 2008].

Ces catégories d'applications ne sont pas exhaustives. En effet, les applications de l'identification audio par empreinte sont très vastes. Elles vont du suivi de diffusion de spots publicitaires à la gestion de bibliothèques de fichiers audio de format MP3 [Cano et al., 2002]. Compte-tenu du récent développement de ces techniques de représentation du signal audio et des méthodes d'identification qui leur sont associées, de nouvelles applications vont certainement émerger suivant l'évolution des systèmes mis en place et les besoins des utilisateurs [Seidel, 2009; Cotton et al., 2010; Moulin, 2010].

Il est possible de décrire un document audio de multiples manières afin d'y associer une information le caractérisant. Ces données associées à la représentation du signal audio, appelées métadonnées, peuvent être de nature soit éditoriale (artiste, album, etc.) soit acoustique ou psycho-acoustique (tempo, rythme, énergie, etc.) [Pachet, 2005]. En tenant compte des caractéristiques acoustiques extraites à partir du signal audio, il est possible de définir différents niveaux de description du signal. Un niveau global de description se concrétisera par l'ajout de métadonnées associé au signal audio alors que les niveaux intermédiaires et locaux sont formés par une analyse du signal et de ses données acoustiques. En fonction du type d'étude du signal choisi, les caractéristiques de description peuvent être de différente nature [Lebossé, 2008] :

- de niveau global. Les caractéristiques de description sont d'ordre sémantique. Le signal audio est analysé par un utilisateur en charge de le décrire pour en extraire une interprétation compréhensible (analyse des émotions par exemple). Les études autour de ce domaine sont fortement liées aux recherches en sciences cognitives et psychologiques.

- de niveau intermédiaire. Les caractéristiques de description sont en charge de catégoriser le signal audio afin de le classer par groupes (style de musique par exemple). Une analyse statistique et une étape préalable d'apprentissage permettent ainsi de définir des classes catégorielles.

- de niveau local. Les caractéristiques de description sont issues de propriétés liées directement à l'analyse acoustique du signal audio (énergie, spectre, etc.). Il s'agit de représenter un contenu audio par une information propre et caractéristique sans pour autant définir une interprétation exploitable par l'utilisateur de l'application.

Par ailleurs, deux grandes familles de méthodes d'identification audio existent :

- les méthodes dites de tatouage audio (*audio watermarking*) [Boney et al., 1996; Cox et al., 1996],
- les méthodes par empreinte audio (*audio fingerprinting*) [Haitsma et al., 2001; Wang et al., 2003].

Une méthode sera préférée par rapport aux autres en fonction des contraintes liées aux dégradations du signal et au type d'application désiré [Cano et al., 2002; Cano et al., 2005]. Les modifications apportées au signal audio dégradé peuvent altérer de manière importante le document original. Toutes ces dégradations doivent être prises en compte dans le développement d'un système d'identification audio robuste [Balado et al., 2007]. Ainsi, dans le cadre du développement d'une méthode d'identification audio, il est nécessaire de déterminer les caractéristiques du signal audio les plus robustes aux perturbations du document audio original. Dans cette première partie, seules des méthodes d'identification audio par empreinte sont décrites et exploitées.

# Chapitre 1. Identification audio par empreinte

L'identification audio par empreinte (*audio fingerprinting*) est une méthode permettant d'associer une signature caractéristique à un signal audio analysé [Haitsma et al., 2001]. Par sa définition, une empreinte audio a pour objectif de représenter un évènement audio de manière non-intrusive, c'est-à-dire que la création de cette empreinte audio n'altère pas le signal original [Cano et al., 2002]. Il s'agit en l'occurrence d'extraire des caractéristiques acoustiques du signal audio et de les stocker dans une base de référence [Haitsma et al., 2002]. Ces caractéristiques sont généralement des descripteurs de niveau local du signal audio [Haitsma et al., 2001; Wang, 2003; Pinquier et al., 2004]. Ces techniques d'identification audio par empreinte font l'objet de travaux de normalisation MPEG-7 pour l'extraction d'informations décrivant le signal audio [Herre et al., 2002]. Cette conception d'identification audio diffère des méthodes de tatouage (*audio watermarking*) qui insèrent les informations utiles à l'identification à l'intérieur du document audio [Boney et al., 1996; Cox et al., 1996].

Dans ce chapitre, les principales propriétés d'une empreinte audio ainsi que la gestion d'une telle empreinte au sein d'un système d'identification audio sont tout d'abord décrites. Puis un ensemble représentatif des techniques de réalisation d'empreintes audio extraites du signal acoustique est brièvement listé. Trois méthodes caractéristiques de la conception d'une empreinte audio sont ensuite présentées. Il s'agit des méthodes proposées par Pinquier [Pinquier, 2004] ainsi que par les sociétés Philips [Haitsma et al., 2001] et Shazam [Wang, 2003]. Enfin, les principales contraintes et limitations de ces systèmes d'identification audio par empreinte sont alors discutées.

## 1.1 Principe de l'identification audio par empreinte

Les systèmes d'identification audio par empreinte calculent une représentation du signal audio sous la forme de signatures caractéristiques appelées empreintes audio (*audio fingerprint*) [Cano et al., 2002]. Dans ce cadre, la technique utilisée pour identifier les éléments audio est basée sur la recherche de caractéristiques acoustiques discriminantes [Lancini et al., 2004]. En particulier, cette technique permet de déterminer si deux éléments audio considérés perceptuellement similaires sont les mêmes documents, ceci même lorsque leur signal acoustique diffère. C'est le cas par exemple entre un morceau de musique présent sur un compact disque audio et sa représentation en fichier au format MP3 [Haitsma et al., 2002]. L'empreinte audio est alors l'élément de base utilisé dans la tâche de l'identification audio. Dans les systèmes d'identification audio par empreinte, l'empreinte audio extraite est une représentation bien plus compacte que le signal acoustique dont elle est issue [Haitsma et al., 2001]. Cette empreinte doit également être robuste à différents processus de traitement audio : compression, filtrage, égalisation, contrôle de la dynamique, conversion analogique-digital [Haitsma et al., 2002]. Cette robustesse répond à des besoins d'application par exemple pour une identification audio en environnement extérieur avec une possible transmission par téléphone fixe ou portable.

### 1.1.1 Contraintes de l'identification audio par empreinte

Les systèmes d'identification audio par empreinte sont soumis à diverses contraintes liées aux déformations du signal audio et aux exigences de l'application désirée. Les déformations du signal audio concernent autant les dégradations liées à l'ajout de bruits dus à la mauvaise qualité de transmission ou dus à des sons superposés au signal d'origine, que les dégradations liées à des changements de vitesse ou d'amplitude de reproduction [Allamanche et al., 2001; Haitsma et al., 2003]. Ces changements de vitesse et d'amplitude limités sont rencontrés lors de la reproduction d'extraits de musique dans des émissions radiophoniques avec des contraintes de temps et de transmission. Les dégradations concernent également les effets de compression audio avec perte d'information [Le Guyader et al., 2000], de type conversion au format audio MP3 par exemple [Haitsma et al., 2002]. De surcroît, suivant le type d'application recherché, le système est soumis à des choix d'implémentation contraints par les performances attendues dépendant de la vitesse d'exécution et de la taille de stockage nécessaire au système [Cano et al., 2005]. Ainsi, des compromis sont nécessaires entre la tolérance aux erreurs de fausse alarme et aux erreurs d'oubli ainsi qu'entre la contrainte de la quantité de données à traiter et la mise en œuvre de la structure de la base de référence. Les méthodes d'identification audio par empreinte ont pour objectif d'extraire les empreintes audio offrant alors le meilleur compromis entre toutes ces contraintes et exigences.

Dans ce cadre, l'IFPI (*International Federation of the Phonographic Industry*) et la RIAA (*Recording Industry Association of America*) ont défini un ensemble de propriétés de référence pour l'évaluation et la comparaison de méthodes d'identification audio par empreinte [RIAA, 2001]. Les principales propriétés référencées pour une telle évaluation sont [Cano et al., 2005] :

- la robustesse et l'invariance. Le document audio doit être identifiable, indépendamment des perturbations subies par le signal audio (compression avec perte de type MP3, décalage temporel et parties tronquées par la diffusion d'un court extrait, etc.) [Haitsma et al., 2002]. Cette propriété doit également pouvoir permettre l'identification du document même à partir de l'analyse d'un extrait seulement. Il s'agit alors de mettre également en œuvre un processus de synchronisation de création de l'empreinte audio lors de l'étape d'identification de l'extrait.

- la fragilité. L'empreinte issue du signal audio et caractérisant le document doit être discriminante. Il s'agit en l'occurrence de ne pas identifier un document audio distinct mais perceptuellement proche. En fonction de l'application choisie, il peut également s'agir de localiser certains types d'altérations du document audio d'origine afin de les discriminer. En ce sens, les caractéristiques de la fragilité sont à l'opposé de celles de la robustesse.

- l'efficacité. Les résultats issus de l'étape d'identification doivent être quantifiables dans leurs succès comme dans leurs échecs. Cette évaluation est effectuée en mesurant les nombres d'identifications correctes, de faux négatifs et de faux positifs.

- la complexité. Il s'agit de maîtriser le coût de calcul nécessaire à la création d'une empreinte à partir du signal audio ainsi que l'espace mémoire nécessaire au stockage des empreintes de la base de référence. Cette complexité tient également compte des procédés mis en œuvre pour la recherche d'une empreinte au sein de la base de référence.

- la sécurité. Les algorithmes mis en œuvre doivent être résistants aux tentatives de contournement. Ce contournement correspond à tout ensemble de manipulations du signal audio et de l'empreinte générée dans l'objectif de transgresser la procédure d'identification mise en place.

Toutes ces propriétés de référence pour l'évaluation servent de recommandation dans la définition des méthodes d'identification audio par empreinte. Il s'agit alors de répondre au

mieux à ces conditions requises pour le développement d'un système d'identification audio par empreinte.

### 1.1.2 Système d'identification audio par empreinte

Dans les principaux systèmes d'identification audio par empreinte, les empreintes audio sont directement calculées à partir de la source du signal acoustique [Haitsma et al., 2001; Wang, 2003; Pinquier et al., 2004]. Ces empreintes ainsi créées sont alors conservées dans une base de référence intégrée au système d'identification [Haitsma et al., 2002]. Elles sont ensuite associées à des données complémentaires. Ces données complémentaires, appelées métadonnées, représentent toute information sémantique additionnelle se rapportant à l'évènement sonore [Cano et al., 2005]. Dans l'idéal, une empreinte audio doit être unique à un évènement acoustique afin d'être en mesure de caractériser cet évènement de manière univoque [RIAA, 2001]. Lors de l'analyse d'un flux audio à identifier, des empreintes audio sont calculées de manière similaire puis comparées à celles contenues dans la base de référence afin de retourner le cas échéant les métadonnées associées servant à l'identification [Cano et al., 2002].

De tels systèmes d'identification audio par empreinte sont utilisés dans la détection d'extraits de musique à travers l'analyse d'un flux audio [Cano et al., 2005] (Figure 1.1). On distingue alors trois composantes fonctionnelles :

- un module de création d'empreinte,
- un module de stockage constitué d'une base de référence,
- un module de comparaison d'empreintes.

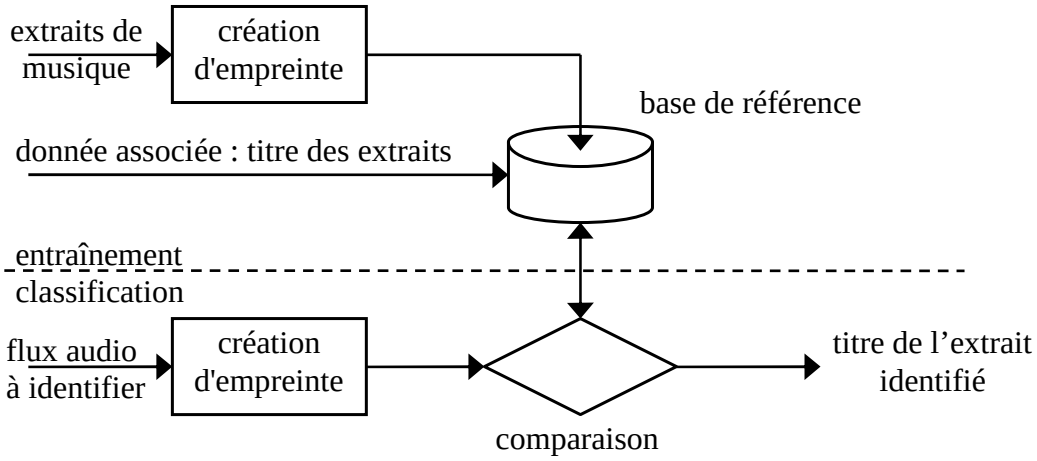



Figure 1.1 : Identification audio par empreinte pour l'identification d'extraits de musique [Cano et al., 2005]

Le module de création d'empreinte permet de sélectionner les caractéristiques du signal acoustique qui servent à définir les empreintes. Ainsi une première étape de calcul de paramètres acoustiques à partir du signal audio est nécessaire. Ces paramètres acoustiques sont ensuite utilisés pour la création des empreintes représentatives du segment de signal audio analysé. Le module de stockage permet ensuite de conserver dans une base de référence les empreintes issues du signal audio de référence et les métadonnées associées. Le module de comparaison permet enfin de mesurer la distance entre des empreintes issues d'un flux audio à identifier et celles présentes dans la base de référence à partir d'un critère de similarité. Ce module de comparaison retourne alors le cas échéant les métadonnées associées aux empreintes de référence considérées similaires à celles issues du flux audio à identifier.

### 1.1.3 Techniques de création d'empreinte

Une empreinte audio est une représentation du signal acoustique [Haitsma et al., 2001]. Il est donc possible de choisir un type de représentation en se basant sur différentes méthodes de paramétrisation acoustique pour créer ces empreintes. Les principales techniques de création d'empreinte ont donc été développées à partir :

- d'une modulation d'énergie du spectre en sous-bandes [Laroche, 2002],
- d'une analyse de magnitudes du spectre en sous-bandes [Pinquier et al., 2004],
- de multiples analyses en composantes principales [Burges et al., 2003],
- d'une représentation sous forme d'état binaire des sous-bandes fréquentielles du spectre [Fragoulis et al., 2001; Papaodysseus et al., 2001],
- d'une mesure en platitude spectrale [Herre, 2004],
- d'un descripteur basé sur les attaques sinusoïdales [Wang, 2003],
- de coefficients issus d'ondelettes pyramidales [Li et al., 2004; Baluja et al., 2008],
- de paramètres MFCC et d'une modélisation par HMM [Gomes et al., 2003],
- d'une dérivée temps-fréquence des énergies en sous-bandes [Haitsma et al., 2001; Mansoo et al., 2006; Bellettini et al., 2010].

La liste de ces techniques n'est pas exhaustive. Cependant, cette liste représente l'étendue globale de l'état de l'art en identification audio par empreinte. Un tableau récapitulatif des principales caractéristiques de ces techniques de création d'empreinte est donné [Lebossé, 2008] (Tableau 1.1).

| méthode | segmentation temporelle | création d'empreinte | indexation | comparaison et reconnaissance |
|---|---|---|---|---|
| [Li et al., 2004] | ondelette pyramidale | coefficients d'ondelettes à différentes échelles | indexation par échelle de résolution | distance euclidienne |
| [Kurth, 2002] | fenêtres recouvrantes | bit de différence d'énergie entre les instants $t$ et $t+1$ | - | distance de Hamming |
| [Burges et al., 2003] | transformée complexe modulée | 64 valeurs obtenues par analyse à composante principale pour 20 s de signal | - | distance euclidienne |
| [Brück et al., 2004] | fenêtres recouvrantes | énergies d'un banc de 8 filtres appliqués à la FFT, stockées sur 16 bits chacune | - | somme des valeurs absolues par distance euclidienne + choix de minima |
| [Haitsma et al., 2002] | fenêtres recouvrantes | bits de différence entre filtres fréquentiels | indexation par table d'empreintes et positions des moments d'apparition | distance de Hamming, calculée sur 256 valeurs binaires |

Tableau 1.1 : Principales techniques d'identification audio par empreinte [Lebossé, 2008]

Plusieurs de ces méthodes se basent notamment sur le principe d'une analyse spectrale avec découpage en sous-bandes fréquentielles. C'est le cas par exemple de la méthode développée par Pinquier [Pinquier et al., 2004] utilisant une analyse de magnitudes du spectre et de celle proposée par la société Philips [Haitsma et al., 2001]. Dans cette dernière méthode, le descripteur par dérivée temps-fréquence des énergies en sous-bandes servant à la création des empreintes et l'accès rapide à la base de référence sont souvent cités. Ces caractéristiques font l'objet de plusieurs études [Haitsma et al., 2003; Cano et al., 2005; Ke et al., 2005]. Cette technique de création d'empreinte audio est utilisée par la société Philips dans une application d'identification par empreinte pour l'identification d'extraits de musique [Haitsma et al., 2002]. Par contre, une analyse différente est utilisée pour le descripteur basé sur les attaques sinusoïdales [Wang, 2003]. En effet, un tel descripteur d'empreintes audio se base sur la

recherche de pics énergétiques. Il est actuellement utilisé dans le même type d'application pour l'identification d'extraits de musique par la société Shazam [Wang, 2006]. Ces trois méthodes exploitées par Pinquier, Philips et Shazam sont alors décrites dans le cadre de leur application dans les sections suivantes.

## 1.2 Méthode de Pinquier

Afin de répondre à la tâche de détection de jingles et des éléments de musiques courts, Pinquier propose une méthode originale d'identification audio par empreinte [Pinquier et al., 2004]. Dans cette méthode, le système développé est adapté à la segmentation automatique par recherche de segments sonores pour segmenter un flux audiovisuel [Pinquier et al., 2002]. Il s'agit alors, dans un flux audio à analyser, de détecter et de localiser des occurrences d'éléments sonores perceptuellement similaires à ceux stockés en mémoire dans la base de référence.

### 1.2.1 Vecteur de représentation acoustique et empreinte

Après une opération de fenêtrage, le signal audio est découpé en trames de quelques dizaines de millisecondes avec un taux de recouvrement de 50 %. Une opération de transformation de type temps/fréquence est alors appliquée sur chaque trame à travers une transformée de Fourier rapide (*Fast Fourier Transform*, FFT) [Cooley et al., 1965]. Puis le spectre obtenu est divisé en $N$ sous-bandes de fréquence selon une échelle perceptuelle de type Bark. L'utilisation d'une telle échelle perceptuelle permet de mieux tenir compte de la sensibilité auditive humaine. Pour chaque trame, l'énergie du signal contenu dans chacune des sous-bandes est alors calculée et conservée dans un vecteur de $N$ valeurs réelles. Chaque valeur d'énergie est ensuite normalisée par rapport à l'énergie moyenne de chacune des sous-bandes afin de retirer tout effet lié à la variation du facteur bruit/intensité. Le vecteur ainsi obtenu par les énergies en sous-bandes est alors une représentation de la trame. Pour chacun des éléments audio de référence, les vecteurs de valeurs réelles sont conservés linéairement en mémoire, les uns à la suite des autres. Une empreinte est ici un segment de vecteurs contigus de taille variable ou fixe selon le choix défini par le système. La définition de telles empreintes permet dans ce cas de comparer les éléments audio à identifier sous la forme de segments de même taille que ceux stockés en mémoire dans la base de référence.

### 1.2.2 Distance entre empreintes

La distance entre empreintes est effectuée sur la taille choisie pour le segment, vecteur par vecteur. Considérant un signal audio à comparer sous la forme d'un segment composé de $n$ vecteurs, une mesure de distance sera donc effectuée sur tous les segments de $n$ vecteurs

possibles parmi l'ensemble des éléments de la base de référence disponibles pour la comparaison. La mesure de distance entre les segments de vecteurs est ici la moyenne des distances euclidiennes entre vecteurs comparés deux à deux linéairement. Compte-tenu du stockage linéaire des vecteurs dans la base de référence, il s'agit donc d'une mesure de distance effectuée par comparaison de segments de $n$ vecteurs suivant une fenêtre glissante de la taille du segment. Cette fenêtre est déplacée pas à pas, c'est-à-dire déplacée d'un vecteur au suivant (Figure 1.2).

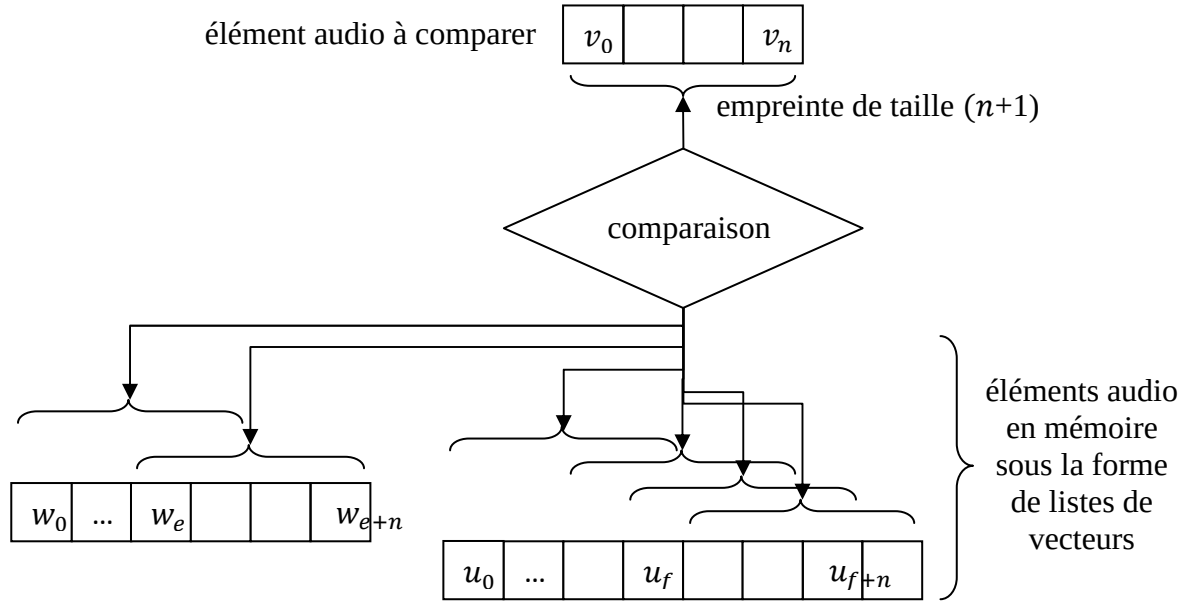



Figure 1.2 : Comparaison segment par segment, vecteur par vecteur (méthode de Pinquier)

Dans cette méthode, lorsque les éléments audio à comparer sont courts et contiennent un nombre restreint de vecteurs, il est possible de comparer l'élément audio complet par rapport à l'ensemble des segments de même dimension dans la base de référence. Dans le cas d'éléments audio longs ou de taille très variable, il est cependant nécessaire de faire appel au choix arbitraire d'une taille fixe d'empreinte pour limiter le nombre de vecteurs dans le segment. Dans ce cas, l'ajout d'une méthode de temporisation par tampon mémoire est nécessaire. La mesure de distance s'effectue alors sur l'ensemble des segments composant l'élément audio à comparer.

Afin de retourner les segments de vecteurs minimisant la distance entre empreintes ainsi définies, deux seuils sont utilisés pour répondre au critère de similarité. Un premier seuil absolu permet de balayer les éléments audio de la base de référence. Les éléments identifiés sont ceux dont la mesure de distance entre empreintes est en dessous de ce premier seuil. Dans ce cas, le critère de similarité entre ces segments de vecteurs sélectionnés est alors

considéré valide. Il s'agit en fait de rejeter les éléments audio de la base de référence dont la distance des empreintes comparées est plus élevée que cette valeur de seuil. Puis, parmi les éléments audio sélectionnés, un second seuil adaptatif permet d'évaluer la qualité de l'identification. Il s'agit de déterminer dans le segment analysé l'endroit ayant un pic maximisant la similarité entre vecteurs puis de mesurer la distance entre les vecteurs à gauche et à droite de ce pic.

Cette méthode de création d'empreinte audio proposée par Pinquier utilise une représentation du signal acoustique sous la forme de segments de vecteurs à valeurs réelles. D'autres méthodes proposent un mode de représentation du signal plus compact sous la forme de sous-empreintes. C'est le cas notamment d'une technique de création d'empreinte audio proposée par la société Philips.

## 1.3 Méthode de Philips

La société Philips a développé un système d'identification audio par empreinte en créant les empreintes audio à partir d'une analyse spectrale en banc de filtres [Haitsma et al., 2001; Haitsma et al., 2002]. Ce système extrait tout d'abord une représentation compacte des trames du signal audio sous la forme de vecteurs binaires. Cette représentation compacte de la trame est appelée sous-empreinte audio (*audio subfingerprint*). Tout au long du document, le terme sous-empreinte sera utilisé pour décrire un vecteur obtenu par représentation acoustique de la trame en réponse aux caractéristiques de l'identification audio par empreinte. Dans le cas présent d'un système développé suivant la méthode de Philips, les sous-empreintes sont regroupées sous la forme de séquences pour former les empreintes.

### 1.3.1 Vecteur de représentation acoustique et empreinte

Les échantillons issus du signal audio numérisé sont regroupés pour former une trame sur laquelle sont appliqués un fenêtrage avec recouvrement et une transformation de Fourier. Un vecteur acoustique est calculé sur une trame de durée de 370 ms avec un taux de recouvrement de 31/32. Chaque vecteur acoustique est utilisé pour générer une sous-empreinte. On obtient alors 86 sous-empreintes de 32 bits par seconde environ. Une décomposition en sous-bandes fréquentielles de la représentation spectrale est ensuite effectuée par un banc de 33 filtres passe-bande. La plage de fréquences utilisée est fixée entre 300 et 2000 Hz et segmentée suivant une échelle de type Bark. L'énergie de chaque bande fréquentielle est alors calculée en sortie de chacun de ces filtres. Afin de rendre la représentation de la trame plus robuste, le signe des différences d'énergie est calculé simultanément le long de l'axe fréquentiel et de l'axe temporel pour deux trames consécutives

(Figure 1.3). Chaque bit représente alors le signe de la variation énergétique à court-terme de deux sous-bandes fréquentielles adjacentes.

Compte-tenu de ces opérations pour la création d'une empreinte, une sous-empreinte est un vecteur de représentation de la trame du signal acoustique constitué d'un certain nombre de valeurs binaires. Considérant $E_{t,i}$ l'énergie de la sous-bande $i$ à la trame $t$, la valeur $B_{t,i}$ du bit d'indice $i$ de la sous-empreinte correspondant $B_t$ est alors calculée comme :

$$B_{t,i} = \begin{cases} 1 \; si \; \left(E_{t,i} - E_{t,i+1}\right) - \left(E_{t-1,i} - E_{t-1,i+1}\right) > 0 \\ 0 \; si \; \left(E_{t,i} - E_{t,i+1}\right) - \left(E_{t-1,i} - E_{t-1,i+1}\right) \leq 0 \end{cases} \quad (1.1)$$

Les sous-empreintes contigües sont fortement similaires car les fenêtres de signal audio sont calculées avec un fort taux de recouvrement. En effet, dans ce cas, les valeurs obtenues par double dérivée temps-fréquence sont proches entre deux sous-bandes fréquentielles adjacentes de deux trames contigües. La concaténation d'un nombre de sous-empreintes fixé par le système forme alors un segment constituant l'empreinte du signal audio sur un intervalle temporel donné.

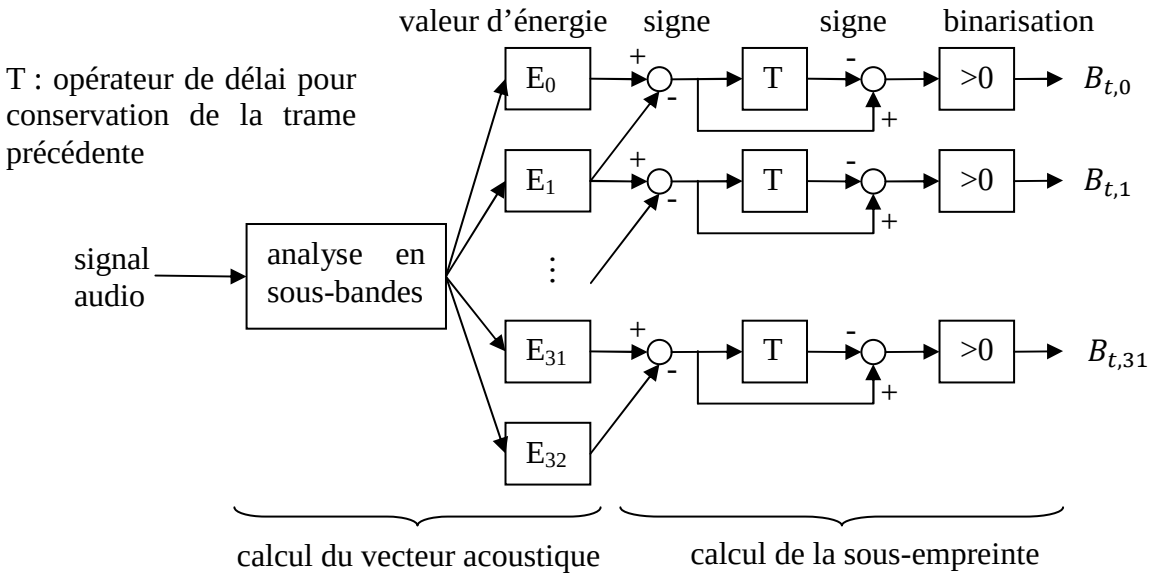



Figure 1.3 : Extraction de la sous-empreinte audio selon la méthode de Philips [Haitsma et al., 2001]

### 1.3.2 Distance entre empreintes et identification

Une fonction de distance permet de comparer les empreintes, formées par des segments de taille fixe de sous-empreintes, les unes par rapport aux autres. Lors de l'entraînement, de telles empreintes sont créées à partir du signal audio de référence puis conservées dans la base de référence avec les métadonnées associées. Lors de l'identification, un segment à identifier est formé par la concaténation des sous-empreintes du signal audio analysé jusqu'à atteindre la taille désirée pour une empreinte. Ce segment de sous-empreintes à identifier est comparé à l'ensemble des empreintes contenues dans la base de référence. La distance entre ce segment et l'empreinte est alors obtenue par le nombre moyen de bits de différence entre leurs sous-empreintes en correspondance. La mesure de distance utilisée est donc le taux moyen de bits d'erreur (*Bit Error Rate*, BER) (Figure 1.4). Le BER est basé sur la distance de Hamming locale entre deux sous-empreintes. Cette distance de Hamming est effectuée à l'aide d'une opération OU-exclusif. En tenant compte de cette mesure, deux empreintes de signal audio sont considérés similaires si le BER entre leurs segments de sous-empreintes correspondants est inférieur à une valeur seuil donnée. Dans le cas de l'application dédiée à la tâche d'identification d'extraits de musique, un taux de différence de 25 % peut par exemple être défini comme valeur seuil pour le BER. Ce seuil sur le BER correspond alors à 8 bits de différence en moyenne entre chaque paire en correspondance de sous-empreintes de 32 bits [Haitsma et al., 2001].

De surcroît, le choix du seuil sur le BER définit la tolérance aux fausses alarmes. Dans le cas de la tâche d'identification audio par empreinte, la notion de fausse alarme correspond à l'association abusive de segments du flux audio à identifier par des empreintes de référence dont les métadonnées ne sont pas celles attendues. Ainsi ce seuil doit permettre l'association d'empreintes d'un même signal audio ayant subi des déformations, tout en discriminant les empreintes issues de signaux audio d'un contenu différent. Le système proposé par la méthode de Philips est robuste à de nombreux types de dégradations dans l'application de détection d'extraits de musique [Haitsma et al., 2002]. Cette robustesse assure la possibilité d'identifier des extraits de musique, même après leur enregistrement sous une forme dégradée du signal (compression avec perte de type MP3 après une acquisition par microphone par exemple). Il permet de surcroît une classification très rapide par le traitement de plusieurs milliers d'extraits de musique en temps réel [Haitsma et al., 2002]. Ce système bénéficie également d'une faible taille de stockage de la base de référence par une représentation acoustique du signal audio limitée à 32 bits par sous-empreinte.

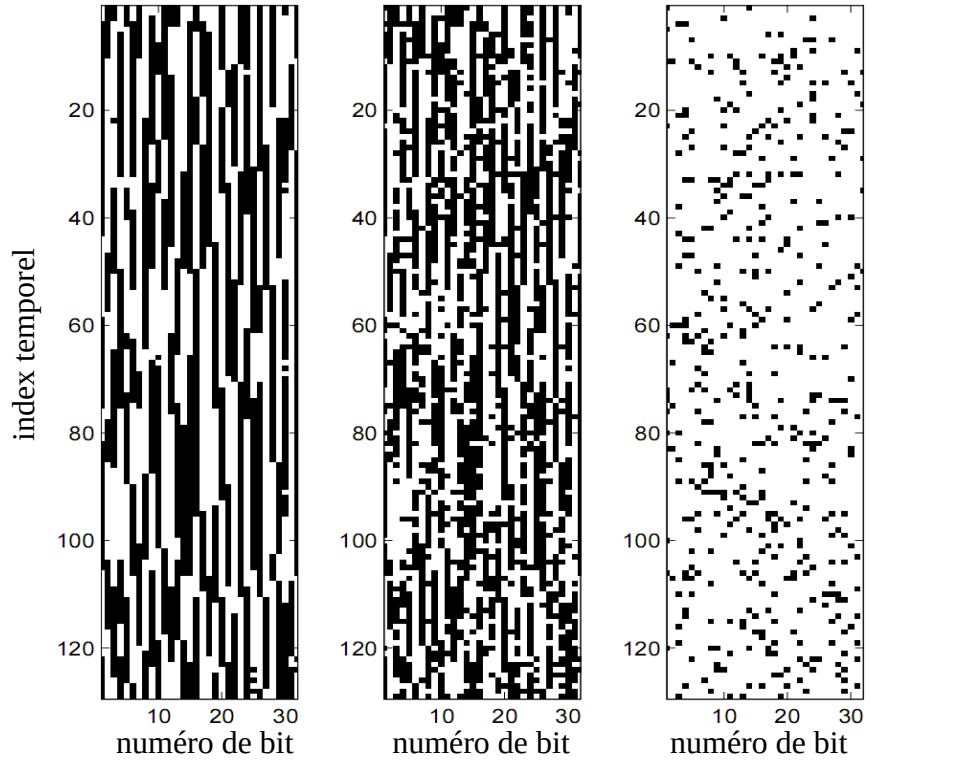



Figure 1.4 : Distance entre empreintes (méthode de Philips) [Haitsma et al., 2001]

### 1.3.3 Amélioration de la vitesse d'identification

Une méthode de réduction de l'espace de recherche est mise en œuvre pour limiter le nombre de comparaisons entre empreintes et ainsi améliorer la vitesse de l'identification. La méthode choisie ici est une indexation des sous-empreintes constituant les empreintes de référence par leur valeur. Selon la taille de la sous-empreinte sous la forme d'un vecteur binaire de $N$ bits, le tableau d'indexation obtenu sera une table à $2^N$ entrées. La taille mémoire allouée pour une sous-empreinte est définie par le système sur 32 bits, soit une taille de tableau d'indextaion de $2^{32}$ valeurs. Ce tableau d'indexation contient les liens de redirection vers les moments d'apparition dans la base des empreintes de référence des sous-empreintes de même valeur que l'index du tableau (Figure 1.5).

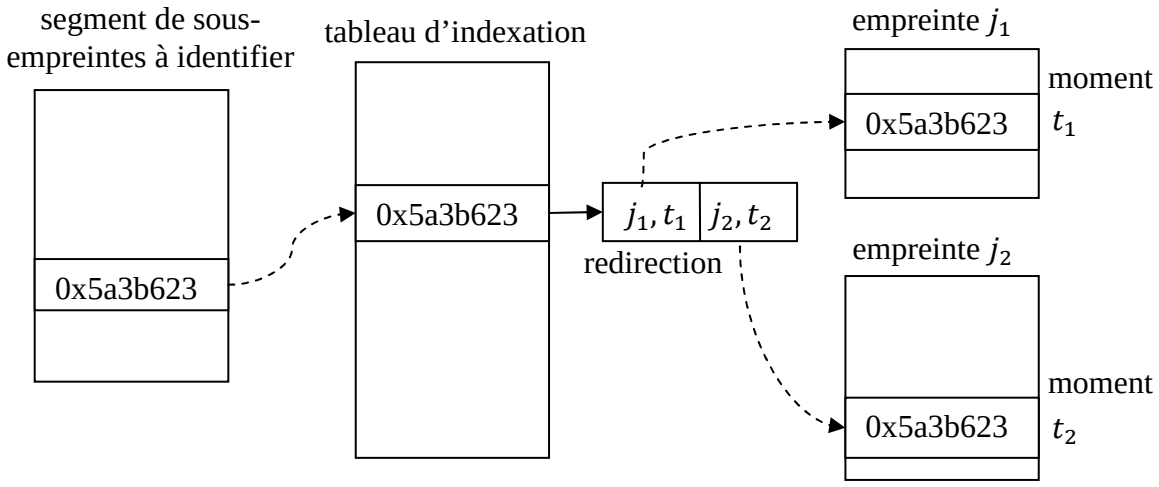



Figure 1.5 : Réduction de l'espace de recherche par un tableau d'indexation [Haitsma et al., 2001]

Cependant, cette méthode d'indexation pour la réduction de l'espace de recherche rencontre plusieurs difficultés d'implémentation. Considérant une sous-empreinte définie sur 32 bits, la taille du tableau d'indexation est très volumineuse. Le chargement d'un tel tableau en mémoire est donc difficile. De surcroît, si on considère que certaines valeurs de sous-empreinte sur 32 bits ne sont pas représentées, alors des portions du tableau d'indexation correspondant seront vides. La solution trouvée est donc d'employer une table de hachage en lieu et place d'un simple tableau d'indexation. Dans une table de hachage, l'indexation n'est plus effectuée sur la valeur de la sous-empreinte mais sur une valeur de correspondance calculée grâce à une fonction de *hash* [Cormen et al., 2001]. Ainsi, la table de hachage nécessite moins d'espace de stockage. La taille de la table de hachage dépend alors du nombre de valeurs de sous-empreintes référencées. En contrepartie, une même valeur de hachage peut être associée à plusieurs valeurs de sous-empreintes. Dans ce cas, il est donc nécessaire de vérifier que le moment d'apparition référencé par le tableau d'indexation pour une valeur de hachage donnée correspond bien à la valeur de sous-empreinte recherchée dans la base de référence.

Une seconde difficulté rencontrée est liée aux déformations possibles d'un même signal d'origine entre le segment de sous-empreintes du signal à identifier et l'empreinte du signal correspondant dans la base de référence. La méthode développée par Philips considère que si les empreintes audio définies par des segments de sous-empreintes représentent un nombre suffisant de trames du signal acoustique, alors il y a quasiment toujours la présence de sous-empreintes identiques pour deux évènements audio considérés similaires [Haitsma et al., 2001]. Cette contrainte de similarité forte peut être relâchée par la recherche des sous-

empreintes ayant au plus un certain nombre de bits de différence défini par le seuil sur le BER choisi. Le seuil sur le BER est alors fixé par le système en fonction du compromis désiré entre la réduction de l'espace de recherche et la limitation du nombre de faux rejets. D'autres techniques de création d'empreinte audio utilisent une représentation acoustique différente du signal audio. C'est le cas notamment de la méthode proposée par la société Shazam.

## 1.4 Méthode de Shazam

Une autre approche, soutenue par la société Shazam, s'appuie sur un algorithme d'étude des pics sinusoïdaux [Wang, 2003]. Il s'agit dans cette méthode de trouver et de caractériser les variations d'amplitudes locales à partir d'une analyse du spectre à court terme. Cette méthode est intégrée à une application commerciale dédiée à l'identification de morceaux de musique [Wang, 2006].

### 1.4.1 Vecteur de représentation acoustique et empreinte

Les échantillons, issus du signal audio numérisé, permettent de retourner les trames utiles pour le calcul d'une analyse spectrale par un fenêtrage avec recouvrement et une transformation de Fourier. Une segmentation par plages de fréquence du spectre ainsi obtenu permet de définir des zones d'analyse sur le plan temps-fréquence. Pour chacune de ces zones, les variations d'amplitude spectrale les plus fortes sont recherchées et considérées comme des points d'intérêt. Ces points d'intérêt correspondent alors à des attaques ou des relâchements sinusoïdaux (Figure 1.6). Un seuil de sélection sur ces variations d'amplitude est ajouté afin de conserver un nombre restreint de points d'intérêt. Afin de représenter les zones du spectre de fréquence de manière uniforme, un seuil différent est utilisé selon les bandes de fréquence considérées [Ogle et al., 2007].

Les points d'intérêts sélectionnés sont combinés deux à deux pour augmenter l'apport d'information et la capacité de discrimination du vecteur de représentation du signal audio [Betser, 2008]. Ainsi, des paires de points d'intérêt sont formées pour calculer un vecteur de représentation acoustique du signal audio. A partir de deux points d'intérêt de couple fréquence-temps $(f_1, t_1)$ et $(f_2, t_2)$, le vecteur de représentation qui leur est associé est obtenu par le triplet $(f_1, f_2, t_2 - t_1)$. L'information temporelle représentée par $t_1$ n'est pas exploitée afin de s'affranchir de la contrainte du temps absolu. Ainsi seul le temps relatif à la distance entre les moments d'apparition des deux points d'intérêt est conservé dans le vecteur de représentation.

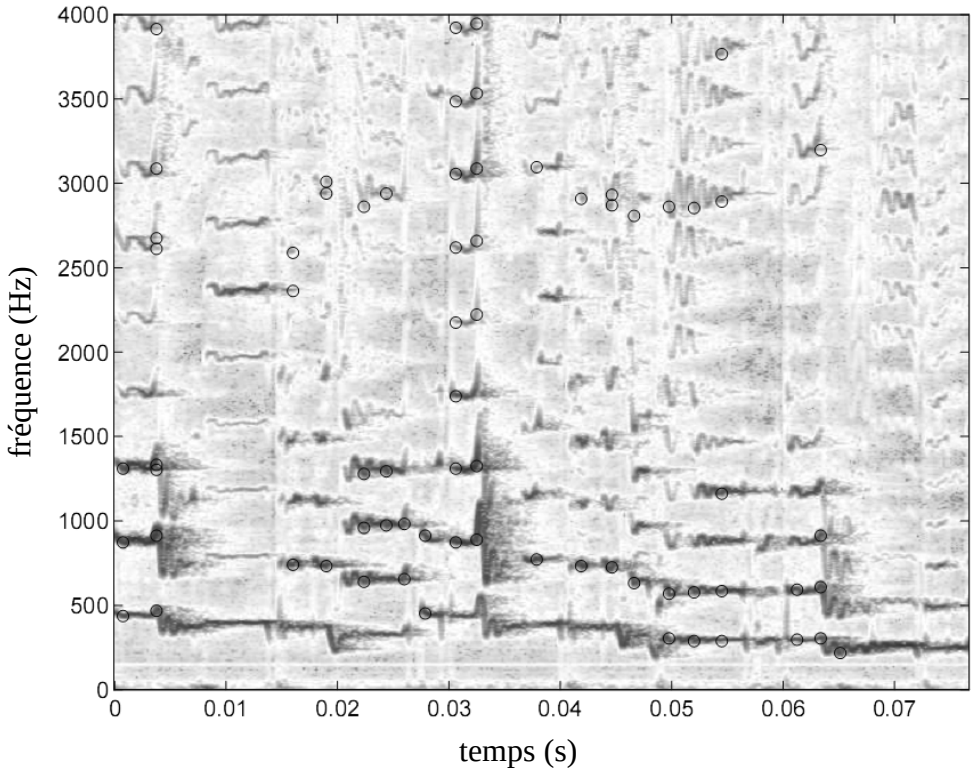



Figure 1.6 : Recherche des points d'intérêt (méthode de Shazam) [Wang, 2003]

Dans la méthode de Shazam, l'algorithme de création d'empreinte audio forme des paires de points d'intérêt en tenant compte de contraintes d'association. Ces contraintes limitent le choix d'appariement à une sélection de points d'intérêt dans le voisinage du point d'intérêt utilisé comme référence autant sur le plan temporel que fréquentiel. Considérant un point d'intérêt de fréquence-temps $(f, t)$, les vecteurs de représentation formés à partir des paires associées à ce point d'intérêt sont choisis de la façon suivante :

- limiter les candidats à l'appariement à un horizon temporel $[t + 1, t + \Delta t]$ et à un horizon fréquentiel $[f - \Delta f, f + \Delta f]$,
- conserver seulement les $K$ candidats contenant les plus grandes valeurs d'énergie.

La recherche de points d'intérêt pour l'appariement s'effectue uniquement sur les évènements acoustiques postérieurs au point d'intérêt de référence. Cette contrainte permet d'empêcher la sélection de doublons et les appariements avec une différence temporelle nulle. La valeur $K$ du nombre de candidats retenus est définie par le système en fonction du nombre de vecteurs de représentation désirés avec un même point d'intérêt de référence (en général, $K \leq 10$) [Wang, 2003].

Les empreintes sont dans ce cas des segments de vecteurs de représentation. Ces empreintes sont formées par la concaténation d'un nombre fixe de vecteurs adjacents issus du même instant temporel ou non. Similairement à la méthode de Philips, l'accès à ces vecteurs de représentation dans la base de référence est assuré par un tableau d'indexation. Une table de hachage permet alors d'obtenir une correspondance des vecteurs de représentation sous la forme de sous-empreintes. A ce moment, une opération de quantification scalaire permet de représenter chacune des trois valeurs réelles d'un vecteur de représentation sur un nombre limité de bits. Une sous-empreinte peut être définie par exemple sur 20 bits en attribuant 8 bits pour la valeur de fréquence du premier point d'intérêt, 6 bits pour la valeur de fréquence du second point d'intérêt et 6 bits pour l'intervalle temporel entre les moments d'apparition de ces points d'intérêt [Ellis, 2009]. Chaque entrée du tableau d'indexation contient alors une liste de références aux empreintes contenant le vecteur de représentation retournant la sous-empreinte et son moment d'apparition. Ainsi cette liste de références est composée d'objets définis sous la forme de couples $(j, t)$ où $j$ est la valeur d'index de l'empreinte dans la base de référence et $t$ le moment d'apparition du vecteur de représentation à l'intérieur de l'empreinte (Figure 1.5, page 21).

### 1.4.2 Distance entre empreintes et identification

La distance entre empreintes est définie selon la concordance de l'histogramme des décalages temporels entre les moments d'apparition des vecteurs de représentation similaires qui composent ces empreintes. Ces vecteurs sont considérés similaires s'ils partagent les mêmes valeurs de sous-empreinte correspondant. Lors de l'identification d'un signal audio, les sous-empreintes obtenues à partir des vecteurs de représentation sont calculés sur le signal à analyser pour un intervalle de temps donné. Compte-tenu du segment obtenu par la concaténation de ces vecteurs, l'identification est alors réalisée pour chaque vecteur de représentation temps-fréquence $(f_1, f_2, t_2 - t_1)$ au temps $t_1$ de la manière suivante :

- sélectionner dans le tableau d'indexation les objets constituant la liste de référence de même valeur que la sous-empreinte correspondante,
- pour chaque objet sélectionné $(j, t)$, calculer le décalage temporel $d = t_1 - t$,
- conserver l'empreinte $j$ comme candidat.

Puis dans un second temps, les vecteurs de représentation du signal audio à identifier sont comparés pour chacune des empreintes $j$ candidates :

- calculer l'histogramme des décalages temporels $d$ entre l'empreinte $j$ candidate et le segment de vecteurs du signal analysé,
- valider l'identification du segment par l'empreinte $j$ candidate si cet histogramme présente une vraisemblance supérieure à un seuil donné.

Durant l'étape d'identification audio, la comparaison par rapport aux empreintes de la base de référence s'effectue par une mesure de distance directement à partir des sous-empreintes. En effet, ces sous-empreintes obtenues par la fonction de hachage choisie conservent les caractéristiques des vecteurs de représentation acoustique du signal audio [Wang, 2003]. Le choix d'une telle sous-empreinte permet donc non seulement de réduire l'espace de recherche mais également de mesurer la distance entre empreintes. Cette mesure de distance est assurée par la correspondance des valeurs de fréquence des points d'intérêt conservés et le calcul du décalage temporel de leur moment d'apparition.

## 1.5 Discussions

Les méthodes d'identification audio par empreinte sont reconnues pour leurs performances dans l'identification d'extraits de musique, autant par la qualité de leur identification que par la faible complexité de leur mise en œuvre [Cano et al., 2005]. Cependant, de nouvelles contraintes apparaissent lors du changement de leur cadre d'utilisation dans l'identification d'autres types d'évènements sonores. Ces contraintes mettent alors en évidence des faiblesses et limitations à l'utilisation de telles méthodes réduisant ainsi leur champ d'application [Ogle et al., 2007].

L'information utile nécessaire à la reconnaissance d'un objet audio ayant subi des altérations est liée à l'énergie contenue dans les principales composantes sinusoïdales de cet objet audio [Fragoulis et al., 2001; Papaodysseus et al., 2001; Wang, 2003]. Cependant, les méthodes basées sur une analyse spectrale en bancs de filtres fréquentiels ne rendent pas totalement compte de cette information perceptive [Haitsma et al., 2001; Pinquier, 2004]. De surcroît, ces méthodes intègrent au sein de leurs empreintes certaines informations du signal considérées moins informatives. On peut alors considérer que ces aspects de création de l'empreinte la rendent moins robuste aux déformations du signal audio. Ainsi l'ajout de bruit non aléatoire et fortement énergétique rend ces techniques inefficaces à l'identification d'objets audio [Ogle et al., 2007]. L'amélioration de la robustesse de ces techniques d'identification audio par empreinte est toujours un sujet d'étude et de recherche [Bellettini et al., 2010; Ramona et al., 2011].

La méthode proposée par Shazam tient compte de ces contraintes en se basant justement sur une analyse des attaques sinusoïdales [Wang, 2003]. Ces attaques sinusoïdales sont des éléments fortement porteurs d'énergie. Cette méthode est cependant destinée à l'identification d'objets audio longs. En effet, un évènement audio court ne présente pas suffisamment de pics sinusoïdaux pour pouvoir être caractérisé de manière unique et efficace par cette méthode [Ogle et al., 2007]. De surcroît, cette méthode exploite l'information des composantes sinusoïdales de forte énergie. Ces composantes sont dans ce cas peu altérées lors de l'apparition de bruits additionnels. Cependant, dans une évaluation récente, il apparaît que cette méthode est moins robuste par rapport à une méthode de type Philips [Haitsma et al., 2001] en présence de bruits additionnels, de bruits convolutifs et d'une compression du signal audio [Ramona et al., 2011].

Par ailleurs, les systèmes d'identification audio par empreinte sont adaptés à la recherche d'exemplaires audio ayant subi des dégradations par rapport à un document original référencé [Cano et al., 2005]. Leur champ d'application peut être élargi à la détection d'évènements similaires dans un cadre de reproduction très contraint, pour la reconnaissance de sonneries de téléphone par exemple. Cependant, ces systèmes conçus pour la recherche de reproduction d'évènements sonores connus ne sont pas adaptés à la détection d'autres types d'évènements non-reproductibles comme ceux rencontrés en reconnaissance automatique de la parole [Ogle et al., 2007]. On peut donc s'interroger sur l'efficacité de ces techniques lorsque ces méthodes sont utilisées au sein de systèmes de reconnaissance automatique de la parole (RAP).

A ce sujet, la méthode d'identification audio par empreinte a été évaluée dans une première approche d'adaptation à la reconnaissance de la parole [Vasiloglou et al., 2004]. Dans cette adaptation présentée par Vasiloglou, il s'agit de reprendre la technique de création d'empreinte audio et le principe d'identification définis dans la méthode proposée par Philips. Cette adaptation est alors évaluée dans le cadre de la reconnaissance de mots isolés en mode mono-locuteur.

# Chapitre 2. Système d'identification audio adaptée à la parole selon Vasiloglou

*Remarque préliminaire* : l'emploi du terme empreinte audio [Haitsma et al., 2002] utilisé dans le cadre de la RAP tout au long de ce document ne fait pas référence aux techniques d'identification basées sur l'empreinte vocale connue dans le domaine de l'authentification du locuteur [Doddington, 1985]. Ce terme fait uniquement référence à une empreinte acoustique calculée à partir d'une analyse du signal audio (Annexe A, page 133).

Une adaptation de la méthode d'identification audio par empreinte est proposée au sein d'une application de reconnaissance d'enregistrements de parole bruités [Vasiloglou et al., 2004]. Le système proposé par Vasiloglou utilise la technique de création d'empreinte audio et la méthode d'identification proposées par Philips [Haitsma et al., 2002] (section 1.3, page 17). Il s'agit dans le cadre de l'étude présentée par Vasiloglou d'évaluer la performance d'une telle adaptation pour la reconnaissance d'enregistrements bruités de mots isolés en mode mono-locuteur.

Le principe général de ce système d'identification de mots isolés est similaire à celui d'un système d'identification audio par empreinte [Haitsma et al., 2002] (Figure 1.1, page 12). Dans le cas présent, il ne s'agit pas d'identifier des extraits de musique mais de reconnaître des mots prononcés isolément les uns des autres. Au sein de ce système, pour chaque occurrence de mot enregistré, l'empreinte correspondante est définie à partir du segment des sous-empreintes qui la composent. Cependant, une telle empreinte ne conserve qu'un unique exemplaire de chaque valeur de sous-empreinte en les ordonnant par ordre croissant. Les empreintes des mots isolés ainsi que leur transcription associée sont conservées dans une base de référence. L'identification d'un mot prononcé consiste alors tout d'abord à créer son empreinte à partir du signal de parole analysé. Ensuite, il s'agit de compter le nombre de sous-empreintes similaires avec chacune des empreintes conservées dans la base

de référence. Le mot retourné pour l'identification est celui associé à l'empreinte de la base de référence qui maximise le nombre de sous-empreintes similaires à l'empreinte issue du signal analysé. L'évaluation du système consiste à mesurer sa performance en précision dans la tâche d'identification de mots. Cette évaluation est effectuée entre les empreintes de référence du signal de parole original conservées en mémoire et les empreintes issues du même ensemble de parole mais dont le signal aura été dégradé par un bruit additionnel. Ce système est détaillé ci-dessous.

## 2.1 Vecteur de représentation acoustique et empreinte

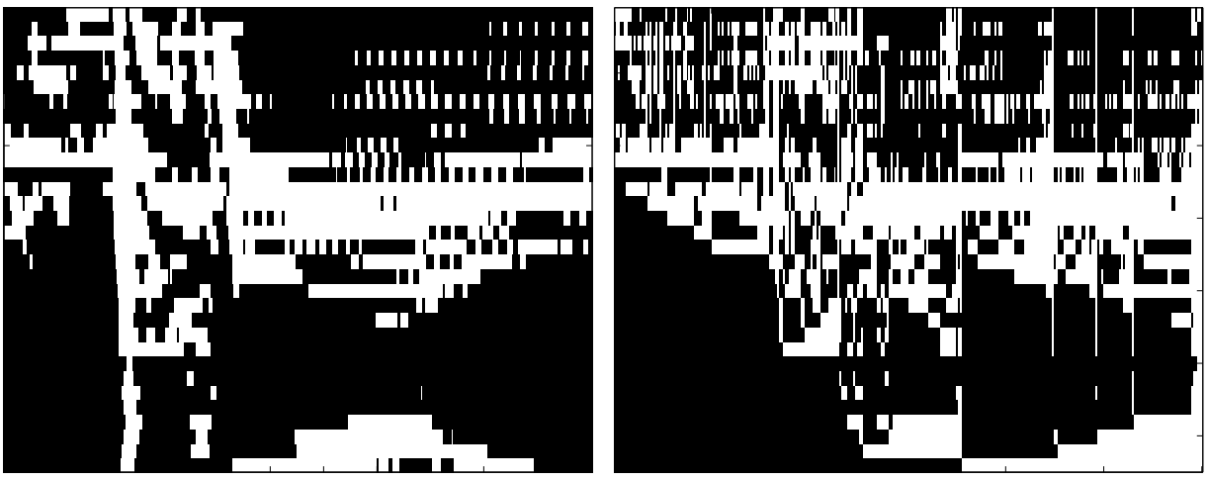
Une sous-empreinte conserve ici la même définition que dans la méthode de Philips [Haitsma et al., 2001]. Chacun des bits d'une sous-empreinte de 32 bits représente, pour une trame donnée, le signe des variations énergétiques simultanément selon les axes temporels et fréquentiels. Ce calcul binaire est effectué à partir d'une représentation spectrale d'une trame du signal de parole segmentée en 33 sous-bandes fréquentielles adjacentes. Un taux de recouvrement élevé permet de détecter tout changement abrupt dans le spectre. Ce choix est conditionné par la durée très courte de production des plosives (Annexe G, page 149). Grâce à ce taux de recouvrement élevé, les sous-empreintes contigües sont fortement similaires. La variation dans le temps des sous-empreintes consécutives est donc lente.

Les trames d'une durée de 18,75 ms sont calculées toutes les 0,625 ms en appliquant une fenêtre de Hanning. Pour un signal audio échantillonné à 16 kHz, ce fenêtrage correspond à des trames d'une taille de 300 échantillons calculées tous les 10 échantillons. Une transformée de Fourier rapide retourne l'analyse spectrale du signal. Un découpage du spectre en sous-bandes fréquentielles est effectué sur l'intervalle de 100 Hz à 2000 Hz, selon une échelle de Mel. Les sous-empreintes calculées par la méthode présentée par Philips montrent une grande robustesse à de nombreux types de dégradation du signal audio [Haitsma et al., 2002]. En se basant sur cette méthode, le calcul des sous-empreintes utilise l'information du signe des variations énergétiques sur les axes temps-fréquence à court terme. Pour discriminer les bandes du spectre correspondant à une zone de silence, un seuil d'énergie $T$ est défini à 0,01 dB. Dans ce cas, considérant $E_{t,i}$ l'énergie de la sous-bande $i$ de la trame $t$, le bit d'indice $i$ de la sous-empreinte $B_t$ est alors calculé comme :

$$B_{t,i} = \begin{cases} 1 \; si \; E_{t+1,i+1} - E_{t,i} > 0 \\ 0 \; si \; E_{t+1,i+1} - E_{t,i} \leq 0 \\ 0 \; si \; E_{t+1,i+1} < T \quad \text{ET} \quad E_{t,i} < T \end{cases} \tag{2.1}$$

Dans le cas présent, une telle sous-empreinte est ainsi créée dans le but d'une meilleure robustesse aux phénomènes de variation vocale liés aux plosives et fricatives par rapport à une sous-empreinte définie selon la méthode de Philips. De surcroît, l'ajout d'un seuil sur les valeurs d'énergie permet de ne pas tenir compte des sous-bandes de fréquence faibles en énergie. Dans ce cas, on considère que ces sous-bandes ne contiennent pas d'information utile du signal de parole. Donc le signal audio présent au sein de ces sous-bandes à faible valeur d'énergie est alors considéré comme du bruit.

Contrairement à la méthode de Philips [Haitsma et al., 2001], les empreintes issues du signal de parole ne sont pas des segments de taille fixe. En effet, les empreintes sont calculées dynamiquement à partir de la segmentation en mots. Cette segmentation est fournie par la transcription textuelle de la base de données audio accompagnant le signal de parole. Pour chaque prononciation de mot enregistré, une empreinte contient alors les différentes valeurs des sous-empreintes issues du signal de parole correspondant à la segmentation de ce mot dans le continuum de parole. Ces valeurs ne sont ni redondantes ni placées au sein de l'empreinte selon l'ordre d'apparition temporelle des trames du signal de parole. Ces sous-empreintes sont en effet triées par leur valeur en ordre croissant et un seul exemplaire de chaque valeur est conservé (Figure 2.1). Ainsi, les informations des moments d'apparition des sous-empreintes et leur nombre d'occurrences ne sont pas conservées. Comme les empreintes ne conservent qu'un seul exemplaire de sous-empreintes identiques, la taille des empreintes obtenues est inférieure ou égale à la taille du segment des sous-empreintes issues de la segmentation du mot.



liste des sous-empreintes du mot *carry* empreinte correspondante au mot *carry*

Figure 2.1 : Création d'une empreinte de type Vasiloglou (mot *carry*) [Vasiloglou et al., 2004]

## 2.2 Distance entre empreintes et identification

La mesure de distance entre deux empreintes est basée sur le nombre de sous-empreintes identiques entre les deux empreintes. La mesure de distance ainsi définie est donc une procédure peu coûteuse. En effet, il s'agit d'identifier les sous-empreintes présentes en commun dans les deux empreintes. Donc considérant la fonction $taille(A)$ retournant le nombre de sous-empreintes de l'empreinte $A$, le nombre de comparaisons à effectuer lors de la mesure de distance entre deux empreintes $A$ et $B$ est $M = \min(taille(A), taille(B))$. La complexité de cette mesure est alors en $O(M)$. Cette faible complexité est obtenue grâce au tri en ordre croissant des sous-empreintes au sein d'une empreinte. Si les sous-empreintes étaient conservées dans leur ordre d'apparition temporelle lors de la formation d'une empreinte, alors la complexité de la mesure de distance entre empreintes dont la taille est du même ordre de grandeur serait en $O(M^2)$.

L'identification d'une empreinte issue du signal de parole à analyser est effectuée par mesure de distance sur l'ensemble des empreintes de référence conservées dans la base de référence. Cette base de référence est accompagnée d'une table de hachage similaire à celle proposée dans la méthode de Philips pour l'amélioration de la vitesse d'identification [Haitsma et al., 2001] (section 1.3.3, page 20). La table de hachage mise en place permet un accès rapide en retournant les différentes positions des moments d'apparition des sous-empreintes dans la base de référence. La procédure d'identification est la suivante :

- choisir une empreinte de référence ayant au moins une sous-empreinte commune avec l'empreinte à identifier,
- compter le nombre de sous-empreintes communes entre les deux empreintes,
- à partir de ce nombre, attribuer un score de similarité entre les deux empreintes,
- conserver dans une liste triée les meilleurs scores de similarité et leur empreinte de référence associée.

Ainsi, lors de l'identification d'une empreinte, les mots retournés seront ceux associés aux empreintes de référence dans la liste triée. Selon l'application visée, les scores de similarité de chacune de ces empreintes dans la liste triée peuvent être utilisés comme résultat direct ou comme rapport de probabilité de prononciation du mot retourné de l'empreinte évaluée pour l'identification.

Parfois, le nombre de sous-empreintes en commun entre les empreintes à comparer est très faible. Si lors de l'étape d'identification, aucune empreinte de référence ne possède en commun plus de 30 % des sous-empreintes de l'empreinte à identifier, alors une nouvelle procédure d'identification est mise en place. Dans ce cas, l'option de recherche 1-bit est

appliquée. Cette option de recherche permet de multiplier les possibilités d'identification par les empreintes de la base de référence. Il s'agit de modifier un par un les bits de chacun des sous-empreintes de l'empreinte à identifier qui ne sont pas en commun avec l'empreinte de référence choisie pour la comparaison. Ainsi si les sous-empreintes comparées deux à deux ne diffèrent que d'un seul bit, ils seront comptabilisés dans le calcul du score de similarité. Cependant, cette option de recherche 1-bit augmente de manière significative la complexité de l'algorithme. En effet, l'espace de recherche est augmenté d'autant que le nombre de modifications possibles bit par bit des sous-empreintes. De surcroît, en modifiant une sous-empreinte par l'option de recherche 1-bit, l'organisation des sous-empreintes par ordre croissant au sein de l'empreinte n'est plus respectée. Il est donc nécessaire de parcourir le segment de sous-empreintes de l'empreinte de référence choisie pour la comparaison afin de déterminer si cette sous-empreinte modifiée est commune aux deux empreintes. La complexité de la mesure de distance entre empreintes dont la taille est du même ordre de grandeur est donc désormais en $O(M^2)$ pour chacune des nouvelles sous-empreintes calculées par l'option de recherche 1-bit.

## 2.3 Expériences

Afin d'étudier le comportement du système proposé par Vasiloglou au sein de différentes applications de RAP, trois expériences sont effectuées :

- reconnaissance d'enregistrements bruités de mots isolés,
- reconnaissance de mots isolés bruités en environnement mono-locuteur,
- reconnaissance de phonèmes isolés en environnement multi-locuteur.

Dans un premier temps, le cadre d'expérimentation de l'expérience originale de l'article de référence est reproduit. Il s'agit de reconnaître des mots isolés à partir d'enregistrements bruités dont le signal de parole propre d'origine du mot correspondant est présent dans la base de référence. Dans un second temps, cette première expérience est étendue à une reconnaissance de mots isolés dans un environnement mono-locuteur. Il s'agit alors de reconnaître des mots isolés bruités dont le signal de parole propre d'autres prononciations de ce mot par le même locuteur est présent dans la base de référence. Dans un troisième temps, le système est évalué pour une tâche de reconnaissance de phonèmes isolés dans un environnement multi-locuteur. Il s'agit dans ce cas de reconnaître des phonèmes issus d'un signal de parole continue mais dont la segmentation est connue. La base de référence utilisée pour le test contient alors le signal de parole de phonèmes prononcés par des locuteurs différents de ceux prononçant les phonèmes à reconnaître. Dans le cadre de cette dernière expérience, de nouvelles méthodes de création d'empreinte sont alors proposées.

### 2.3.1 Reconnaissance d'enregistrements bruités de mots isolés

Dans cette première expérience, le cadre d'expérimentation proposé par Vasiloglou est reproduit [Vasiloglou et al., 2004]. Dans l'article de référence, l'ensemble de la base de données KED-TIMIT est segmenté en mots [CSTR, 2001] (Annexe B, page 135). La base de données d'origine est convertie afin de traiter les mots prononcés isolément les uns des autres. Une empreinte est alors calculée à partir du signal de parole pour chacun de ces mots isolés. L'ensemble constitué des empreintes et de leur étiquette linguistique de mot associé est conservé dans la base de référence.

Cette première expérience et son évaluation sont définies en respectant le plus possible les indications fournies par l'article de référence. Différents ensembles de données sont choisis afin d'étudier l'influence des mots courts sur les performances de la reconnaissance. Par ailleurs, un même mot peut être représenté par plusieurs occurrences suivant le nombre de fois où il a été prononcé dans la base de données. Trois séries de données utilisées pour cette expérience sont ainsi constituées par :

- la série $S_1$ formée par tous les mots de la base de données (3275 mots prononcés),
- la série $S_2$ formée par les mots de deux lettres et plus (3167 mots prononcés),
- la série $S_3$ formée par les mots de trois lettres et plus (2831 mots prononcés).

Grâce à la transcription permettant la segmentation des mots, les empreintes de référence et celles issues du signal de parole à identifier sont calculées de manière identique. Concrètement, les sous-empreintes extraites du signal de parole des fichiers audio originaux sont rassemblées sous la forme d'empreintes de mots. Ces empreintes initiales définissent la base de référence. Les dégradations du signal audio original sont effectuées en ajoutant un bruit blanc à diverses valeurs de rapport signal à bruit (*Signal to Noise Ratio*, SNR). Un bruit blanc comme évènement sonore est la réalisation acoustique d'un processus aléatoire dans lequel la densité spectrale de puissance est la même pour toutes les fréquences [Hugonnet et al., 1998]. Dans le cadre de nos expériences, il est nécessaire de vérifier que le bruit blanc choisi respecte bien cette propriété mais à moyen terme uniquement. Si cette propriété de répartition est respectée à court terme, c'est-à-dire d'une trame à la suivante, ce bruit blanc sera discriminé lors de la création de l'empreinte. En effet, les effets d'un bruit blanc ayant une répartition identique sur chacune des sous-bandes fréquentielles d'une trame à la suivante sont annulés compte-tenu du mode de calcul d'une sous-empreinte par dérivée d'énergie sur deux trames consécutives. Donc en tenant compte de cette contrainte, la base d'évaluation des empreintes du signal de parole à identifier est construite à partir des mêmes fichiers audio que ceux de référence. Cependant, le signal de parole propre est modifié par l'ajout d'un signal

audio de bruit blanc gaussien réparti sur une plage de fréquences correspondant au champ auditif. Ainsi pour chacune des trames du signal audio utilisé pour l'évaluation, les valeurs pondérées d'échantillons issus d'un bruit blanc gaussien sont additionnées aux valeurs des échantillons du signal de parole d'origine (Annexe D, page 139).

Il s'agit ensuite d'évaluer dans quelle mesure l'identification de l'empreinte issue de la base d'évaluation retourne bien le mot recherché. Une première évaluation porte sur la présence du mot recherché parmi les résultats de meilleur score de la liste triée retournée par le système. Une seconde évaluation porte sur la présence de ce mot parmi ceux des résultats des 10 meilleurs scores retournés par la liste triée. Comme plusieurs résultats peuvent être retournés pour un seul et même score, la liste triée peut contenir un nombre beaucoup plus important de mots correspondant aux 10 meilleurs scores. De surcroît, pour chacune de ces évaluations, les résultats sont obtenus avec activation de l'option de recherche 1-bit et en mode normal sans cette option. Différentes valeurs de pondération du bruit blanc gaussien sont également choisies parmi les paramètres d'évaluation. Les résultats obtenus sont fournis dans un tableau récapitulatif pour des valeurs de SNR à 20 dB et à 0 dB (Tableau 2.1).

| **performance %** / **données** | **SNR** | **sans option de recherche** | | **option de recherche 1-bit** | |
|---|---|---|---|---|---|
| | | **meilleur** | **10 meilleurs** | **meilleur** | **10 meilleurs** |
| série $S_1$ (3275 mots) | 20 dB | 92,3 | 98,1 | 89,5 | 98,5 |
| | 0 dB | 56,8 | 93,8 | 42,0 | 88,8 |
| série $S_2$ (3167 mots) | 20 dB | 93,2 | 98,5 | 90,7 | 98,6 |
| | 0 dB | 58,2 | 94,2 | 43,1 | 88,5 |
| série $S_3$ (2831 mots) | 20 dB | 94,1 | 98,7 | 91,9 | 98,6 |
| | 0 dB | 61,6 | 94,8 | 46,1 | 89,0 |

Tableau 2.1 : Résultats de la reconnaissance d'enregistrements bruités de mots isolés par empreinte de Vasiloglou (KED-TIMIT)

Pour l'ensemble des expériences sur les séries $S_1$, $S_2$ et $S_3$, le taux de présence du mot recherché dans la liste des 10 meilleurs scores dépasse 98 % lorsque le SNR est à 20 dB. Avec ce SNR, si l'option de recherche 1-bit n'est pas activée, le taux de présence du mot recherché en meilleur score dans la liste augmente lorsque les petits mots sont supprimés. Dans ce cas, le taux de présence évolue en effet de 92 % à 94 % environ. Cependant dans ce cas, lorsque cette option de recherche 1-bit est activée, les performances se dégradent légèrement à chaque fois, diminuant de 2 % à 3 % environ. En effet, dans ce cas précis, l'augmentation de l'espace de recherche pour la mesure du score apportée par cette option permet la validation d'empreintes indésirables parmi celles de meilleur score.

Dans le cas où le paramètre du SNR est à 0 dB, les résultats sont plus mitigés. Tout d'abord, la présence d'un bruit ajouté au signal de parole à un tel niveau énergétique entraîne de grandes modifications des sous-empreintes constituant l'empreinte. Ainsi, lorsque l'option de recherche 1-bit n'est pas activée, le taux de présence du mot recherché en meilleur score dans la liste est beaucoup plus faible par rapport à un SNR à 20 dB, situé autour de 57 % à 61 %. Cependant, dans ce cas, le taux de présence du mot recherché dans la liste des 10 meilleurs scores reste élevé, autour de 94 %. Comme pour les cas précédents lorsque le SNR est à 20 dB, l'option de recherche 1-bit ne permet pas non plus une plus forte présence du mot recherché en première position dans la liste. Au contraire, lorsque le SNR est à 0 dB, le taux de présence du mot recherché en meilleur score se dégrade fortement, descendant autour de 42 % à 46 %.

Les résultats de cette expérience sont encourageants dans la recherche d'une adaptation des méthodes d'identification audio par empreinte pour des applications de RAP notamment lorsque le SNR considéré est en faveur du signal de parole. Cependant, nous n'avons pas pu retrouver des résultats similaires à ceux présentés par Vasiloglou. Dans l'article de Vasiloglou, les performances d'une telle identification de mots isolés d'un signal bruité à 0 dB sont supérieures à 70 % en se limitant au meilleur score retourné par la liste triée. Cette performance est obtenue par l'utilisation de l'option de recherche 1-bit qui améliore dans ce cas l'efficacité du système. Nous avons reproduit l'option de recherche 1-bit à partir de la description fournie par l'article de référence. Dans notre cas toutefois, nous obtenons une dégradation des performances par rapport à une évaluation sans cette option de recherche, contrairement aux résultats attendus.

Enfin, les conditions d'expérimentation ne sont pas adaptées à une application de reconnaissance de la parole en conditions réelles comme le cadre de la RAP en parole

continue. En effet, les bases de données de référence pour l'apprentissage et de test pour la reconnaissance sont issues du même signal de parole, seul un bruit additionnel les différencie. Donc les variations de prononciation d'un mot ne sont pas prises en compte dans ce type d'expérience. Les performances de reconnaissance sont alors de l'ordre de 100 % lorsque l'énergie de ce bruit additionnel est très faible par rapport à l'énergie du signal d'origine. Ces performances décroissent avec la diminution du SNR augmentant ainsi l'importance du bruit par rapport au signal d'origine. Nous désirons poursuivre ce type d'expérience d'identification audio par empreinte appliquée à la parole, en étendant le champ d'évaluation à une reconnaissance de mots isolés bruités. Cette reconnaissance est effectuée sur des mots isolés bruités dont le signal de parole propre du mot recherché n'est pas contenu dans la base de référence.

### 2.3.2 Reconnaissance de mots isolés bruités en environnement mono-locuteur

Dans cette seconde expérience, les principales conditions d'expérimentation de l'expérience précédente en section 2.3.1 sont reprises. Toutefois, dans cette nouvelle expérience, pour chacun des mots à reconnaître issus d'un signal de parole bruité, le signal de parole propre correspondant à l'enregistrement de ce mot n'est pas présent dans la base de référence. Ainsi, l'expérience porte sur une reconnaissance de mots isolés bruités dans un environnement mono-locuteur. Dans ce cas, les seuls mots admis pour l'évaluation sont ceux prononcés plusieurs fois dans la base de données d'origine. En effet, cette contrainte est nécessaire pour assurer la présence d'au moins une prononciation du mot à reconnaître dans la base de référence lorsque l'enregistrement de ce mot est retiré de la base.

En se basant sur les ensembles de données précédentes en section 2.3.1, les séries de données utilisées pour cette expérience sont ainsi constituées par :

- la série $S_1$ formée par tous les mots de la base de données (3275 mots prononcés dont 1891 mots ayant au moins une seconde prononciation),
- la série $S_2$ formée par les mots de deux lettres et plus (3167 mots prononcés dont 1783 mots ayant au moins une seconde prononciation),
- la série $S_3$ formée par les mots de trois lettres et plus (2831 mots prononcés dont 1450 mots ayant au moins une seconde prononciation).

Les séries de données sont identiques à l'expérience précédente en section 2.3.1. Seul le signal de parole propre correspondant au mot recherché est retiré de la base lors de l'évaluation de sa reconnaissance par le système. Le même type de bruit blanc gaussien que celui utilisé dans l'expérience précédente en section 2.3.1 est appliqué au signal de parole des mots à reconnaître. Les évaluations portent sur la présence du mot recherché dans la liste des

résultats retournés, similairement à la première expérience. Les résultats obtenus sont fournis dans un tableau récapitulatif pour des valeurs de SNR à 20 dB et à 0 dB (Tableau 2.2).

| performance % / données | SNR | sans option de recherche | | option de recherche 1-bit | |
|---|---|---|---|---|---|
| | | meilleur | 10 meilleurs | meilleur | 10 meilleurs |
| série $S_1$ (1891 mots) | 20 dB | 86,8 | 96,7 | 82,1 | 97,0 |
| | 0 dB | 42,3 | 90,5 | 29,0 | 85,7 |
| série $S_2$ (1783 mots) | 20 dB | 88,21 | 97,1 | 83,8 | 96,7 |
| | 0 dB | 43,8 | 90,8 | 30,2 | 85,0 |
| série $S_3$ (1450 mots) | 20 dB | 88,6 | 97,6 | 84,6 | 96,2 |
| | 0 dB | 47,2 | 91,1 | 33,1 | 85,5 |

Tableau 2.2 : Résultats de la reconnaissance de mots isolés bruités par empreinte de Vasiloglou (KED-TIMIT)

Bien que les résultats soient inférieurs à ceux obtenus dans les conditions similaires à la première expérience, les remarques effectuées sur les résultats précédents en section 2.3.1 sont de nouveaux valables (section 2.3.2, page 35). Le taux de présence du mot recherché dans la liste des 10 meilleurs scores reste cependant élevé, au dessus de 96 % lorsque le SNR est à 20 dB et au dessus de 85 % pour un SNR à 0 dB. Ces derniers résultats sont encourageants pour poursuivre l'évaluation du système au sein d'un environnement multi-locuteur. Une nouvelle expérience est donc mise en place pour une reconnaissance de phonèmes isolés dont les locuteurs multiples de la base d'apprentissage sont différents de ceux de la base de test.

### 2.3.3 Reconnaissance de phonèmes isolés en environnement multi-locuteur

Dans cette troisième expérience, le système initialement proposé par Vasiloglou est évalué pour une tâche de reconnaissance de phonèmes isolés dans un environnement multi-locuteur. Dans ce cas, seuls les meilleurs scores de la liste triés sont retournés comme résultats par le système durant son évaluation. Pour chaque empreinte à identifier, plusieurs phonèmes distincts peuvent être proposés en résultat lorsque la procédure d'identification du système détermine plusieurs empreintes de meilleur score.

L'expérience décrite consiste alors à évaluer la performance de diverses méthodes de création d'empreinte à partir du système d'identification audio par empreinte présenté par Vasiloglou. A cet effet, cinq méthodes de création d'empreinte sont alors définies par modifications successives :

i) méthode de Vasiloglou sur 32 bits (EV32),
ii) méthode EV32 avec une paramétrisation adaptée pour la parole (AP32),
iii) méthode AP32 avec une sous-empreinte définie sur 16 bits (AP16),
iv) méthode AP32 avec conservation de la séquence de sous-empreintes (APSL),
v) méthode APSL avec un seuil modifié pour le calcul de sous-empreinte (APSM).

Cette troisième expérience porte donc sur l'évaluation du système développé autour de cinq méthodes de création d'empreinte pour la reconnaissance de phonèmes isolés. Ces différentes méthodes conçues par modifications successives de la création de l'empreinte sont décrites de manière détaillées ci-dessous. Puis un tableau récapitulatif recense les différentes caractéristiques particulières à chacune de ces méthodes (Tableau 2.3).

La méthode EV32 représente la méthode originale de création d'empreinte présentée par Vasiloglou. Dans la méthode AP32, cette méthode originale EV32 est modifiée par le choix de variables de paramétrisation acoustique reconnues adaptées pour la RAP. Il s'agit de calculer les sous-empreintes tous les 10 ms à partir de fenêtres d'échantillons du signal de parole d'une durée de 25 ms. Ce choix est justifié par la volonté de discriminer le signal de parole stationnaire tout en conservant une représentation des variations locales du signal. De surcroît, le calcul d'une sous-empreinte est effectué sur une analyse spectrale d'une plage de fréquences allant de 40 Hz à 3700 Hz. Ce choix est justifié par la plage de fréquences de production de la parole.

Dans la méthode AP16, la seconde méthode AP32 est modifiée en réduisant la taille d'une sous-empreinte de 32 bits à 16 bits. L'effet d'une telle réduction est double. D'une part, les sous-bandes fréquentielles du spectre utilisées pour les calculs d'énergie sont plus larges. D'autre part, l'espace de recherche pour l'identification d'empreintes à une distance donnée les unes des autres est réduit. Dans la base de référence, la possibilité de rencontrer des empreintes ayant des sous-empreintes identiques est plus élevée lorsque ces sous-empreintes sont définies sur 16 bits et non plus sur 32 bits.

Dans la méthode APSL, la troisième méthode AP16 est modifiée pour conserver linéairement l'intégralité de la séquence de sous-empreintes qui compose le segment temporel correspondant à un phonème donné. Ainsi, les valeurs de sous-empreinte sont possiblement redondantes et sont placées au sein de l'empreinte selon l'ordre d'apparition temporelle des trames du signal de parole. Dans ce cas, la mesure de distance initiale entre empreintes définie par Vasiloglou est modifiée au profit d'une mesure de BER selon une méthode similaire au *Perceptual Audio Hashing* [Haitsma et al., 2001]. Cette mesure de distance de BER est basée sur la distance locale de Hamming. Comme les empreintes peuvent être de taille différente, cette mesure de distance est effectuée linéairement sur chacune des sous-empreintes en correspondance, du début de chacune des deux empreintes à comparer jusqu'à la fin de la plus petite des deux empreintes. Afin de tenir compte de la différence de taille entre empreintes comparées, la distance obtenue est ensuite pondérée par le ratio de la plus grande des deux empreintes par rapport à la plus petite. Dans ce cas, la mesure du meilleur score correspond à la distance minimale résultante.

Dans la dernière méthode APSM, la quatrième méthode APSL est modifiée en simplifiant le calcul de la sous-empreinte. De surcroît, le seuil statique sur l'énergie proposé par Vasiloglou à l'équation est ici remplacé par un seuil de distance relative. Compte-tenu de l'énergie $E_{t,i}$ de la sous-bande $i$ de la trame $t$, le bit de valeur $B_{t,i}$ d'indice $i$ de la sous-empreinte $B_t$ à calculer est alors défini comme :

$$B_{t,i} = \begin{cases} 1 \; si \; E_{t+1,i+1} > T'.E_{t,i} \\ 0 \; sinon \end{cases} \tag{2.2}$$

Le seuil $T'$ définit ici la capacité à activer le bit d'indice $i$ de la sous-empreinte calculée sur la trame $t$ lorsque les valeurs d'énergie $E_{t+1,i+1}$ et $E_{t,i}$ sont proches. Ce seuil $T'$ permet de conserver une continuité des valeurs binaires d'un vecteur au suivant lorsque les sous-bandes fréquentielles contigües ont des valeurs proches d'une trame à la suivante. Pour cette évaluation, ce seuil est empiriquement fixé à $T' = 0{,}98$ afin de maximiser les performances sur le sous-ensemble de développement.

<table>
<tr><th>méthode</th><th>paramétrisation</th><th>taille de sous-empreinte</th><th>création d'empreinte</th><th>calcul binaire de sous-empreinte</th></tr>
<tr><td>EV32</td><td>trame de 18,75 ms calculée tous les 0,625 ms<br><br>plage de fréquence de 100 à 2000 Hz</td><td rowspan="2">32 bits</td><td rowspan="3">valeurs uniques triées par ordre croissant</td><td rowspan="3">$1\ si\ E_{t+1,i+1} - E_{t,i} > 0$<br><br>$0\ si\ E_{t+1,i+1} < T$<br>ET $E_{t,i} < T$</td></tr>
<tr><td>AP32</td><td rowspan="4">trame de 25 ms calculée tous les 10 ms<br><br>plage de fréquence de 40 à 3700 Hz</td></tr>
<tr><td>AP16</td><td rowspan="3">16 bits</td></tr>
<tr><td>APSL</td><td rowspan="2">séquence linéaire temporelle</td><td rowspan="2">$1\ si\ E_{t+1,i+1} > T'.E_{t,i}$</td></tr>
<tr><td>APSM</td></tr>
</table>

$T$ : seuil sur la valeur d'énergie
$T'$ : coefficient pondérateur

Tableau 2.3 : Caractéristiques des différentes méthodes de création d'empreinte

Pour l'évaluation de ces cinq méthodes de création d'empreinte audio, la base de données TIMIT [Fischer et al., 1986] est choisie (Annexe B, page 135). Cette base de données est plus complète que KED-TIMIT et fournit des phrases prononcées par de multiples locuteurs. L'ensemble phonétique choisi est composé de 39 phonèmes, suivant la définition du CMU/MIT [Lee et al., 1989b]. Cette base de données est segmentée en trois ensembles disjoints pour i) l'apprentissage (4 heures, 462 locuteurs), ii) le développement (30 minutes, 56 locuteurs) et iii) le test (1 heure, 112 locuteurs). Les locuteurs présents dans chacun de ces ensembles sont distincts. La phase de développement permet d'adapter les différents paramètres variables du système pour la définition des différentes méthodes de création d'empreinte audio. La phase de test final permet alors de mesurer les résultats obtenus en fonction des cinq méthodes de création d'empreinte audio définies.

Durant l'évaluation, les zones de silence ne sont pas prises en compte. L'évaluation porte sur la mesure de précision moyenne obtenue pour les résultats retournés. Cette précision moyenne est définie comme la moyenne des précisions locales dans la reconnaissance de chacune des empreintes à identifier. La précision locale est obtenue pour chaque empreinte à identifier par le ratio du nombre de bons phonèmes parmi l'ensemble des phonèmes retournés. L'information du nombre moyen de phonèmes retournés est également fournie (Tableau 2.4).

| méthode | développement | | test | |
|---|---|---|---|---|
| | **précision moyenne (%)** | **nombre moyen de résultats** | **précision moyenne (%)** | **nombre moyen de résultats** |
| EV32 | 6,1 | 15,0 | 6,1 | 15,1 |
| AP32 | 19,3 | 13,4 | 19,9 | 13,1 |
| AP16 | 28,5 | 12,3 | 28,7 | 12,2 |
| APSM | 28,1 | 1,6 | 28,6 | 1,7 |
| APSL | 28,3 | 1,6 | 28,7 | 1,7 |

Tableau 2.4 : Résultats de la reconnaissance de phonèmes isolés selon les méthodes de création d'empreinte (TIMIT)

Les performances du système pour la reconnaissance de phonèmes isolés en environnement multi-locuteur sont très mauvaises avec les empreintes de Vasiloglou EV32 avec une précision de 6 %. Avec les empreintes issues de la méthode AP32, une nette amélioration de ces performances est perçue lorsque l'ensemble de la plage de fréquences de production de la parole est prise en compte, atteignant presque une précision de 20 %. Dans ce cas, les sous-empreintes sont calculées en représentant une plus grande étendue du signal utile de parole. Le choix de ne produire que 100 sous-empreintes par seconde permet par ailleurs de limiter la redondance de l'information utile disponible lors du calcul d'une sous-empreinte par rapport à la précédente. Ainsi, le nombre de sous-empreintes contigües identiques ou fortement similaires est réduit. Une autre conséquence est que la taille des empreintes est alors diminuée par une diminution du nombre de sous-empreintes fortement similaires.

Avec les empreintes issues de la méthode AP16, une amélioration des performances du système est de nouveau présente avec une précision dépassant 28 %. Dans ce cas, la réduction de la taille d'une sous-empreinte sur 16 bits permet de diminuer la résolution de la représentation acoustique du signal de parole. Enfin, avec les empreintes issues des méthodes APSM et APSL, le mode de création d'empreinte spécifique à Vasiloglou est délaissé au profit d'une représentation temporelle linéaire du signal de parole. Cette représentation est proche de la méthode initialement proposée par Philips (section 1.3.2, page 19). Les résultats obtenus en précision sont alors sensiblement équivalents de ceux obtenus avec les empreintes issues de la méthode AP16 tandis que la complexité de création d'empreinte est moins élevée.

En effet, dans ces empreintes issues des méthodes APSM et APSL, les étapes de tri par ordre croissant de valeur des sous-empreintes et de suppression des doublons ne sont alors plus effectuées. Enfin, le nombre moyen de résultats retournés par phonème à reconnaître diminue fortement alors que ce nombre moyen était relativement stable pour les trois premières méthodes de création d'empreinte. Ce nombre moyen est situé pour ces trois premières méthodes autour de 12 à 15 phonèmes retournés comme résultats par phonème à reconnaître. Désormais autour de 1,5 en moyenne pour ces deux dernières méthodes de création d'empreinte, la forte diminution de ce nombre moyen montre une grande différence de la taille de la liste des résultats retournés par le système.

## 2.4 Discussions

Un système similaire à celui proposé par Vasiloglou a été développé [Vasiloglou et al., 2004]. A partir de ce système, nous avons défini plusieurs expériences. Dans un premier temps, nous avons reproduit l'expérience présentée dans l'article de référence. Cette expérience a porté sur la reconnaissance d'enregistrements bruités de mots isolés. Les performances du système développé sont de l'ordre de grandeur de celles attendues. Nous n'avons cependant pas été en mesure d'atteindre les résultats présentés dans le document de référence. Dans un second temps, l'expérience initiale est étendue à la reconnaissance de mots isolés en environnement mono-locuteur. Bien que les résultats soient en retrait par rapport à la première expérience, un tel système s'avère adapté à ce type de tâche de RAP. Cependant, nous pensons que la base de données initialement utilisée pour les expériences n'est pas idéalement adaptée à la reconnaissance de mots isolés. En effet, KED TIMIT est une base de données de parole continue [CSTR, 2001]. Dans le cas d'une reconnaissance en mots, il est alors nécessaire de tenir compte des approximations de la segmentation fournie par la transcription et des variations d'élocution liées aux enchaînements des mots les uns à la suite des autres.

De surcroît, parmi les types de dégradation proposés par Vasiloglou pour ses expériences, nous n'avons pas traité des effets de variation temporelle. La variation temporelle artificielle appliquée au signal de parole reflète dans ce cas précis une dégradation acoustique du signal d'origine liée uniquement à la durée des évènements sonores qui le composent. Cette dégradation n'exprime pas la notion de variation de la vitesse d'élocution. Dans le système présenté par Vasiloglou, les sous-empreintes contigües sont fortement similaires grâce à un taux de recouvrement élevé de 29/30. Ainsi, comme seule une occurrence de chaque sous-empreinte est conservée dans l'empreinte concernée, peu importe le nombre de répétitions de cette sous-empreinte dans le signal de parole analysé. En effet, pour la création de l'empreinte, il s'agit à partir du signal de parole analysé de conserver au moins une occurrence de chaque valeur de sous-empreinte. La variation temporelle indiquée

permet donc d'augmenter ou de diminuer le nombre de sous-empreintes contigües similaires représentant les variations d'énergies en sous-bandes fréquentielles.

Dans un troisième temps, le système est évalué dans une tâche de reconnaissance de phonèmes isolés en environnement multi-locuteur. Les premiers résultats obtenus à partir des empreintes de Vasiloglou sont très mauvais. Les performances du système sont améliorées en modifiant successivement :

- la paramétrisation acoustique pour une adaptation au signal utile de parole,
- la taille de sous-empreinte pour une réduction de l'espace de recherche,
- la création de l'empreinte pour une conservation de l'information temporelle,
- le calcul de la sous-empreinte pour le remplacement du seuil fixe sur la valeur d'énergie vers un seuil relatif sur les variations énergétiques.

Ces dernières performances sont maintenues en simplifiant le mode de création d'une empreinte pour une conservation des séquences linéaires de sous-empreintes selon leur ordre d'apparition temporelle. Enfin, le mode de calcul d'une sous-empreinte est simplifié avec l'ajout d'un seuil relatif sur les valeurs d'énergie pour assurer une meilleure continuité des valeurs binaires lors du calcul d'une sous-empreinte à la suivante.

Le système proposé par Vasiloglou est un travail original et une première approche d'adaptation de l'identification audio par empreinte à la reconnaissance de la parole. Nous décidons de reprendre ce travail d'adaptation pour concevoir notre propre système d'identification audio par empreinte dédié à une tâche particulière de la RAP. Ce système original est alors évalué sur différents méthodes de calcul de sous-empreinte dans une tâche de décodage acoustico-phonétique (DAP) en parole continue et en mode multi-locuteur.

# Chapitre 3. Système d'identification audio pour le DAP

Le travail présenté par Vasiloglou [Vasiloglou et al., 2004] est une première approche d'adaptation d'un système d'identification audio par empreinte à la RAP en mots isolés dans un environnement mono-locuteur. Nous avons adapté ce système pour répondre aux tâches de reconnaissance de mots isolés en environnement mono-locuteur et de reconnaissance de phonèmes isolés en environnement multi-locuteur. Nous désirons désormais savoir si un tel système est adaptable au cadre de la RAP en parole continue dans un environnement multi-locuteur. Nous développons alors notre propre système d'identification audio par empreinte dédié à cette tâche de RAP. Notre système possède plusieurs différences majeures avec celui proposé par Vasiloglou. Désormais, autant la base de référence pour l'apprentissage que le signal de parole à analyser pour la reconnaissance sont constitués de parole continue. Dans ce cas, l'information de segmentation de début et fin de l'empreinte n'est pas connue. Il est alors difficile de reprendre les méthodes de création d'empreinte précédemment évaluées (Chapitre 2, page 27). Une conséquence immédiate de cette absence de segmentation est que la procédure d'identification et la mesure de distance entre empreintes de Vasiloglou sont beaucoup plus complexes. En effet, dans le système proposé par Vasiloglou, les principes de distance entre empreintes sont basés sur une taille d'empreinte connue et finie.

Afin de comparer les performances du système développé avec celles d'un système de RAP issu de l'état de l'art autour d'une tâche commune, l'application choisie est un décodage acoustico-phonétique (DAP). L'évaluation porte sur les performances du système en sortie du module de comparaison, sans intervention d'un modèle de langage. Etant donné les bonnes performances du système d'identification audio par empreinte développé par Philips à la tâche de détection d'extraits de musique [Haitsma et al., 2002], un système similaire est mis en œuvre pour notre application. Dans notre système d'identification audio par empreinte

appliquée à la tâche de DAP, une représentation de l'intégralité de la base de données de parole segmentée utilisée comme base de référence est conservée. Les données complémentaires associées à la représentation du signal de parole dans la base de référence sont des unités linguistiques correspondant aux phonèmes.

Le principe d'identification de notre système est basé sur la concaténation de ces phonèmes en sortie du système pour déterminer les phonèmes prononcés dans un flux de parole continue. L'étape d'identification correspond alors au décodage d'un flux de parole de test. Durant cette étape d'identification, la recherche de plus longs segments associant sous-empreintes de signal de parole d'apprentissage et sous-empreintes de signal de parole de test permet de retourner des ensembles de phonèmes. La séquence de ces ensembles de phonèmes retournés permet ensuite de déterminer les phonèmes prononcés.

## 3.1 Principe général

Le principe général de ce système de DAP est similaire au principe d'identification audio par empreinte [Cano et al., 2005]. Dans notre cas, le signal audio est le signal acoustique de parole et les données associées sont les phonèmes. Compte-tenu du nombre limité de phonèmes possibles, un dictionnaire intermédiaire permet de coder les phonèmes sous la forme de numéros d'index lors de leur association avec les segments de sous-empreintes. La base de données utilisée par le système, désignée par base de référence, contient alors ce dictionnaire, les données de référence pour l'apprentissage et une méthode d'accès rapide aux données. Le module de calcul des sous-empreintes permet de traiter le signal de parole afin d'en déterminer les caractéristiques de représentation du signal. Le module de comparaison permet de procéder à l'identification phonétique du signal de parole à analyser (Figure 3.1).

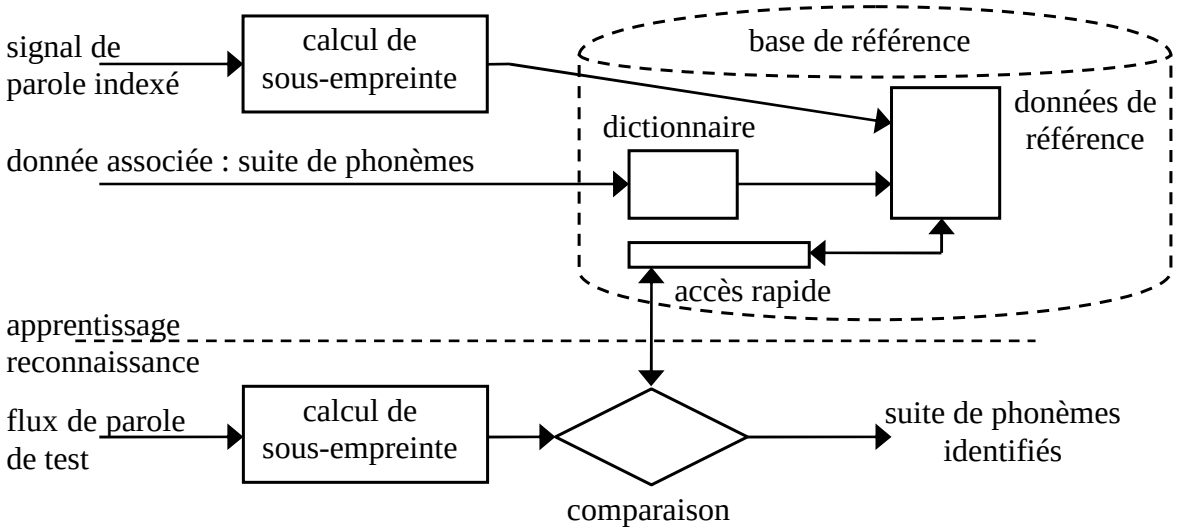



Figure 3.1 : Identification audio par empreinte adaptée au DAP

Dans le système d'identification audio par empreinte que nous proposons pour le DAP, le principe d'utilisation d'une base de référence est conservé. Ainsi, une représentation compacte de la base de données de parole phonétiquement indexée est conservée en mémoire dans une base de référence. Cette base de référence permet d'associer une représentation du signal acoustique et sa donnée complémentaire sous la forme d'une étiquette linguistique, ici le phonème. Durant la phase de reconnaissance correspondant au DAP, un évènement acoustique similaire au signal de parole de test est recherché à travers la base de référence. Le système effectue un appariement efficace entre le signal de parole de test et les données d'apprentissage présentes dans la base de référence. Selon un critère de similarité défini, les trames du signal de parole représentées sous la forme de sous-empreintes sont comparées deux à deux entre le signal de parole de test et celui de la base de référence. Ainsi, le segment de sous-empreintes de la base de référence dont la distance est la plus proche du segment correspondant dans ce signal de parole de test permet de retourner son/ses donnée(s) phonétique(s) associée(s) (Figure 3.2).

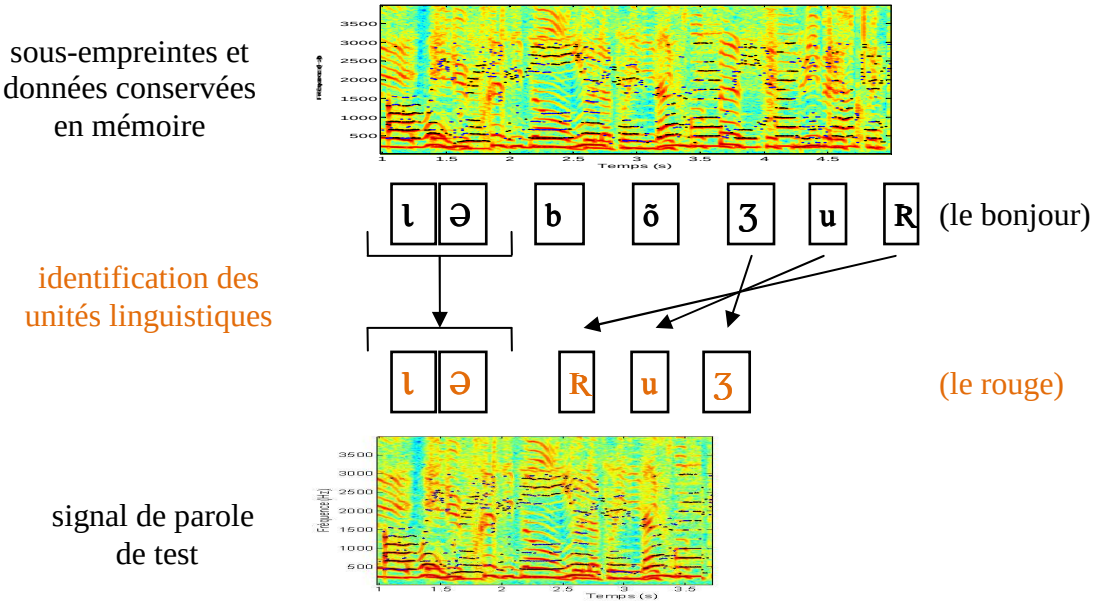



Figure 3.2 : Principe d'identification par empreinte appliquée au DAP

La notion d'empreinte est relativement similaire à celle utilisée en identification audio par empreinte. Pourtant, la production du signal de parole et la représentation du message linguistique induisent certaines caractéristiques originales dans la création d'empreinte adaptées. En effet, une empreinte est ici une séquence dynamiquement créée de sous-empreintes concaténées. Dans le système développé, le début et la fin d'une empreinte sont donnés par ceux du phonème correspondant fourni par la segmentation de la transcription. Ainsi, la taille d'une empreinte est variable. Par ailleurs, dans le cas d'une représentation phonétique, l'ensemble des mots du vocabulaire pouvant être prononcés pour former le message linguistique est représenté par la concaténation d'un nombre limité de phonèmes. Ainsi, on admet une certaine redondance de l'information phonétique disponible dans la base de référence. Durant la phase d'indexation pendant l'apprentissage, ces sous-empreintes et les phonèmes associés fournis par la segmentation sont stockés dans la base de référence. Ainsi, quelque soit le nombre de prononciations d'un phonème donné, toutes les séquences de sous-empreintes correspondantes sont conservées. Donc chaque phonème est associé à une ou plusieurs séquences de sous-empreintes dans la base de référence. Durant la phase d'identification pendant la reconnaissance, les empreintes de test sont dynamiquement générées. Ces empreintes représentent les séquences de sous-empreintes dont les débuts et fins sont obtenus par la différence temporelle entre la position de la sous-empreinte appariée au sein de l'empreinte issue de l'apprentissage et les débuts et fins de cette dernière. De surcroît, il est nécessaire d'adapter les étapes de calcul et de comparaison des sous-empreintes

issues de l'analyse du signal de parole selon le type de représentation des paramètres acoustiques et la mesure de distance associée.

## 3.2 Relâchement de la contrainte d'identification

Dans la méthode de Philips [Haitsma et al., 2001], lors de l'étape d'identification d'un signal de parole de test, seuls sont considérés pour la comparaison les segments de sous-empreintes de la base de référence possédant au moins une sous-empreinte identique au segment du signal audio de test. Cette contrainte permet de grandement restreindre le nombre de segments de sous-empreintes à comparer afin de garantir un coût de calcul efficace vis-à-vis d'une comparaison exhaustive de l'ensemble des segments possibles dans la base de référence. En contrepartie, cette contrainte est parfois trop discriminante lorsque les sous-empreintes sont faiblement représentées dans la base de référence. Cette forte discrimination est rencontrée lors d'une quantité limitée de données disponibles pour l'indexation et lors du choix d'une grande taille de sous-empreinte. Par exemple, la dimension d'une sous-empreinte définie sur 32 bits s'étend sur plus de 4 milliards de valeurs différentes. Une telle discrimination a alors pour contrainte de ne pas atteindre des segments de sous-empreintes de la base de référence fortement similaires à ceux issus du signal de parole de test mais dont aucune sous-empreinte n'est identique.

Dans notre système, nous proposons de relâcher la contrainte d'identification sur la sous-empreinte permettant l'appariement entre les segments de vecteurs du signal de parole de test et ceux présents dans la base de référence. Ce relâchement de la contrainte d'identification consiste à élargir le critère d'acceptation pour l'appariement aux sous-empreintes dont la distance de Hamming est inférieure au seuil sur le BER fixé par le système. Tout comme la méthode de Philips, la base de référence possède un tableau d'accès rapide. Pour chaque valeur de sous-empreinte, ce tableau indexe les différentes positions de ce vecteur dans la base de référence. Ainsi, lors de la recherche d'une sous-empreinte donnée, il s'agit d'élargir cette recherche à la liste des valeurs de sous-empreinte dont la distance à cette sous-empreinte recherchée est inférieure à une distance de Hamming donnée.

Nous considérons que la recherche de l'ensemble des valeurs binaires à une certaine distance d'une valeur de référence revient à construire un arbre de recherche. Les nœuds de cet arbre sont des valeurs binaires. Les arêtes entre les nœuds sont construites si nécessaires et permettent de relier les nœuds ayant une distance de Hamming unitaire entre eux. Cette construction dans l'espace des distances de Hamming revient à développer un graphe sous la forme d'un cube partiel sans cycle avec la valeur binaire de référence comme racine et le seuil sur le BER comme distance maximale. Soit $A$ une valeur binaire définie sur $N$ bits et $A_i$ le bit

d'indice *i* de *A*. On considère alors *A* comme la racine de l'arbre. Etant donné le seuil *S* sur la distance de Hamming, l'arbre est construit itérativement en largeur d'abord pour une profondeur *p* de 1 à *S*. Soient *F* la liste des feuilles de l'arbre et *H* la liste des nœuds à retourner, l'algorithme permettant de trouver l'ensemble des valeurs binaires à une distance inférieure ou égale à *S* de la valeur *A* peut s'écrire selon le pseudo-code suivant :

```
Prérequis :      inverse(V, i)   retourne la valeur V avec le bit d'indice i inversé

Initialisation : F = H = {A}
                 flipped[A] = 0

Algorithme :   Pour ( p = 1 ; p <= S ; p = p + 1 )
                     F' = Ø
                     Pour chaque élément E de F
                          Pour ( i = flipped[E] + 1 ; i <= N ; i = i + 1 )
                               X = inverse(E, i)
                               F' = F' + {X}
                               flipped[X] = i
                     F = F'
                     H = H + F'
               Retourner H
```

Figure 3.3 : Pseudo-code de l'algorithme de relâchement de la contrainte d'identification

Au sein de cet algorithme, la liste *flipped* permet de conserver pour chaque noeud l'indice du bit inversé par rapport à la valeur de son père dans l'arbre. Ainsi, il s'agit de parcourir en largeur d'abord les valeurs existantes dans l'arbre à un certain niveau *p* afin de générer les feuilles correspondantes au niveau suivant *(p+1)*. Le calcul des valeurs binaires à une distance de Hamming donnée d'une valeur de référence dépend uniquement de la taille binaire de la valeur de référence et de la distance de Hamming maximale acceptée. Compte-tenu de la taille d'une valeur binaire définie sur $N$ bits et le seuil $S$ sur la distance de Hamming maximale autorisée, la complexité d'un tel algorithme est donc exprimée en $O(S,N) = S.\log^2(N)$. Cette complexité est réduite par rapport à celle d'une recherche linéaire au sein d'un arbre *N*-aire de profondeur *S* exprimée en $O(S,N) = N^S$. Le graphe ci-dessous représente l'arbre construit par cet algorithme, pour une valeur de référence 0 définie sur 4 bits avec un seuil sur la distance de Hamming fixé lui aussi à 4 bits (Figure 3.4).

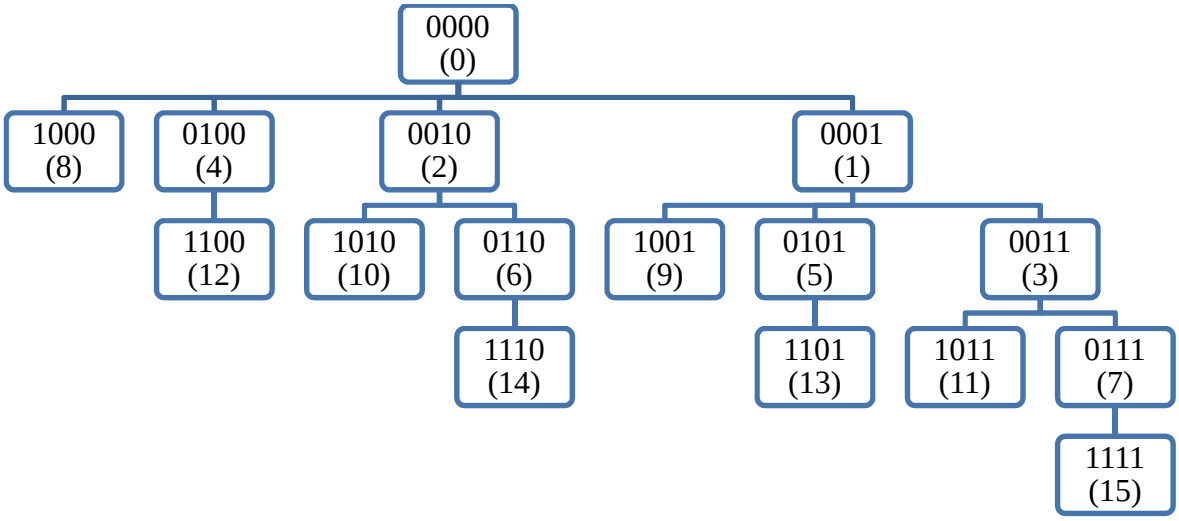



Figure 3.4 : Arbre de distance de Hamming de profondeur 4 bits pour la valeur 0

La liste des éléments retournés est la liste de l'ensemble des nœuds de l'arbre ainsi généré. Cet algorithme permet donc de calculer de manière efficace toutes les valeurs binaires à une distance de Hamming donnée d'une valeur binaire de référence. Afin d'optimiser l'utilisation de cet algorithme dans le programme développé, une seule liste est calculée lors de l'initialisation du système. En se basant sur la taille d'une sous-empreinte, cette liste contient l'ensemble des valeurs binaires à une distance maximale de Hamming donnée par le seuil de BER par rapport à la valeur de référence 0. Ensuite, une opération OU-exclusif est effectuée entre chacune des valeurs binaires de cette liste initiale et la sous-empreinte localement considérée pour l'identification. Cette opération permet de retourner la liste des valeurs binaires utilisées pour l'appariement dans la recherche des empreintes à comparer issues de la base de référence.

## 3.3 Vecteur de représentation acoustique

Dans cette étude, une sous-empreinte basée sur des paramètres issus de l'identification audio par empreinte est comparée à différentes formes de sous-empreinte basées sur des paramètres acoustiques de type MFCC [Bogert et al., 1963]. Les sous-empreintes sont calculées pour chaque trame du signal de parole à partir des paramètres acoustiques. Dans ce cas, une fenêtre de Hamming [Enochson et al., 1968] est appliquée toutes les 10 ms sur une trame du signal de parole d'une longueur de 25 ms. A ce moment, 100 trames par secondes sont calculées, assurant pour chacune d'entre elles un taux de recouvrement de 60 % de la précédente trame. A partir de l'analyse du spectre à court terme (*Fast Fourier Transform*, FFT) [Cooley et al.,

1965], diverses méthodes de calcul de sous-empreinte sont proposés. Cinq méthodes de calcul de sous-empreinte sont alors définies comme :

- sous-empreinte adaptée de l'identification audio par empreinte (APSM),
- sous-empreinte à base de MFCC et quantification vectorielle (QV),
- sous-empreinte à base de MFCC et quantification supervisée uniforme (QU),
- sous-empreinte à base de MFCC et quantification supervisée non-uniforme (QN),
- sous-empreinte à base de MFCC et les deux plus proches clusters (2C).

Dans les méthodes de calcul de sous-empreinte à base de paramètres acoustiques MFCCs, la sous-empreinte résultante conserve une forme de représentation acoustique du signal de parole dans un espace réel multidimensionnel.

### 3.3.1 Sous-empreinte adaptée de l'identification audio par empreinte (APSM)

Dans l'évaluation précédente de reconnaissance de phonèmes isolés (section 2.3.3, page 36), la méthode APSM est celle retournant les empreintes avec les meilleures performances. Cette méthode est alors reprise pour son calcul de sous-empreinte. Dans cette méthode, le calcul de sous-empreinte ne dépend que de la taille de la sous-empreinte. En effet, une sous-empreinte est ici obtenue à partir des valeurs d'énergie en sous-bandes fréquentielles du spectre de la trame courante et de celles du spectre de la trame juste précédente. Compte-tenu de la taille d'une sous-empreinte définie sur $N$ bits, la complexité d'une telle méthode de calcul de sous-empreinte est alors exprimée en $O(N)$.

### 3.3.2 Sous-empreinte à base de MFCC et quantification vectorielle (QV)

Chaque trame du signal de parole en entrée est représentée par un vecteur acoustique à 39 dimensions. Le vecteur est constitué des 12 premiers coefficients MFCCs plus $C(0)$ comme composante d'énergie de la trame et leurs dérivées premières et secondes. Une quantification vectorielle non-supervisée par segmentation suivant le principe des k-moyennes est utilisée. Cette quantification est appliquée sur chaque vecteur acoustique $x$ avec son plus proche centroïde $C$ de moyenne et variance $(\mu, \sigma^2)$. La distance $d(x, C)$ entre le vecteur $x$ à 39 dimensions et le centroïde $C$ est basée sur le logarithme d'une fonction de densité de probabilité $D(x, C)$ (*Probability Density Function*, PDF) [Young et al., 2006] telle que :

$$\ln D(x, C) = -\frac{1}{2}\left[\sum_{j=1}^{39} \frac{(x_j - \mu_j)^2}{\sigma_j^2} + \ln\left((2\pi)^{39} \prod_{j=1}^{39} \sigma_j^2\right)\right] \qquad (3.1)$$

Dans le cas présent, les centroïdes sont indépendants les uns des autres. On considère alors que la variance est ici locale au centroïde associé au vecteur considéré. Donc en supprimant les coefficients pondérateurs et les termes constants, nous choisissons d'utiliser la distance simplifiée $d(x, C)$ suivante :

$$d(x, C) = \sum_{j=1}^{39} \left( \frac{(x_j - \mu_j)^2}{\sigma_j^2} - \ln \frac{1}{\sigma_j^2} \right) \tag{3.2}$$

Le premier terme de cette équation (3.2) correspond à une mesure de distance de Mahalanobis entre le vecteur $x$ et le centroïde $C$ de moyenne et variance $(\mu, \sigma^2)$ [Mahalanobis, 1936]. En effet, on peut exprimer cette distance de Mahalanobis sous la forme d'une distance euclidienne centrée normalisée $d_M(x, C)$ :

$$d_M(x, C) = \sqrt{\sum_{j=1}^{39} \frac{(x_j - \mu_j)^2}{\sigma_j^2}} \tag{3.3}$$

Donc compte-tenu de l'expression de cette distance de Mahalanobis $d_M(x, C)$, notre distance de PDF simplifiée $d(x, C)$ peut alors s'écrire ainsi :

$$d(x, C) = d_M(x, C)^2 - \sum_{j=1}^{39} \ln \frac{1}{\sigma_j^2} \tag{3.4}$$

Le second terme de cette équation (3.2), présent également dans l'équation (3.5), est une constante associée au centroïde. Ce terme est un reliquat de la valeur du déterminant issu de la PDF [Young et al., 2006]. A des fins d'optimisation, ce terme peut être préalablement calculé pour chacun des centroïdes concerné. Ce terme ainsi calculé sera alors réutilisé lors de l'utilisation de notre mesure de distance $d(x, C)$ au sein de l'équation (3.2).

Durant la segmentation par k-moyennes, les centroïdes sont initialisés par une répartition équilibrée des vecteurs acoustiques suivant leur distribution le long de l'axe du coefficient $C(0)$. Le nombre de centroïdes correspond au nombre de valeurs possibles d'une sous-empreinte. Afin d'assurer un rapport de représentativité homogène des centroïdes lors de leur construction, un équilibrage de leur poids est appliqué entre chaque itération de la quantification. Le poids d'un centroïde est donné par le nombre de vecteurs utilisés pour le

calcul de ce centroïde. Cet équilibrage est effectué en supprimant tout d'abord les centroïdes de faible poids. Puis pour chaque centroïde supprimé, les centroïdes de poids fort sont divisés en deux classes distinctes et de taille égale. Les centroïdes obtenus par cette quantification sont ensuite conservés dans un tableau. Puis une approximation par recherche du plus proche voisin est effectuée en utilisant l'équation précédente (3.2) afin de trouver le plus proche centroïde de chaque vecteur acoustique. Une valeur de sous-empreinte est alors obtenue pour chaque vecteur acoustique en retournant la valeur d'index de son plus proche centroïde dans le tableau des centroïdes. Considérant ainsi le nombre de centroïdes calculés en relation avec la taille d'une sous-empreinte définie sur $N$ bits, la complexité d'une telle méthode de calcul de sous-empreinte est exprimée en $O(2^N)$. En effet, une sous-empreinte est ici obtenue en mesurant la distance du vecteur acoustique correspondant à l'ensemble des centroïdes calculés.

### 3.3.3 Sous-empreinte à base de MFCC et quantification supervisée uniforme (QU)

Pour une meilleure classification des vecteurs acoustiques lors de la quantification, la segmentation fournie par la transcription phonétique des données de l'apprentissage est désormais utilisée. Les vecteurs acoustiques issus du signal de parole sont calculés de la même manière que pour la méthode QV. Cependant, la quantification suivant le principe des k-moyennes est maintenant effectuée pour chaque ensemble de vecteurs acoustiques associés à un phonème donné. Ainsi, la quantification de chacun de ces ensembles est obtenue en appliquant l'équation précédente (3.2) avec la même initialisation des centroïdes que dans la méthode QV. Le nombre de centroïdes associés à chacun des phonèmes dépend alors du nombre de valeurs possibles d'une sous-empreinte divisé par le nombre de phonèmes possibles.

Dans le cas présent, l'information phonétique n'est utilisée que pour le calcul des centroïdes. En effet, la classe phonétique des vecteurs acoustiques du signal de parole de test lors de la reconnaissance n'est pas connue. Afin d'appliquer le même principe de calcul de sous-empreinte quelque soit le signal de parole analysé entre l'apprentissage et le test, nous choisissons de ne pas exploiter l'information phonétique pour la représentation acoustique du signal. Les sous-empreintes sont donc obtenues de la même manière que dans la méthode QV. Comme dans la méthode précédente QV, compte-tenu de la taille d'une sous-empreinte définie sur $N$ bits, la complexité de la présente méthode de calcul de sous-empreinte est également exprimée en $O(2^N)$.

### 3.3.4 Sous-empreinte à base de MFCC et quantification supervisée non-uniforme (QN)

Afin d'étendre les possibilités d'identification, il est possible d'augmenter le nombre d'occurrences des sous-empreintes dans la base de référence. Cet objectif peut être atteint en appliquant une méthode de densification de l'espace de recherche. Dans ce but, à partir de chaque centroïde obtenu par la méthode QU, un algorithme d'expansion permet de retourner des grappes contenant un ou plusieurs centroïdes. Une grappe, appelée cluster, est ainsi construite en appliquant l'équation précédente (3.2) entre le vecteur $\mu$ de la moyenne d'un centroïde $C$ et tous les autres centroïdes. Dans ce cas, toutes les mesures de distance entre $C$ et tous les autres centroïdes sont conservées dans une liste avec leur centroïde correspondant. A partir de cette liste de distances triée par ordre croissant, un cluster est construit en sélectionnant tous les plus proches centroïdes de la liste de même phonème que $C$ jusqu'à atteindre un centroïde de phonème différent. A ce moment, une méthode d'édition est utilisée afin de fusionner les clusters ayant au moins un centroïde en commun. Finalement, les clusters ne contenant qu'un unique centroïde sont considérés non pertinents et sont supprimés, ainsi que le centroïde qu'ils contiennent (Figure 3.5).

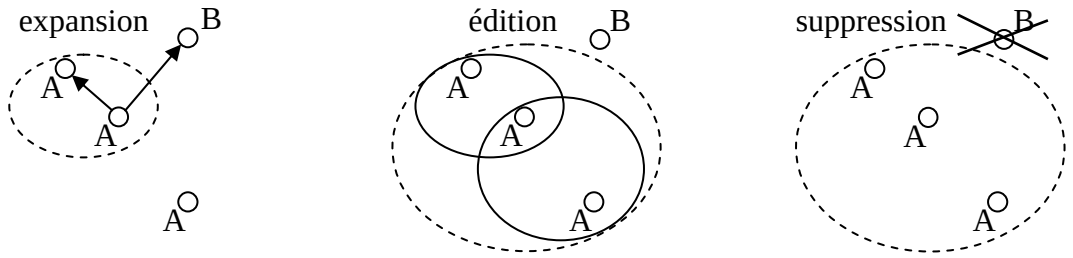



Figure 3.5 : Méthode de construction des clusters

Après l'ensemble des opérations d'expansion, d'édition et de suppression, un tableau des clusters permet de conserver les clusters restants et la liste de leurs centroïdes associés. Les centroïdes restants sont utilisés pour la recherche du plus proche voisin de chaque vecteur acoustique du signal de parole analysé en utilisant l'équation précédente (3.2). Une sous-empreinte est alors obtenue pour chaque vecteur acoustique en retournant la valeur d'index du cluster contenant le plus proche centroïde du vecteur. Considérant le nombre de centroïdes centroïdes restants et compte-tenu de la taille d'une sous-empreinte définie sur $N$ bits, la complexité de cette méthode de calcul de sous-empreinte est exprimée en $O(2^N)$. Une sous-empreinte est ici obtenue en mesurant la distance du vecteur acoustique correspondant à l'ensemble des centroïdes restants.

### 3.3.5 Sous-empreinte à base de MFCC quantifié avec les deux plus proches clusters (2C)

Afin de poursuivre la démarche de densification de l'espace de recherche, la taille des clusters considérés non pertinents est progressivement augmentée. Cet objectif peut être atteint en appliquant une méthode de condensation pour supprimer les petits clusters, peu représentatifs de l'ensemble des données d'apprentissage. A partir de la précédente méthode QN, la procédure de construction des clusters est appliquée de manière récursive sur les centroïdes conservés tout en incrémentant la taille des clusters à supprimer. Cette méthode de condensation s'exécute tant que les clusters restants représentent tous les phonèmes. A la fin de ce processus, les listes des centroïdes et clusters au début de la dernière itération sont retournées, c'est-à-dire lorsque tous les phonèmes sont encore représentés par les centroïdes conservés.

Afin de représenter l'espace de recherche ainsi densifié de manière plus précise, les sous-empreintes sont obtenues à partir des valeurs d'index des deux premiers clusters distincts contenant leurs plus proches centroïdes. Tout d'abord, la même mesure de distance définie par l'équation précédente (3.2) est appliquée entre les vecteurs et les centroïdes conservés. Puis à partir du tableau des clusters conservés et la liste de leurs centroïdes associés, les valeurs d'index des deux premiers clusters distincts contenant les centroïdes les plus proches de chaque vecteur sont utilisés pour définir la sous-empreinte correspondante. Concrètement, le centroïde le plus proche du vecteur acoustique sera utile pour déterminer le plus proche cluster. La liste des centroïdes les plus proches du vecteur sera ensuite parcourue jusqu'à atteindre le premier centroïde associé à un cluster différent. Ce centroïde sera alors utile pour déterminer le second plus proche cluster du vecteur acoustique. Ainsi, si par exemple 200 clusters sont conservés par cette méthode de condensation, le vecteur sera défini sur 16 bits. Les 8 bits de poids fort du vecteur représenteront alors la valeur d'index du plus proche cluster. Les 8 bits de poids faible représenteront quant à eux la valeur d'index du second plus proche cluster. Considérant le nombre de centroïdes centroïdes restants et compte-tenu de la taille d'une sous-empreinte définie sur $N$ bits, la complexité de cette dernière méthode de calcul de sous-empreinte est également exprimée en $O(2^N)$ pour les mêmes raisons que dans la méthode QN.

## 3.4 Empreinte adaptée au DAP et meilleure séquence phonétique

Dans le présent système, les empreintes sont des phonèmes représentés par des segments de taille variable de sous-empreintes dans la base de référence. Ces empreintes sont conservées séquentiellement dans la base de référence, selon l'ordre de leur moment d'apparition phonème par phonème, phrase par phrase. Lors de la reconnaissance, la recherche de sous-empreintes similaires est effectuée pour chaque trame du signal de parole de test. Un accès rapide à la base de référence est fourni en utilisant une méthode similaire au *Perceptual Audio Hashing* issu de la méthode de Philips [Haitsma et al., 2001]. Pour une trame donnée issue du signal de parole de test, une empreinte retenue pour l'identification est une empreinte de la base de référence qui minimise la distance avec le segment correspondant dans le signal de parole de test. A cet effet, le segment correspondant est obtenu par un alignement de l'empreinte sur les sous-empreintes du signal de parole de test. Cet alignement est assuré par un recalage de la position de la trame courante sur la position de la sous-empreinte correspondante dans l'empreinte issue de l'apprentissage.

La mesure de distance entre empreintes est définie selon la méthode de calcul de sous-empreinte choisie. Dans la méthode APSM, cette distance est donnée par le BER, en extension de la distance de Hamming locale sur les sous-empreintes. Un seuil sur ce BER est ajouté afin de valider ou non l'identification par l'empreinte comparée suivant la distance retournée. Dans les méthodes QV, QU, QN et 2C, cette mesure de distance est obtenue par le ratio du nombre de sous-empreintes différentes sur le nombre total de sous-empreintes contenus dans l'empreinte. Dans ces dernières méthodes, aucun seuil n'est défini et l'empreinte minimisant la distance est toujours retenue pour l'identification. L'algorithme de relâchement de la contrainte d'identification n'est utilisé que dans la méthode APSM. Dans toutes les autres méthodes, la contrainte d'identification est la présence de la sous-empreinte issue de la trame courante du signal de parole de test au sein des empreintes candidates à l'identification dans la base de référence.

Pour toutes les méthodes de calcul de sous-empreinte, cette procédure d'identification est appliquée trame après trame sur le signal de parole de test. Ainsi différentes empreintes de la base de référence peuvent être retenues pour l'identification d'un même ensemble de trames. Un algorithme de programmation dynamique est alors utilisé pour choisir la meilleure séquence d'empreintes et ainsi retourner les phonèmes correspondants (Annexe E, page 141). Nous développons au sein de ce système de DAP certaines idées originales adaptées de travaux de recherche en synthèse de la parole par concaténation d'unités phonétiques [Sigasaka, 1988]. Il s'agit de rechercher les plus longues séquences phonétiques afin de minimiser le nombre de points de concaténation d'éléments audio discontinus. Cette

recherche de plus longues séquences permet ainsi de réduire les effets de perturbation liés à la coarticulation. Parmi toutes les empreintes retenues durant l'identification, la recherche de la meilleure séquence phonétique représentative est alors effectuée grâce à un algorithme de programmation dynamique de type level-building Dynamic Time Warping [Myers et al., 1981] (Annexe E, page 141). Les contraintes sur cet algorithme sont la minimisation de la distance de similarité et la minimisation du nombre de changements de séquences d'empreintes. Ainsi, cet algorithme effectue une mise en correspondance de séquences d'empreintes issues de la base de référence sur un segment du signal de parole de test (Figure 3.6).

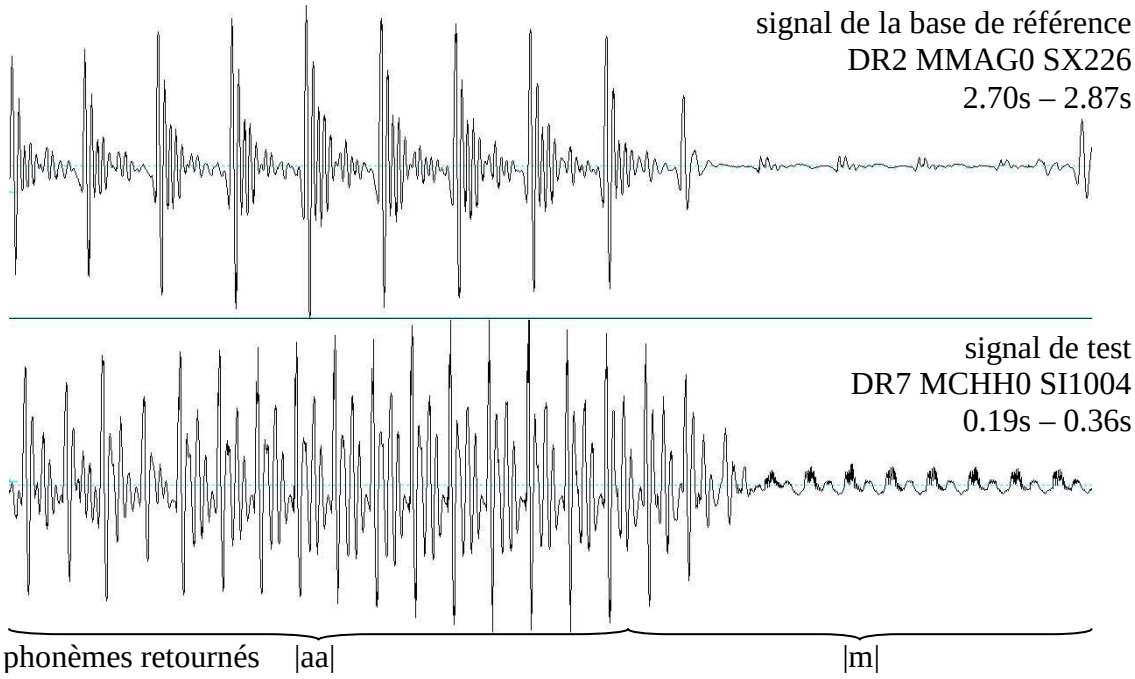



Figure 3.6 : Exemple de décodage d'une séquence phonétique (méthode QN, base TIMIT)

## 3.5 Expériences

Le système développé est évaluée pour sa performance dans la tâche de DAP. Les bases de données de signal de parole utilisées pour l'apprentissage et le test sont définies à partir de bases de données reconnues. Les paramètres variables définis par le système sont tout d'abord adaptés pour maximiser les performances sur un sous-ensemble de développement. Puis une expérience finale permet de valider les résultats obtenus en fonction des méthodes de calcul de sous-empreinte choisies.

### 3.5.1 Bases de données utilisées et type de sous-empreinte

Dans les expériences suivantes, les bases de données utilisées sont TIMIT [Fischer et al., 1986], NTIMIT [Jankowski et al., 1990] et BREF80 [Lamel et al., 1991] (Annexe B, page 135 et Annexe C, page 137). Les bases de données TIMIT et NTIMIT sont anglophones tandis que la base de données BREF80 est francophone.

L'ensemble phonétique choisi pour les deux bases de données TIMIT et NTIMIT est composé de 39 phonèmes, suivant la définition du CMU/MIT [Lee et al., 1989b]. Les bases TIMIT et NTIMIT sont segmentées en trois ensembles disjoints pour i) l'apprentissage (4 heures, 462 locuteurs), ii) le développement (30 minutes, 56 locuteurs) et iii) le test (1 heure, 112 locuteurs). Dans le cas de BREF80, l'ensemble phonétique est composé de 35 phonèmes. La base de données BREF80 est segmentée en trois ensembles disjoints pour i) l'apprentissage (9 heures, 72 locuteurs), ii) le développement (20 minutes, 3 locuteurs), iii) le test (40 minutes, 5 locuteurs).

Une taille commune de sous-empreinte de 16 bits est choisie afin de prévenir tout sur-apprentissage des données issues de l'ensemble d'apprentissage. L'ensemble des valeurs possibles pour la création d'une sous-empreinte en fonction de la méthode de création choisie et du nombre de phonèmes nécessaires à la description de la base de données sélectionnée est présenté dans un tableau récapitulatif (Tableau 3.1). Dans la méthode QU, la quantification vectorielle appliquée à chacun des phonèmes retourne 1680 centroïdes pour les sous-ensembles d'apprentissage de TIMIT et NTIMIT ainsi que 1872 centroïdes pour le sous-ensemble d'apprentissage de BREF80. Pour les méthodes QN et 2C, le nombre et la taille des clusters dépendent des centroïdes restants suite à la densification de l'espace de recherche.

| base de données et nombre d'éléments / type de sous-empreinte | | TIMIT | NTIMIT | BREF80 |
|---|---|---|---|---|
| APSM | binaires | 65 536 | 65 536 | 65 536 |
| QV | centroïdes | 65 536 | 65 536 | 65 536 |
| QU | centroïdes | 65 520 | 65 520 | 65 520 |
| QN | clusters<br>centroïdes | 2 224<br>33 730 | 3 053<br>24 765 | 1 113<br>44 933 |
| 2C | clusters<br>centroïdes | 67<br>28 750 | 140<br>18 522 | 35<br>42 032 |

Tableau 3.1 : Nombre d'éléments générés selon la méthode de calcul de sous-empreinte et la base choisies

Dans les méthodes QN et 2C, le nombre de clusters et de centroïdes restants après l'étape de densification de l'espace de recherche dépend fortement de la base de données choisie. En effet, dans le cas de TIMIT, il reste un peu plus de 50 % des centroïdes considérés pertinents pour la méthode QN. Ce taux signifie que pour près d'un centroïde sur deux obtenus dans la méthode QU, leur plus proche centroïde est de phonème différent. Ces centroïdes se retrouvent alors isolés lors de la densification de l'espace de recherche dans la méthode QN. Dans cette méthode QN, ce taux de centroïdes pertinents descend même jusqu'à 38 % dans le cas de NTIMIT alors qu'il est de 69 % dans le cas de BREF80. La considération de pertinence des centroïdes dans cette méthode est donc fortement liée aux conditions d'enregistrements de la base de données et à la qualité du signal de parole qui en résulte. En effet, la principale différence entre TIMIT et NTIMIT est la présence de bruits additifs et convolutifs additionnels [Kamper et al., 2009]. De surcroît, BREF80 est considérée comme une base de données de signal de parole propre grâce à l'enregistrement du signal audio en environnement contrôlé [Lamel et al., 1991; Besacier et al., 2001].

Ce contraste dans la réduction du nombre de centroïdes en fonction de la base de données augmente dans la méthode 2C. Ainsi, la méthode 2C permet de considérablement réduire le nombre de clusters, permettant une indexation de ces clusters sur 8 bits ou moins. Dans ce cas, après application de la méthode de condensation, seuls 44 % des centroïdes sont conservés pour TIMIT. Ce taux de conservation est de 28 % pour NTIMIT et de 64 % pour BREF80. Ce faible taux dans NTIMIT montre un éparpillement de la répartition de clusters phonétiques de petite taille dans l'espace de recherche. Ces clusters et leurs centroïdes

associés considérés non pertinents sont donc supprimés lors des itérations successives de l'algorithme de condensation. Dans le cas de BREF80, cette réduction très importante du nombre de clusters retourne la présence d'un unique cluster par phonème.

### 3.5.2 Evaluation du système développé

Les expériences consistent en l'évaluation du système dans une tâche de DAP, selon le type de sous-empreinte choisie. Les résultats obtenus sont comparés à ceux d'un système classique à base de chaînes de Markov cachées (*Hidden Markov Model*, HMM), sans modèle de langage [Young et al., 2006] (Annexe F, page 145). A partir des mêmes vecteurs MFCCs utilisés dans la méthode QV, chaque HMM définissant un modèle phonétique est construit sur 3 états contenant chacun un modèle de mélange de 512 gaussiennes. Le nombre total de gaussiennes ainsi définies est de 59904 gaussiennes pour chaque sous-ensemble d'apprentissage de TIMIT et NTIMIT ainsi que de 53760 gaussiennes pour le sous-ensemble d'apprentissage de BREF80.

L'évaluation du système est définie en deux temps :

- la phase de développement permet d'adapter les paramètres variables du système et de choisir les types de sous-empreinte retournant les meilleures performances,
- la phase de test final permet de valider les résultats selon les types de sous-empreinte.

Les critères utilisés pour la mesure de performance durant l'évaluation sont :

- le nombre de phonèmes dans la transcription évaluée ($N$),
- le nombre d'erreurs de suppression (*Deletion errors*, $D$),
- le nombre d'erreurs de substitution (*Substitution errors*, $S$),
- le nombre d'erreurs d'insertion (*Insertion errors*, $I$),
- la performance globale (*Accuracy*, $A$).

Les résultats de ces évaluations sont exprimés en taux moyens de ces critères rapportés sur le nombre total de phonèmes dans les transcriptions évaluées. Dans ce cas, la performance globale $A$ est obtenue par :

$$A = 1 - \frac{D + S + I}{N} \tag{3.5}$$

Dans notre système d'identification audio par empreinte dédiée au DAP, les empreintes d'une durée inférieure ou égale à 30 ms ne sont pas chargées dans la base de référence. Cette restriction permet de limiter un grand nombre d'erreurs d'insertion dues aux phonèmes très courts. Dans la procédure de programmation dynamique, le coût d'un changement de séquence d'empreintes d'une valeur de 1 est ajouté au coût issu de la mesure de distance entre empreintes de valeur réelle entre 0 et 1. Dans la méthode APSM, le choix d'un petit seuil de BER contraint le critère de similarité. Dans ce cas, peu d'empreintes sont sélectionnées pour l'identification et un grand nombre d'erreurs de suppression apparaît. Lorsque cette contrainte sur le seuil de BER est relâchée, beaucoup d'erreurs d'insertion sont admises. Le meilleur ajustement sur ce seuil de BER est donné pour une valeur de BER à 0,1. Les résultats d'expérience sur le sous-ensemble de développement de chacune des bases de données sont obtenus pour ces paramètres de programmation dynamique et de seuil sur le BER (Tableau 3.2).

| base % / type | TIMIT | | | | NTIMIT | | | | BREF80 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | *D* | *S* | *I* | *A* | *D* | *S* | *I* | *A* | *D* | *S* | *I* | *A* |
| **HMM** | 4,9 | 21,7 | 12,2 | **61,2** | 6,1 | 32,9 | 17,3 | **43,7** | 3,6 | 15,8 | 6,2 | **74,4** |
| APSM | 38,1 | 37,2 | 7,1 | 17,6 | 41,1 | 38,3 | 7,6 | 13,0 | 46,4 | 29,6 | 5,0 | 19,0 |
| QV | 22,0 | 39,9 | 11,9 | 26,2 | 16,4 | 48,2 | 21,8 | 13,6 | 14,4 | 33,5 | 11,0 | 41,1 |
| QU | 19,6 | 31,8 | 8,6 | 40,0 | 16,9 | 44,0 | 18,2 | 20,9 | 14,5 | 27,2 | 9,4 | 48,9 |
| **QN** | 18,0 | 27,3 | 8,5 | **46,2** | 19,2 | 40,0 | 12,6 | **28,2** | 16,6 | 21,6 | 6,5 | **55,3** |
| **2C** | 17,1 | 27,2 | 8,3 | **47,4** | 19,0 | 39,9 | 12,0 | **29,1** | 10,4 | 24,9 | 10,3 | **54,4** |

Tableau 3.2 : Résultats du DAP selon les différents types de représentation définis (développement)

Les sous-empreintes à base de paramètres MFCCs quantifiés sont de bien meilleurs vecteurs de représentation du signal de parole pour la tâche de DAP que celles issues de la méthode APSM. Dans la méthode QU, le nombre d'erreurs de substitution est réduit par rapport à la méthode QV grâce à une meilleure représentation de l'espace de recherche. Les sous-empreintes issues des méthodes QN et 2C retournent les meilleurs résultats. Dans la méthode QN, le nombre d'erreurs de substitution diminue, en particulier grâce à une augmentation de la taille des séquences d'empreintes retournées par l'algorithme de programmation dynamique. D'un autre coté, les erreurs de substitution sont parfois dues à un

mauvais choix de séquences d'empreintes contenant des phonèmes mal identifiés. Cependant, l'usage de ces techniques de DAP retourne des performances contrastées par des résultats bien en retrait par rapport à un système classique de DAP à base de modèles HMMs multigaussiens.

Une expérience finale est effectuée sur chacun des sous-ensembles de test pour le système de référence à base de HMM ainsi que pour les méthodes QN et 2C (Tableau 3.3).

| base / % / type | TIMIT | | | | NTIMIT | | | | BREF80 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | *D* | *S* | *I* | *A* | *D* | *S* | *I* | *A* | *D* | *S* | *I* | *A* |
| **HMM** | 4,6 | 21,7 | 11,9 | **61,7** | 6,1 | 32,9 | 17,3 | **43,7** | 3,7 | 15,7 | 5,8 | **74,8** |
| **QN** | 18,0 | 27,8 | 8,7 | **45,5** | 19,2 | 40,1 | 12,9 | **27,8** | 11,0 | 24,0 | 12,0 | **53,0** |
| **2C** | 17,2 | 27,7 | 8,3 | **46,8** | 18,7 | 40,3 | 12,5 | **28,5** | 11,0 | 23,9 | 10,9 | **54,2** |

Tableau 3.3 : Résultats du DAP selon les différents types de représentation choisis (test)

Les résultats sur le test final confirment les performances obtenues durant le développement. Le choix d'exclure les empreintes d'une durée de 30 ms et moins diminue certes le nombre d'insertions mais au détriment d'une augmentation du nombre de suppressions. Le compromis obtenu par ce choix n'est pas optimal.

## 3.6 Discussions

Un premier travail proposé par Vasiloglou présentait une adaptation d'un système d'identification audio par empreinte pour une tâche de RAP en mots isolés [Vasiloglou et al., 2004] (Chapitre 2, page 27). Dans ce présent chapitre, nous avons poursuivi ces travaux d'adaptation. Nous avons alors développé une nouvelle approche de DAP utilisant les propriétés et techniques de l'identification audio par empreinte [Haitsma et al., 2001].

Le principe général d'un système classique d'identification audio par empreinte est conservé (Figure 3.1, page 45). Le signal de parole et les unités phonétiques sont alors les données utilisées pour la formation des empreintes. Dans notre système, la base de référence conserve linéairement les sous-empreintes au-fur-et-à-mesure de leur calcul sur le signal de

parole analysé. La transcription phonétique du signal de parole ainsi indexé vient compléter les données d'apprentissage. Dans notre système, une empreinte est donc formée par la séquence de sous-empreintes correspondant au segment temporel d'un phonème donné. Par ailleurs, la méthode d'accès rapide par tableau d'indexation et fonction de hachage est conservée en rapport à la méthode d'identification audio par empreinte présentée par Philips (section 1.3, page 17). Lors de l'étape d'identification, un alignement est effectué pour la comparaison entre les segments de sous-empreintes en se basant sur la position du vecteur utilisé pour l'accès rapide et les informations issues de la transcription phonétique. La contrainte d'identification a également été assouplie en permettant une augmentation de l'espace de recherche des empreintes issues de l'apprentissage à comparer pour l'identification. Ainsi la sous-empreinte initiant la comparaison doit désormais avoir une distance comprise dans le seuil de BER et ne plus être strictement identique. Une méthode de programmation dynamique permet en outre d'assembler les empreintes retenues pour l'identification phonétique afin de favoriser le choix d'empreintes contigües dans la base de référence. Cette étape supplémentaire de programmation dynamique permet d'assembler des séquences de phonèmes afin de limiter les effets de perturbation liés à la coarticulation.

Par ailleurs, nous avons adapté le principe de calcul de sous-empreinte pour une meilleure représentation du signal de parole. Des sous-empreintes compactes sont alors définies à partir de méthodes de quantification de paramètres MFCCs et de densification de l'espace de recherche. Ainsi, pour une taille de sous-empreinte définie sur 16 bits, l'espace nécessaire au stockage de la base de référence est inférieur à un méga-octet par heure de signal de parole. Ces nouvelles sous-empreintes retournent de bien meilleurs résultats que celles issus de méthodes classiques d'identification audio par empreinte. Ces sous-empreintes issues de valeurs d'indexation perdent toutefois certaines caractéristiques liées à l'indépendance des bits qui les composent. De surcroît, les performances de notre système basé sur les principes de l'identification audio par empreinte restent en retrait par rapport à un système de référence construit par des HMMs à base de mélanges de gaussiennes.

Dans la suite de cette étude, il s'agit donc de caractériser certaines propriétés de représentation acoustique du signal de parole. Cette caractérisation permettra alors d'évaluer dans quelles mesures les sous-empreintes proposées par certaines méthodes d'identification audio par empreinte peuvent être adaptées pour leur utilisation au sein d'un système de RAP.

# Conclusion de la première partie

Les méthodes d'identification audio par empreinte sont reconnues pour leurs performances dans la tâche de détection d'extraits de musique [Haitsma et al., 2001; Wang, 2003; Pinquier et al., 2004]. Bien que le choix de paramètres acoustiques représentatifs du signal audio varie d'une méthode à l'autre (section 1.1.3, page 13), les techniques utilisées pour la conservation des données, leur accès et leur comparaison avec un signal audio de test sont communes (Figure 1.1, page 12). Ainsi, au sein d'un système d'identification audio par empreinte, trois modules sont présents pour :

- le calcul des sous-empreintes issues de l'analyse acoustique du signal,
- la conservation des empreintes et métadonnées associées dans une base de référence,
- la comparaison des empreintes avec un flux audio de test.

Cependant, les techniques développées par ces méthodes ne sont pas adaptées à la détection d'autres types d'évènements sonores comme pour la reconnaissance de la parole [Ogle et al., 2007]. Une première adaptation de telles techniques pour la RAP a été proposée par Vasiloglou [Vasiloglou et al., 2004]. Il s'agit dans cette étude de se baser sur les travaux d'identification audio par empreinte présentés par Philips [Haitsma et al., 2002] pour concevoir un système d'identification de mots isolés dans un environnement mono-locuteur. L'évaluation d'un tel système est effectuée sur l'identification de versions bruitées d'empreintes de signal de parole conservées dans une base de référence. Deux expériences complémentaires ont été effectuées pour évaluer ce système dans les tâches de reconnaissance de mots isolés en environnement mono-locuteur (section 2.3.2, page 35) et de reconnaissance de phonèmes isolés en environnement multi-locuteur (section 2.3.3, page 36).

Nous avons décidé de poursuivre cette voie de recherche en construisant un système d'identification audio par empreinte adapté à la tâche de DAP en parole continue dans un environnement multi-locuteur. Durant ces travaux, nous avons développé une technique de relâchement de la contrainte d'identification (section 3.2, page 47). Grâce à cette technique,

l'espace de recherche des empreintes issues de l'apprentissage à comparer pour l'identification est augmenté afin de diminuer le risque de faux rejet. De surcroît, une étape de programmation dynamique est ajoutée afin de sélectionner les empreintes utilisées pour retourner leur phonème associé durant l'identification. Cette étape est adaptée de travaux de recherche en synthèse de la parole par concaténation d'unités phonétiques [Sigasaka, 1988]. Elle permet d'assembler des séquences de phonèmes afin de réduire les effets de perturbation liés à la coarticulation. De surcroît, de nouvelles méthodes de calcul de sous-empreinte sont définies. Ces sous-empreintes sont basées sur des paramètres acoustiques de type MFCC [Bogert et al., 1963]. Différentes méthodes de quantification et de densification de l'espace de recherche sont proposées pour le calcul de ces sous-empreintes (section 3.3, page 49). Durant l'évaluation de notre système à la tâche de DAP, l'usage de ces nouvelles sous-empreintes permet de meilleurs résultats qu'avec des sous-empreintes issues de l'identification audio par empreinte classique (section 3.5, page 57). Cependant, ces nouveaux vecteurs perdent certaines caractéristiques liées à l'indépendance des bits qui les composent. De surcroît, quelque soit la méthode de calcul de sous-empreinte utilisée durant l'évaluation, notre système reste toujours moins performant qu'un système de référence construit autour de modèles HMMs à base de mélanges de gaussiennes (Annexe F, page 145).

Au cours du développement de notre système de DAP, nous avons soulevé certains points de divergence dans le traitement des données entre les tâches de l'identification audio d'une part et de la RAP d'autre part. En effet, on peut considérer que l'identification audio dans le cadre de la musique et de la parole traitent de problématiques différentes. Pour l'identification audio, il s'agit d'un processus d'appariement entre un signal d'origine et une version dégradée de ce signal. L'identification audio par empreintes est une technique robuste à des dégradations du signal audio par des éléments extérieurs liés à l'acquisition, la conservation et la restitution de ce signal. En RAP, il s'agit de regrouper sous une même catégorie des séquences de signal de parole uniques et variables qui sont perçues similaires. Les paramètres acoustiques mis en œuvre dans un système de RAP, comme le DAP par exemple, doivent alors être adaptés à cette singularité afin de développer un système de reconnaissance robuste aux différentes variabilités du signal de parole. Il est donc nécessaire de définir et de décrire cette variabilité ainsi que les principes mis en œuvre dans la représentation acoustique du signal de parole pour le développement de systèmes de RAP. Nous désirons ensuite poursuivre cette démarche d'adaptation de l'identification audio à la reconnaissance de la parole à travers le développement d'une application d'identification acoustico-phonétique.

# Seconde partie

Dans cette seconde partie, nous introduisons le cadre général de la problématique de l'identification dans la reconnaissance automatique de la parole (RAP). Il s'agit en l'occurrence d'étudier dans quelle mesure les paramètres de représentation du signal acoustique issus de l'identification audio par empreinte peuvent être adaptés à la reconnaissance de la parole. Dans un premier temps, les principes généraux de la RAP sont rappelés. Une réflexion autour de l'étude de la variabilité est abordée et les principaux paramètres acoustiques usuels en RAP sont décrits. Dans un second temps, la robustesse de paramètres acoustiques choisis est évaluée sur deux types de représentation usuelle :

- des paramètres classiquement utilisés en RAP,
- des paramètres issus de l'identification audio par empreinte.

Les systèmes de RAP voient leur performance diminuer de manière significative lorsque leurs conditions d'entraînement et d'apprentissage sont différentes de celles de leur utilisation [Lippmann, 1997]. L'analyse des performances de ces systèmes dans de telles conditions permet d'évaluer la robustesse de ces systèmes de RAP [Moreno et al., 1994; Sirigos et al., 1997]. Les causes de variabilité entre ces conditions sont diverses [Thosar et al., 1976; Russell et al., 1983; Ahmed et al., 1985]. Elles peuvent être liées aux principes de production du signal de parole ; c'est la variabilité dite intrinsèque [Benzeghiba et al., 2007]. Dans ce cas, la variabilité intrinsèque est celle présente lors de l'émission de deux signaux de parole distincts, qu'ils contiennent un même message linguistique ou deux messages différents. Cette variabilité est due tout autant aux différences de prononciation entre locuteurs (variabilité inter-locuteur) qu'aux variations subie entre deux prononciations d'un même message par un locuteur unique (variabilité intra-locuteur) [Benzeguiba et al., 2006b]. Dans un système de RAP, cette variabilité intrinsèque peut être perçue comme une distorsion de la production du signal de parole entre les phases d'apprentissage et de test.

D’autres causes de variabilité peuvent être également liées à l’environnement acoustique ainsi qu’à la chaîne d’acquisition et de transmission du signal sonore. Ces causes de variabilité sont liées aux principes de diffusion et de réception du signal de parole ; il s’agit alors de la variabilité dite extrinsèque [Benzeghiba et al., 2006]. En effet, l’environnement acoustique joue le rôle de composante perturbatrice sur le signal de parole d’origine en sortie du conduit vocal. Cette composante est le plus souvent indépendante du signal de parole émis. Le bruit généré par l’environnement acoustique peut être considéré stationnaire ou au contraire très variable. La variabilité introduite par la nature de ce bruit est donc complétée par une variabilité provoquée par l’évolution de ce bruit dans le temps [Jie et al., 2009]. De surcroît, le circuit d’enregistrement pour l’acquisition du signal audio introduit également une distorsion du signal de parole [Hansen et al., 2001]. En plus des bruits électriques variables, il peut également s’agir de variations dans la chaîne de production utilisée pour l’enregistrement et la rediffusion. Ainsi, un changement de microphone entre l’enregistrement des phases d’apprentissage et de test peut modifier la forme générale du spectre du signal de parole [Menéndez-Pidal et al., 2001]. Toutes ces variations provoquent alors une diminution des performances d’un système de RAP [Lippmann, 1997].

Dans la conception d’un système complet de RAP, un module de description du signal de parole permet d’extraire les informations caractéristiques et pertinentes du signal acoustique, avec la possibilité d’y associer une information extérieure [Young et al., 2006]. Ces informations caractéristiques peuvent être définies selon trois facteurs d’échelle [Boite et al., 2000] :

- au niveau global. Le signal de parole analysé est décrit sur une échelle de l’ordre de plusieurs secondes. Les propriétés définies ainsi ne sont obtenues qu’à partir de l’étude d’une longue durée du signal de parole. Le genre du locuteur, le rythme d’élocution ou l’émotion portée dans la voix sont des exemples de tels descripteurs globaux [Ververidis et al., 2006]. Lors de la reconnaissance d’un message linguistique, la portée sémantique, c'est-à-dire le sens du message transmis, est une forme de descripteur global [Haton et al., 1976]. C’est également à ce niveau que s’inscrivent les métadonnées.

- au niveau intermédiaire, dit niveau suprasegmental. Le signal de parole est décrit sur une échelle de l’ordre de la seconde. Ce type de descripteur permet de mettre en évidence certaines caractéristiques acoustiques du signal de parole. Le signal peut alors être segmenté en parties distinctes par exemple par détection de texture en analysant l’enveloppe spectrale ou de rupture suite à une étude de l’évolution de l’énergie [Collet et al., 2005]. Chacune de ces parties est classée dans une catégorie, définie en fonction

de l'application souhaitée. Une segmentation en groupes de mots pouvant reconstituer le message prononcé est considérée comme une forme de descripteur intermédiaire [Mercier et al., 1982]. L'étude de la prosodie du signal de parole s'effectue également à ce niveau suprasegmental [Lindblom et al., 1973; Suaudeau et al., 1994].

- au niveau local, dit niveau segmental. Le signal de parole est décrit sur une échelle de l'ordre du centième de seconde. Ce type de descripteur permet une représentation acoustique fine du signal de parole analysé. De nombreux types de paramètres acoustiques sont définis en fonction de la sélection des caractéristiques du signal de parole choisies pour discriminer ce signal [Davis et al., 1980; Mokbel, 1992; Eyben et al., 2010]. La recherche des rapports d'énergie dans le spectre du signal analysé à l'échelle d'une trame de quelques dizaines de millisecondes permet de définir un descripteur local [Eyben et al., 2010]. Par ailleurs, une segmentation phonétique permettant de caractériser le signal de parole sous la forme d'un nombre limité d'unités linguistiques s'inscrit dans ce niveau segmental [Lindblom et al., 1973].

Au sein d'un système de RAP, les principales techniques de reconnaissance de la parole sont développées pour retrouver l'information au niveau global. Ces techniques sont le plus souvent basées sur deux modules bien distincts [Young et al., 2006] :

- un module de calcul des paramètres acoustiques au niveau local,
- un module de reconnaissance de forme nécessitant souvent un passage par le niveau intermédiaire.

Ces modules servent à traiter l'information acquise lors de l'enregistrement du signal de parole. Considérant le locuteur humain comme producteur du signal de parole, le principe fonctionnel général de la RAP peut être résumé en quatre parties distinctes (Figure 3.7) [Boite et al., 2000] :

- acquisition du signal. Le message émis sous forme d'un signal acoustique est enregistré suivant une chaîne d'acquisition. Le signal de parole est numérisé sous la forme d'échantillons sonores.

- calcul des paramètres acoustiques. Le signal enregistré est transformé en une séquence de vecteurs de paramètres d'observations acoustiques par un module de calcul des paramètres acoustiques.

- apprentissage (facultatif). Un apprentissage, également appelé phase d'entraînement, exploite les données issues d'un signal de parole de référence. Cet apprentissage peut permettre une adaptation du module de reconnaissance des formes afin d'améliorer les performances du système développé lors du décodage.

- décodage. Cette partie se matérialise sous la forme d'un test correspondant à la tâche de reconnaissance ou d'identification selon le type d'application visée. Un module de reconnaissance des formes s'applique à interpréter des segments de paramètres acoustiques en une séquence de termes linguistiques afin de déterminer le message prononcé.

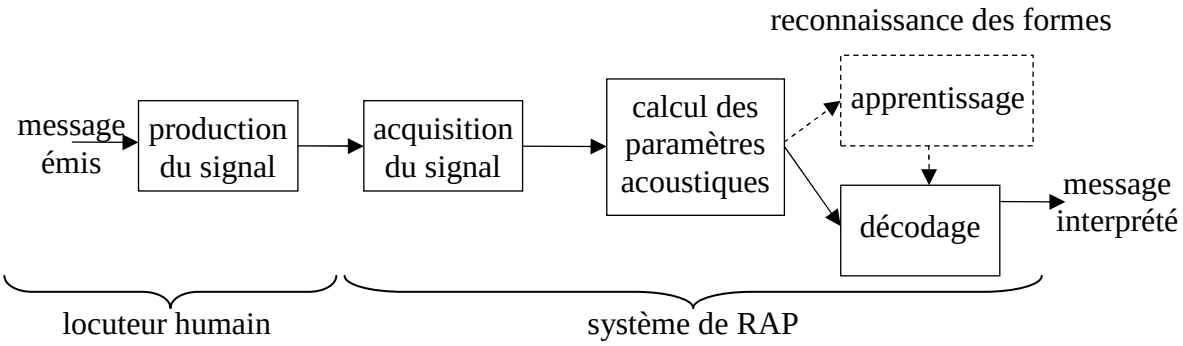



Figure 3.7 : Principe fonctionnel général de la RAP

En tenant compte de ces deux modules de paramètres acoustiques et de reconnaissance des formes, nous présentons d'abord le principe général d'un système de RAP. Les variabilités acoustiques du signal de parole et leur traitement par un système de RAP sont étudiés. Nous abordons dans ce cadre les principes de modélisation acoustique du signal de parole et la représentation des éléments linguistiques qui sont exprimés à travers ce signal. Compte-tenu de la variabilité du signal acoustique de parole, de nombreux efforts ont été effectués afin d'améliorer les performances des systèmes de RAP [Benzeghiba et al., 2007]. L'extraction des caractéristiques du signal de parole lors de l'acquisition du signal et du calcul des paramètres acoustiques est une étape décisive dans la chaîne du traitement des données audio d'un système de RAP. Cette seconde partie s'attache à en exprimer les principaux enjeux.

Dans un premier temps, les différents types de variabilité du signal de parole sont décrits. Les principales possibilités de représentation acoustique du signal de parole sous forme de paramètres acoustiques sont alors énoncées. Dans un second temps, la robustesse aux différents types de variabilité est étudiée sur divers paramètres acoustiques de représentation du signal de parole. Ces paramètres peuvent autant être issus de représentations acoustiques usuelles en RAP comme se baser sur des méthodes de calcul de paramètres acoustiques adaptées de l'identification audio par empreinte. La robustesse de tels paramètres aux différents types de variabilités acoustiques est enfin évaluée au sein d'une tâche d'identification acoustico-phonétique.

# Chapitre 4. Variabilité du signal de parole

Le signal de parole émis par un locuteur ne contient pas uniquement le message linguistique prononcé. En effet, ce signal acoustique varie fortement en fonction par exemple des propriétés physiologiques du locuteur ou de considérations géographiques qui influent sa prononciation [Meyer et al., 2010]. Toutes ces informations ne sont pas utiles à la reconnaissance automatique de la parole (RAP). Elles viennent même parfois perturber la bonne compréhension du message prononcé. Ces informations perturbatrices, propres au signal de parole, sont regroupées sous l'expression de « variabilité intrinsèque » [Benzeghiba et al., 2007].

Par ailleurs, un évènement sonore peut être interprété comme une oscillation des molécules d'air de la sortie du système de diffusion jusqu'à l'entrée de l'oreille interne [Boite et al., 2000]. Dans ce cas, le signal de parole peut donc être appréhendé comme un signal acoustique variable émis par un locuteur et perçu par un auditeur. Ce signal est transmis par une fluctuation locale de la pression acoustique dans l'air. Les perturbations dues à l'environnement comme le bruit extérieur et les distorsions dues au canal de transmission sont communes à toute analyse de signal acoustique. Ces perturbations extérieures à la production de la parole sont regroupées sous l'expression de « variabilité extrinsèque » [Benzeghiba et al., 2006].

Considérant ces différents types de variabilité, un système de RAP est robuste à un ensemble restreint de conditions d'utilisation [Boite et al., 2000]. Ces conditions limitent le cadre applicatif dans lequel le système mis en place peut être efficace. Pourtant, un système de RAP peut être amené à fonctionner dans des conditions différentes de celles pour lesquelles il a été entraîné avec des données d'apprentissage [Lippmann, 1997]. Dans de telles conditions, les performances d'un système de RAP peuvent diminuer [Moreno et al., 1994]. La recherche de l'amélioration de la robustesse des systèmes de RAP par rapport à ces types de variabilité fait l'objet d'un grand nombre de travaux. Certaines approches s'emploient à

obtenir une paramétrisation du signal de parole qui réduit l'influence du canal de transmission [Lee, 1997; Menéndez-Pidal et al., 2001; Morales et al., 2009]. D'autres approches développent des mécanismes d'adaptation aux conditions d'évaluation et de suppression des effets de l'environnement acoustique [Baker et al., 1986; Neumeyer et al., 1995; Kristjansson et al., 2001; Jiucang et al., 2009; Raut et al., 2009; Rennie et al., 2011]. Le but d'un système de RAP est alors d'extraire l'information linguistique du signal de parole, représenté le plus souvent sous forme d'unités linguistiques, en dépit des variabilités extrinsèque et intrinsèque de ce signal acoustique [Benzeghiba et al., 2007].

Dans un premier temps, les différents types de variabilités extrinsèque et intrinsèque sont décrits en fonction de leurs propriétés de perturbation au sein du signal de parole. Puis dans un second temps, le choix de paramètres acoustiques et leurs différentes propriétés sont présentés. La capacité de tels paramètres à représenter les particularités du signal de parole est discutée en fonction de la manière dont ces paramètres sont utilisés au sein d'un système de RAP. Les paramètres acoustiques choisis pour décrire le signal de parole dans les travaux de ce présent document sont alors détaillés.

## 4.1 Variabilité intrinsèque

La variabilité intrinsèque de la parole est liée à l'origine biologique de sa production [Boite et al., 2000]. Le signal de parole ne transmet pas uniquement le message linguistique mais également un grand nombre d'informations sur le locuteur lui-même : genre, âge, origines régionales et sociales, état de santé et état émotionnel [Huang et al., 2001]. Toutes ces informations dépendent du locuteur et des conditions de prononciation du message. Ces informations conditionnent les facteurs de variabilité intrinsèque.

### 4.1.1 Production et perception du signal de parole

La production de la parole est formée par un flux d'air traversant l'appareil phonatoire composé de surfaces vibrantes et de cavités résonnantes [Fant, 1960]. Dans ce cas, le signal de parole peut être modélisé sous la forme d'un modèle dit source-filtre. La source se réfère alors au flux d'air généré par les poumons et passant par le larynx. Le filtre se réfère quant à lui au conduit vocal composé des différentes cavités situées entre la glotte et les lèvres. L'onde acoustique qui porte le signal de parole est le résultat de l'excitation des cavités nasales et/ou orales par une ou deux sources acoustiques [Calliope, 1989]. Ces sources peuvent être un flux laryngé ou des bruits d'explosion et de friction produits dans la cavité orale [Landercy et al., 1982]. Dans ce cas, les mouvements et actions de l'appareil phonatoire (glotte, trachée, langue, joues, lèvres) composent les réalisations acoustiques [Fant, 1960]. Le signal de parole est donc formé par l'enchaînement de ces réalisations acoustiques. Le signal

de parole formé par ces mouvements continus n'est donc pas un signal audio considéré stationnaire [Levinson et al., 1983].

Les systèmes physiologiques de production d'un signal de parole par la voix et de sa perception par l'oreille humaine sont spécifiquement conçus en adéquation [Boite et al., 2000]. Ainsi, l'appareil auditif perçoit les sons compris dans une plage de fréquences allant de 20 Hz à 20 kHz environ [Vogel et al., 2007]. De plus, la perception de l'intensité sonore diffère suivant la fréquence du signal sonore. En effet, les niveaux de sensibilité (seuil d'audition minimal) et de douleur (seuil d'audition maximal) varient en fonction de la fréquence et sont les plus sensibles pour la plage de fréquences allant de 200 Hz à 5000 Hz environ [Fletcher et al., 1933]. Cette plage de fréquences correspond à celle dans laquelle le signal de parole transmet la majeure partie de ses informations acoustiques, soit une plage de fréquences allant de 170 Hz à 4000 Hz [Zwicker et al., 1981] (Figure 4.1).

Par ailleurs, la capacité de reconnaissance de la parole peut être perçue comme un phénomène de catégorisation de séquences du flux de parole [Liberman et al., 1957]. Cette catégorisation permet de retourner des mots ou unités linguistiques nécessaires à la compréhension du message linguistique transmis. Hors la réalité physique d'un signal acoustique continu ne correspond pas à ce schéma de catégorisation. Ainsi, cette aptitude à catégoriser correspond à la faculté d'adaptation de la représentation acoustique d'un signal de parole continu en regroupements d'objets considérés similaires sous forme de classes de données utiles [Reeves et al., 1998].

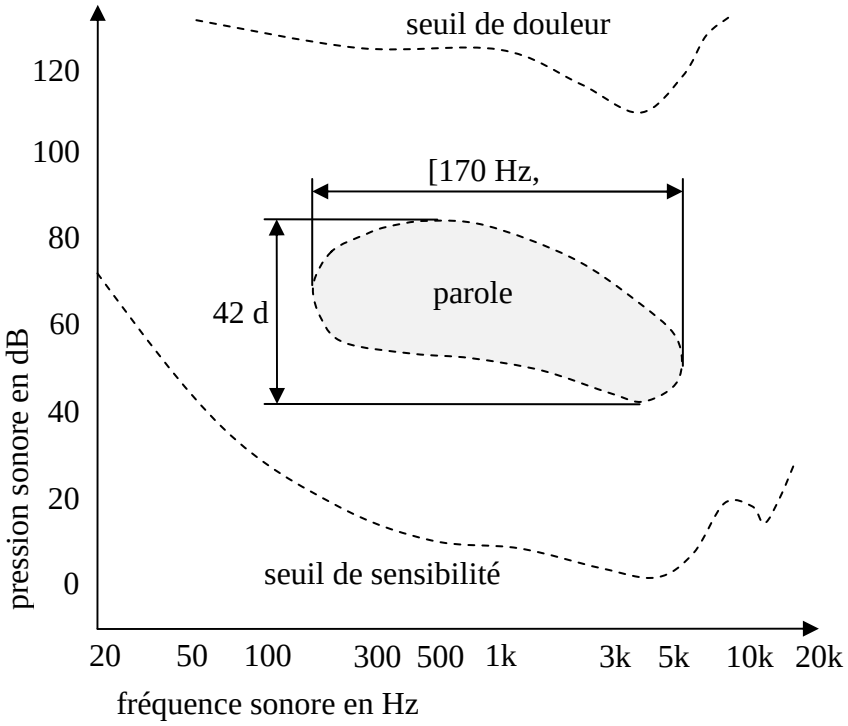



Figure 4.1 : Audiogramme de production de la parole, seuils auditifs [Fletcher et al., 1933; Zwicker et al., 1981]

### 4.1.2 Taxinomie des sons : phonologie

Il est possible de modéliser le signal acoustique de la parole par un nombre limité d'unités linguistiques [MacNeilage, 1973; Scully, 1987]. Ces unités sont alors considérées comme des réalisations élémentaires, le plus souvent représentatives de la langue de prononciation du message linguistique. Les plus petites unités acoustiques de ces réalisations élémentaires sont les phonèmes [Nolan, 1983]. Ces phonèmes permettent de distinguer la prononciation afin de déterminer le message linguistique émis par le signal de parole. Par exemple, en français, les sons /p/ et /t/ sont représentés par deux phonèmes distincts car leur discrimination permet de distinguer des mots différents (pas/tas, pot/tôt, etc.). Les sons produits par ces réalisations acoustiques élémentaires peuvent donc être rattachés à différentes classes phonétiques. Ces classes phonétiques regroupent alors les réalisations acoustiques possédant certaines caractéristiques communes facilement discriminables [Adamczewski et al., 1977] (Annexe G, page 149).

La notation phonétique permet d'associer à chaque mot une prononciation idéale sous la forme d'une séquence de phonèmes. L'Alphabet Phonétique International (API) est une classification phonétique exhaustive et très répandue [IPA, 1999]. Cet alphabet exploite des caractères particuliers afin de ne pas être confondu avec les caractères de langue écrite. D'autres alphabets phonétiques existent en fonction de la langue de destination [Fischer et al., 1986; Lamel et al., 1991]. Ces alphabets permettent alors un codage en ASCII afin de faciliter leur implémentation informatique [Lee et al., 1989b].

### 4.1.3 Coarticulation et variation phonétique

L'appareil phonatoire est soumis à des contraintes mécaniques [Mackenzie, 1997]. Ces contraintes ont pour effet de limiter les variations rapides des parties mobiles (forme des conduits, position de la langue et des lèvres). Les caractéristiques acoustiques des phonèmes prononcés varient suivant le mouvement du conduit vocal lors de l'articulation [Koreman et al., 1999]. En effet, en tenant compte de ces contraintes, la réalisation d'un phonème est tout autant influencée par celle des phonèmes précédents qu'elle influence celle des phonèmes suivants. Même dans une parole continue bien articulée, la production d'un phonème donné résulte d'un mouvement continu de l'appareil articulatoire. Ce mouvement continu assure la prononciation d'un phonème, suite à l'évolution de la configuration articulatoire des phonèmes le précédant et en fonction de la configuration articulatoire des phonèmes le suivant [Hardcastle et al., 2010]. Ce phénomène est connu sous le nom de coarticulation [O'Shaughnessy, 1974]. Donc la connaissance du contexte de l'élément phonétique est considérée comme un élément majeur pour discriminer la catégorisation phonétique. En effet, un phonème est classé dans une catégorie phonétique donnée non seulement par ses propriétés

propres mais également par la perception de ce phonème dans son contexte [Lakoff, 1987; Miller, 1994; Ladefoged et al., 1996]. Dans le domaine particulier de l'étude de la phonétique, ces effets de coarticulation sont considérés extrinsèques au phonème étudié car ils débordent du cadre de description de la variabilité à l'intérieur du phonème [Tohkura et al., 1992]. Dans le cadre de notre étude dans le domaine de la RAP, ces effets de coarticulation sont considérés intrinsèques au signal de parole car ils ne dépendent pas de facteurs extérieurs à sa production.

Pour un même locuteur, la réalisation acoustique d'un phonème donné peut varier en durée mais aussi suivant la forme du conduit vocal dépendant entre autres du contexte de l'élocution et des caractéristiques propres au locuteur [Kuwabara, 1997; Mokhtari, 1998]. Les amplitudes de variation phonétique varient selon que la parole soit lue, promptée, spontanée ou conversationnelle. Selon le style de parole employé, divers effets de prononciation peuvent apparaître, souvent liés à une réduction de l'articulation [Lindblom, 1990; Sotillo et al., 1998]. De surcroît, le débit de parole influe également sur la prononciation du message [Mirghafori et al., 1995; Siegler et al., 1995; Martinez et al., 1997]. La forme du conduit vocal est également propre à chaque individu [Haton et al., 1991]. Cette forme conditionne les propriétés de l'onde d'une réalisation acoustique [Janse, 2004]. Un même phonème peut donc être représenté par une grande variété de réalisations acoustiques.

Il apparaît également dans les bases de données de RAP en parole continue que les mots à forte probabilité d'occurrence dans la langue sont souvent peu articulés [Lemaire, 2007]. C'est le cas des mots d'usage grammatical, dits mots-outils, dont le rôle est réduit à une utilité syntaxique [Bahl et al., 1989]. De surcroît, en plus des effets de coarticulation, les réalisations acoustiques de la prononciation d'un message linguistique peuvent faire également l'objet de substitutions ou de suppressions phonétiques [Warren, 1970; Fosler-Lussier et al., 1999; Duez, 2003; Adda-Decker et al., 2005]. Dans ce cas, l'information manquante au niveau acoustique peut être alors retrouvée au niveau syntaxique ou au niveau sémantique [Meunier, 2005]. Par ailleurs, la modélisation par phonèmes en contexte permet de représenter diverses variations par la mise en place de références acoustiques précises représentant le signal de parole. Ces références tiennent également compte des contraintes de coarticulation dépendantes du contexte [Farnetani, 1997]. Pour cette raison, les transcriptions manuelles des bases de données de signal de parole phonétiquement annotées sont souvent accompagnées d'un minimum d'indications de prononciation [Blanche-Benveniste, 1999]. Ces contraintes sur les effets de coarticulation sont d'autant plus grandes au niveau de la frontière entre les mots où de nombreuses perturbations peuvent apparaître comme des liaisons, des pauses et respirations ou encore des hésitations [Hwang et al., 1989]. Ainsi, pour prendre en compte ces perturbations dans les données de signal de parole servant à l'apprentissage, un alignement

automatique de la transcription phonétique de la base de données peut sensiblement améliorer la qualité de la transcription [Boula de Mareuil et al., 2002].

### 4.1.4 Variabilité liée au locuteur

Les variations observées lors de différentes réalisations acoustiques d'un même message linguistique peuvent être issues de [Yang et al., 1996] :

- la variabilité intra-locuteur. Un phonème prononcé par un même locuteur n'aura jamais la même réalisation acoustique. Compte-tenu de la nature mécanique de production de la parole, le processus de réalisation des phonèmes n'est pas déterministe [Diller, 1979]. Cette réalisation dépend entre autres de sa position dans le contexte phonétique, de la vitesse d'élocution [Janse, 2004], de l'état émotionnel du locuteur [Caraty et al., 2010; Montacié et al., 2011].

- la variabilité inter-locuteur. La forme des ondes acoustiques émises ainsi que leur enchaînement dépendent entre autres des caractéristiques morphologiques du locuteur [Huang et al., 1991; Kubala et al., 1994]. Les effets de coarticulation, d'intonation et les modulations d'amplitude du signal de parole sont de surcroît fortement liés à l'origine géographique et sociale de chaque individu [Garvin et al., 1963; Haton et al., 1991; Lawson et al., 2003].

Dans l'ensemble de ces cas, le signal de parole correspondant à diverses prononciations de la même phrase sera différent. On peut ainsi regrouper les effets de variabilité intrinsèque en trois classes caractéristiques selon [Crouzet, 2000] :

- l'étude des propriétés physiques individuelles,
- l'étude de la modulation à long terme de la voix,
- l'étude de la prononciation du message linguistique.

Par ailleurs, des effets perturbateurs comme un bruit de fond sonore peut influencer la prononciation du locuteur par une élévation de la voix [Lombard, 1911]. Dans ce cas, les propriétés spectrales du signal de parole changent. En effet, à ce moment, la largeur moyenne des bandes de fréquence des formants diminue tandis que l'amplitude et la durée des voyelles ainsi que la hauteur de la fréquence fondamentale augmentent [Junqua, 1995]. Ces effets perturbateurs ne dépendent pas de la prononciation du message linguistique mais peuvent avoir une influence sur l'évolution de cette prononciation. L'ensemble de ces effets extérieurs au locuteur modifiant le signal de parole d'origine est rassemblé sous le terme de variabilité extrinsèque.

## 4.2 Variabilité extrinsèque

La variabilité extrinsèque de la parole est liée aux conditions de transmission et d'acquisition du signal de parole [Benzeghiba et al., 2006]. Comme tout signal audio, le signal de parole émis transite par un milieu intermédiaire avant d'être perçu par le système auditif [Boite et al., 2000]. Une transduction de la pression acoustique est alors effectuée dans l'oreille interne humaine [Calliope, 1989]. Lors d'un enregistrement audio, la variation de pression est captée par un microphone puis convertie en une grandeur électrique [Mariani, 2002]. Ce milieu intermédiaire composé de l'air puis le cas échéant du matériel d'enregistrement et de restitution n'est pas neutre. On considère alors comme principe général que tout canal de transmission contient des sources de bruit perturbant le signal transmis [Shannon et al., 1949]. Les perturbations liées à la transmission du signal viennent corrompre le signal de parole émis en sortie du conduit vocal (Figure 4.2). Ces perturbations sont de divers ordres :

- des bruits additifs peuvent s'ajouter au signal de parole. Ces bruits additifs peuvent être dus à la qualité de transmission (bruit aléatoire) ou à la superposition d'évènements audio additionnels (environnement, mélange de voix) [Grenier et al., 1981].

- des bruits convolutifs peuvent modifier la forme de l'onde du signal de parole par des effets acoustiques de type écho par exemple [Hermansky et al., 1993; Ehlers et al., 1997]. Les perturbations dues au canal de transmission du signal de parole, par exemple par le passage à travers un réseau téléphonique filaire ou GSM, sont également génératrices de bruit convolutif [Hermansky et al., 1999].

- l'enregistrement puis la restitution du signal de parole peuvent modifier ce signal suite à sa conversion sous forme d'onde électrique, sous forme analogique ou encore sous forme numérique (fonction de transfert du microphone, numérisation) [Menéndez-Pidal et al., 2001].

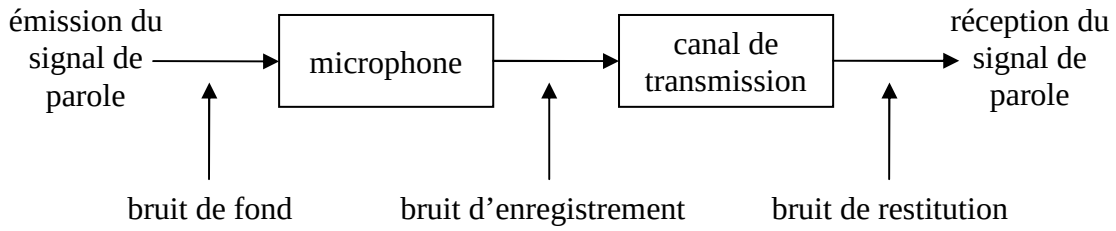



Figure 4.2 : Chaîne générale de bruitage d'un signal de parole

Un bruit additif peut être distingué lors de la présence d'un bruit de fond perturbant le signal de parole initial pendant sa transmission. Si ce bruit de fond est considéré stationnaire à moyen terme, tel un bruit gaussien, alors on admet que la répartition de son énergie est constante à travers l'ensemble de l'amplitude fréquentielle selon l'échelle temporelle utilisée [Treurniet et al., 1994]. Dans ce cas, à l'échelle de la parole, un bruit additif est considéré stationnaire dès qu'il est stable en fréquence à partir d'une échelle de grandeur de l'ordre de 200 ms [Boite et al., 2000]. A cette échelle, le bruit de fond peut se détecter par la présence d'énergies moyennes cumulatives dans certaines plages de fréquence [Hellwarth et al., 1968]. A ce moment, des techniques de compensation du signal de parole peuvent être mises en œuvre afin d'annuler la présence de ces énergies moyennes [Chen et al., 2001]. Cependant, un bruit additif créant une perturbation évoluant en fréquence à cette échelle est difficile à discriminer. Parmi ces autres types de bruit additifs, le bruit de type impulsion est caractéristique par sa forme théorique d'impulsion de Dirac tels un bruit de marteau piqueur ou celui d'un claquement de porte [Vaseghi et al., 1995]. Par ailleurs, l'intervention simultanée d'autres locuteurs que celui porteur du message dans le signal de parole est également considérée comme un bruit perturbateur nécessitant une adaptation du système de RAP [Divoux et al., 1990]. Cette interférence est connue sous le nom d'effet « cocktail party » [Hong et al., 2000]. Ce type de bruit est difficilement détectable car ses caractéristiques spectrales et temporelles sont proches de celle du signal de parole à analyser [Denbigh et al., 1994].

Un bruit convolutif peut être distingué lors de la présence d'effets d'écho, de réverbération, de délai ou encore lors d'une modification du signal de parole par une fonction de transfert perturbatrice [Claes et al., 1996]. L'ensemble de ces effets produit un mélange acoustique du signal de parole initial. Par exemple, la réverbération, appelée effet de salle, résulte en majeure partie d'un mélange audio à partir du signal de parole d'origine et de la réflexion des ondes sonores de ce signal sur les parois de l'environnement [Harris et al., 1990]. Ces caractéristiques sont liées à la configuration géométrique de la salle et à la capacité d'absorption acoustique de la nature des matériaux composant les surfaces. Le traitement adéquat de ces effets nécessite alors une analyse particulière de l'harmonicité du signal résultant [Culling et al., 1994; Kingsbury, 1998].

Lors de l'enregistrement, un effet de compression du signal de parole peut apparaître dans la chaîne d'acquisition [Menéndez-Pidal et al., 2001]. Cet effet génère une réduction de la dynamique du signal. A ce moment, les faibles niveaux de pression acoustique du signal de parole restent inchangés alors que les hauts niveaux de pression acoustique sont réduits en fonction d'une courbe de filtre statique. Lors de cette dégradation, la dynamique du signal de parole est amoindrie, détériorant ainsi la détection des formants. Par ailleurs, l'effet

d'égalisation modifie le signal de parole d'origine en atténuant ou en amplifiant certaines fréquences du signal [Mauuari, 1998]. Ainsi les rapports d'énergie entre sous-bandes de fréquence du signal de parole peuvent être altérés par l'application d'un tel effet d'égalisation.

De surcroît, la numérisation du signal de parole durant son acquisition convertit ce signal continu en une séquence de nombres binaires [Young et al., 2006]. Il s'agit de mesurer à des intervalles de temps réguliers l'amplitude de l'onde acoustique produite par le signal de parole. Cette numérisation s'effectue en deux temps. Tout d'abord, un échantillonnage permet de découper de manière régulière le signal de parole acquis lors de l'enregistrement. Une séquence d'échantillons successifs permet alors de représenter la forme d'onde du signal acoustique. Il est donc nécessaire d'adapter la fréquence d'échantillonnage afin de conserver les caractéristiques utiles de la forme d'onde originale [Boite et al., 2000]. Lors d'une analyse spectrale, le signal de parole analysé est représenté sous la forme d'une somme de sinusoïdes. Il est alors nécessaire de tenir compte de l'effet de crénelage produit par un repli du spectre durant cette opération afin d'éviter toute confusion dans la représentation par sinusoïdes. Ainsi, suivant le théorème de Nyquist-Shannon, la fréquence d'échantillonnage doit être au moins égale au double de la plus grande des fréquences composant le signal utile [Shannon, 1949]. Afin d'assurer la restitution de la plage complète de fréquence utile du signal de parole autour de 4 kHz, la fréquence d'échantillonnage doit être supérieure ou égale à 8 kHz. Dans un second temps, les échantillons du signal de parole sont conservés sous la forme de valeurs binaires. Ces valeurs binaires sont obtenues par une quantification scalaire de la valeur des échantillons. Ainsi, la dynamique du signal de parole est représentée par ces valeurs quantifiées.

Le stockage du signal de parole peut également s'effectuer par l'usage d'outils de compression avec perte, comme par exemple la sauvegarde sous forme de fichier au format MP3 [Shlien, 1994; Rault et al., 1995]. Ce type de compression dite destructrice est réalisé en perdant une certaine partie de l'information du signal [Le Guyader et al., 2000]. Ainsi, le signal obtenu est différent du signal de parole d'origine échantillonné. Les techniques utilisées à cet effet visent à analyser le signal afin de déterminer les sons inaudibles à l'oreille humaine pour les supprimer [Pan, 1995]. En théorie, les caractéristiques utiles du signal de parole ne devraient pas être affectées par ce type de compression. En pratique, il est nécessaire de vérifier que la dégradation du signal audio ne détruise pas tout ou partie de l'information utile issue du signal de parole. Ainsi, un encodage de format MP3 à 64 kbps permet une réduction d'environ 25 fois la taille d'un fichier de signal audio initialement enregistré en qualité compact disque audio. Toutefois, dans ce cas, la qualité d'écoute du fichier sonore est alors dégradée.

Les perturbations liées à la variabilité extrinsèque modifient donc le signal de parole original $s(n)$ en y ajoutant des paramètres additifs et convolutifs de telle sorte que le signal résultant $s'(n)$ est obtenu par [Boite et al., 2000] :

$$s'(n) = s(n) * c(n) + a(n) \quad (4.1)$$

avec $c(n)$ la réponse impulsionnelle d'un filtre inconnu et $a(n)$ la somme des bruits additifs.

La variabilité du signal de parole et la difficulté de sa reconnaissance automatique dépendent alors de l'ensemble des possibilités de ces variabilités extrinsèque et intrinsèque. Ainsi, dans un système de RAP, le signal acoustique de parole initial émis en sortie du conduit vocal ne peut pas être directement exploité. De surcroît, ce signal de parole n'est pas une réalisation idéale de la prononciation du message linguistique émis. Il est alors nécessaire d'extraire les paramètres acoustiques utiles à la RAP à partir du signal issu de l'acquisition sonore. Ce signal contient des perturbations du signal de parole provenant autant de la variabilité intrinsèque que de la variabilité extrinsèque. Cette paramétrisation acoustique intègre des mécanismes pour rendre la RAP robuste à certains types de variabilité [Mokbel, 1992]. Dans la section suivante, diverses représentations acoustiques du signal de parole sont présentées. Ces représentations sont utilisées comme paramètres acoustiques pour les systèmes de RAP.

## 4.3 Paramétrisation acoustique

Dans un système de RAP, les paramètres acoustiques permettant de décrire le signal de parole sont généralement définis sur une échelle d'information de niveau local. Le signal continu de parole est fourni en entrée du système de RAP après une conversion sous la forme d'échantillons sonores. Une suite de vecteurs représentatifs, appelés vecteurs acoustiques ou vecteurs d'observation, est alors retournée en sortie du module de paramétrisation acoustique. Les paramètres acoustiques définis pour la représentation acoustique du signal de parole devraient respecter les critères de [Deviren, 2004] :

- pertinence. Les paramètres acoustiques doivent représenter de manière précise le signal de parole. Leur nombre doit cependant rester limité afin de conserver un coût de calcul raisonnable lors de leur exploitation dans les modules de calcul des paramètres acoustiques et de reconnaissance des formes.

- discrimination. Les paramètres acoustiques doivent représenter de manière caractéristique les différents éléments représentatifs des unités linguistiques afin de les rendre facilement distinctes.

- robustesse. Les paramètres acoustiques doivent résister aux effets perturbateurs liés aux distorsions du signal de parole émis [Milner et al., 2011].

Dans le processus de traitement du signal acoustique d'un système de RAP, un découpage du signal de parole analysé retourne une séquence de segments d'échantillons sonores appelés trames. La durée de ces trames est choisie de telle sorte que le signal de parole est considéré stationnaire [Boite et al., 2000]. Cette segmentation permet alors d'extraire les propriétés locales du signal de parole. Le continuum de parole est donc représenté par une suite de vecteurs d'observation calculés sur des trames du signal de courte durée par exemple de l'ordre de 20 ms, par fenêtre glissante asynchrone ou synchrone au pitch [Young et al., 2006]. Les vecteurs d'observation peuvent représenter le signal de parole sous la forme de différents types de coefficients qui constituent les paramètres acoustiques. Ces paramètres sont choisis pour être le plus utile à la représentation du signal de parole dans l'objectif de décrire le message linguistique. Se basant sur l'analyse des caractéristiques physiologiques de l'oreille [Dallos, 1973], de nombreux types de paramètres acoustiques sont utilisés dans la littérature pour la RAP [Davis et al., 1980; Mokbel, 1992; Eyben et al., 2010]. Parmi les principaux types de paramètres exploités dans les systèmes de RAP, on peut distinguer :

- les coefficients d'analyse temporelle. Ces paramètres acoustiques permettent de représenter les variations de pression acoustique. L'analyse du taux de passage à zéro (*Zero Crossing Rate,* ZCR) fournit des indications sur les transitions entre signaux de parole voisés et non-voisés [Chen, 1988]. De surcroît, l'énergie du signal de parole est souvent utilisée pour détecter les zones de silence ou pour distinguer les zones de parole de celles de musique par exemple [Saunders, 1996; Scheirer et al., 1997; Panagiotakis et al., 2005]. En effet, l'énergie à court terme du signal de parole est beaucoup plus variable que celle d'un signal de musique [Scheirer et al., 1997].

- les coefficients spectraux. Ces paramètres acoustiques représentent les éléments distinctifs du spectre du signal. Il s'agit le plus souvent d'extraire les informations d'énergie contenue dans différentes sous-bandes fréquentielles du spectre à court terme d'une trame donnée. Par exemple, les paramètres LSFs (*Line Spectral Frequencies*) permettent de bien représenter les structures de formant et possèdent de bonnes propriétés d'interpolation [Paliwal et al., 1991]. Dans d'autres méthodes, les paramètres TRAPs (*TempoRAl Patterns*) sont définis sur des segments de trames [Hermansky et al., 1998]. Ces paramètres TRAPs sont des vecteurs temporels de l'ordre de la seconde contenant les valeurs d'énergie de chaque trame pour une sous-bande de fréquence donnée [Hermansky et al., 1999; Grézl et al., 2007].

- les coefficients cepstraux. Ces paramètres acoustiques permettent de caractériser le signal de parole par une discrimination de la réponse fréquentielle à partir d'une analyse issue d'opérations sur le spectre du signal [Holden et al., 1976]. Le plus souvent, cette discrimination est obtenue en utilisant une échelle de fréquence non linéaire représentative de l'intégration fréquentielle de la perception auditive humaine [Juang et al., 1987]. Un ensemble de paramètres cepstraux est très largement utilisés en RAP, il s'agit des paramètres MFCCs (*Mel Frequency Cepstral Coefficients*) [Picone, 1993] basés sur les échelles de Mel.

- les coefficients par codage prédictif. Ces paramètres acoustiques sont construits à partir d'une modélisation du signal par prédiction linéaire par exemple de type LPC (*Linear Predictive Coding*) [Atal et al., 1974; Makhoul, 1975; Tremain, 1982; Kroon et al., 1992]. Dans le domaine cepstral, la prédiction linéaire prend la forme de modèle LPCC (*Linear Predictive Cepstral Coding*) [Miet, 2001]. Dans le domaine spectral, on retrouve les modèles de paramètres acoustiques de type :
    - PLP (*Perceptual Linear Predictive*) [Hermansky, 1990; Woodland et al., 1993],
    - RASTA PLP (*RelAtive SpecTrAl PLP*) [Hermansky et al., 1991; Hermansky et al., 1994],
    - NPC (*Neural Predictive Coding*) [Gas et al., 2000].

Dans le cadre de cette étude, seuls les paramètres acoustiques issus de systèmes de RAP de type énergie en sous-bande et de type MFCC seront utilisés. L'espace de représentation du signal de parole ainsi obtenu est muni d'une mesure de distance euclidienne adaptée à ces paramètres acoustiques. Cette mesure de distance est utilisée comme critère de similarité au sein de l'algorithme de comparaison du système de RAP considéré.

### 4.3.1 Energie en sous-bande

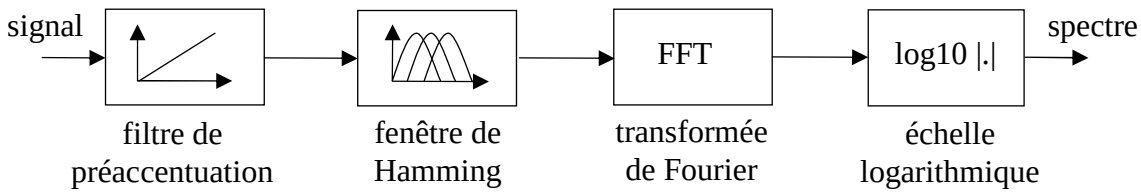



Figure 4.3 : Schéma fonctionnel du calcul du spectre d'un signal de parole

Afin d'obtenir des paramètres acoustiques à partir de l'analyse de l'énergie contenue dans le spectre, des séquences d'échantillons du signal de parole sont regroupées en fenêtres de quelques dizaines de millisecondes. La taille de ces fenêtres est choisie telle que le signal est considéré localement quasi-stationnaire. Une opération de préaccentuation du signal de parole permet d'accentuer les hautes fréquences du signal et de filtrer ainsi la composante continue afin de compenser l'effet de radiation aux lèvres [Boite et al., 2000] (Figure 4.3). Cette préaccentuation s'applique avec un coefficient $\alpha$ sur les $n$ échantillons $x$ du signal vocal telle que :

$$x_i = x_i - \alpha . x_{i-1}, \forall i \in [2; n] \tag{4.2}$$

Ce coefficient $\alpha$ est souvent défini à $\alpha = 0{,}97$ [Young et al., 2006]. Un recouvrement partiel des fenêtres d'échantillons permet de minimiser les sauts de valeurs calculées d'une fenêtre à la suivante. A cet effet, un fenêtrage de Hamming peut être par exemple appliqué [Enochson et al., 1968]. Un tel type de fenêtrage assure une fenêtre moyenne entre un lob principal étroit représentant la fréquence fondamentale du signal de parole et un amortissement modéré des lobes secondaires dus à la distorsion du signal par le fenêtrage. Le coefficient pondérateur du fenêtrage de Hamming est alors soumis à chacun des $n$ échantillons de la trame du signal de parole selon :

$$hamming(i) = 0{,}54 - 0{,}46 . \cos\left(\frac{2\pi . i}{n}\right), \forall i \in [1; n] \tag{4.3}$$

Un autre type de fenêtrage populaire est la fenêtre de Hann, appelée également fenêtre de Hanning [Blackman et al., 1959]. Le coefficient pondérateur du fenêtrage de Hann soumis à chacun des $n$ échantillons de la trame du signal de parole s'exprime alors comme :

$$hann(i) = 0{,}50 - 0{,}50 . \cos\left(\frac{2\pi . i}{n}\right), \forall i \in [1; n] \tag{4.4}$$

L'analyse spectrale du signal de parole est obtenue par transformée de Fourier rapide (*Fast Fourier Transform*, FFT), calculée à court terme [Cooley et al., 1965]. Cette technique permet de représenter la puissance acoustique des différentes fréquences composant le signal audio. Le spectrogramme résultant représente ainsi la distribution énergétique dans le plan temps/fréquence. Cependant, le spectre ne peut pas être directement utilisé comme paramètre

acoustique à cause du nombre élevé de dimensions qui le composent. On considère alors l'énergie $E(n)$ correspondant à la puissance du signal de parole contenue dans les $n$ échantillons $x$ de la trame $t$ du signal de parole telle que :

$$E(t) = \sum_{i=1}^{n} x_i^2 \tag{4.5}$$

Dans le cas particulier d'un découpage du spectre du signal de parole en sous-bandes fréquentielles, les énergies en sous-bandes sont obtenues en calculant l'énergie du signal de parole contenue dans chacune des sous-bandes données. L'énergie globale contenue dans les trames du signal de parole est alors obtenue par la somme de ces énergies en sous-bandes fréquentielles.

### 4.3.2 Paramètre MFCC

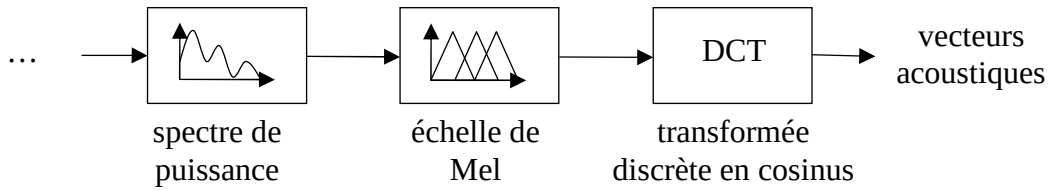



Figure 4.4 : Schéma fonctionnel du calcul des paramètres acoustiques de type MFCC

Un vecteur acoustique MFCC est formé de coefficients cepstraux obtenus à partir d'une répartition fréquentielle selon l'échelle de Mel [Bogert et al., 1963] (Figure 4.4). L'utilisation d'échelles de fréquence non-linéaires, telles les échelles de Mel [Stevens et al., 1937] ou Bark [Zwicker, 1961], permettent une meilleure représentation des basses fréquences qui contiennent l'essentiel de l'information linguistique pour la majeure partie du signal de parole. La correspondance entre les valeurs de fréquence en Hertz $F_{Hertz}$ et en Mel $F_{mel}$ est calculée par [O'Shaughnessy, 1987] :

$$F_{mel} = 2595.log\left(1 + \frac{F_{Hertz}}{700}\right) \tag{4.6}$$

Par ailleurs, il est possible de calculer des coefficients cepstraux à partir d'une répartition fréquentielle linéaire sans utiliser une échelle de Mel mais en conservant la répartition linéaire des échelles de fréquence. Ces coefficients sont alors appelés LFCCs (*Linear Frequency Cepstral Coefficients*) [Rabiner et al., 1993].

Afin de séparer la source spectrale de la réponse fréquentielle, l'opération de méthode cepstrale se base sur la propriété du logarithme qui permet de transformer un produit en addition. Une transformée discrète en cosinus (*Discret Cosinus Transform*, DCT) permet ainsi d'obtenir les $N$ coefficients cepstraux désirés [Ahmed et al., 1974]. Considérant $f$ la fonction de transformation spectrale, le $k^{\text{me}}$ coefficient cepstral $C(k)$ est donc obtenu par :

$$C(k) = \sqrt{\frac{2}{N}} \sum_{i=1}^{N} f(i).\cos\left(\frac{\pi k}{N}(i - 0.5)\right) \tag{4.7}$$

Cette analyse a pour avantages un nombre réduit de coefficients par vecteur acoustique et un faible indice de corrélation entre ces différents coefficients. Les coefficients MFCCs sont réputés plus robustes que ceux issus d'une analyse spectrale [Lockwood et al., 1992]. De nombreux systèmes de RAP exploitent ce type de paramètres acoustiques [Schurer, 1994; Mokbel et al., 1995; Fineberg et al., 1996; Chatterjee et al., 2011]. Les coefficients de type MFCC sont souvent associés à la valeur d'énergie contenue dans la trame de signal de parole appelée sous le terme de coefficient $C(0)$ [Young et al., 2006]. De surcroît, l'utilisation des dérivées premières et secondes de ces coefficients fournit de l'information utile sur la dynamique du signal de parole. En effet, l'information complémentaire apportée par le filtrage temporel introduit par les dérivées des coefficients MFCCs permet une plus grande robustesse des paramètres acoustiques dans les systèmes de RAP face à l'usage des seuls coefficients MFCCs statiques [Yang et al., 2007]. Dans ces conditions, ces paramètres acoustiques prennent souvent la forme de vecteurs de 39 coefficients formés par les 12 premiers coefficients MFCCs, l'énergie $C(0)$ et leurs dérivées premières et secondes. Les valeurs des coefficients correspondant à ces dérivées premières sont alors estimées par développement limité d'ordre 2 telles que [Levy, 2006] :

$$c'_k(t) = \frac{c_k(t+1) - c_k(t-1) + 2.\big(c_k(t+2) - c_k(t-2)\big)}{10} \tag{4.8}$$

avec $c_k(t)$ le $k^{\text{me}}$ coefficient du vecteur acoustique de la trame $t$ et $c'_k(t)$ sa dérivée première.

Les dérivées secondes sont calculées de la même manière à partir des dérivées premières résultantes. Par ailleurs, ce calcul de dérivée est généralisable et valable quelque soit le type de vecteur acoustique choisi.

En revanche, la représentation cepstrale conventionnelle du signal de parole par les seuls coefficients MFCCs ne conserve pas l'information de la phase obtenue par estimation de l'énergie du spectre [Leonard, 1984]. Cette perte d'information ainsi que la réduction de la résolution spectrale diminue la quantité d'information utile disponible pour la RAP. Bien que

l'information de phase soit retrouvée par l'apport du coefficient additionnel $C(0)$, elle semble facultative à la bonne reconnaissance de nombreux phonèmes. Cette information complémentaire apporte toutefois un complément utile dans la classification de certaines consonnes [Liu et al., 1997]. Par ailleurs, il est possible de re-synthétiser un message intelligible sur de la parole propre à partir d'une analyse des seuls coefficients MFCCs, c'est à dire à partir des spectres et cepstres en échelle de Mel [Demuynck et al., 2004]. Donc dans le cas de parole propre, un signal d'excitation basé sur une analyse du pitch est utilisé pour cette opération de re-synthèse [Collen et al., 2007]. Dans ce cas, l'information initiale de phase n'est alors pas utile. Par contre, dans le cas d'un signal de parole bruitée, les informations de phase et de résolution spectrale fine sont très utiles pour la bonne reconnaissance des composantes du message linguistique [Peters et al., 1999; Murty et al., 2006].

## 4.4 Choix des paramètres et gestion de la variabilité

Compte-tenu des multiples réalisations acoustiques possibles d'un même phonème [Lindblom, 1990], la prise en compte de la variabilité comme facteur discriminant [Elman et al., 1986] est préférée aux méthodes de recherche des invariants dans le signal de parole [Stevens et al., 1978; Blumstein, 1986]. Les systèmes de RAP développent donc diverses approches de représentation du signal de parole par paramètres acoustiques afin de maximiser la robustesse de ces paramètres aux effets de variabilités extrinsèque et intrinsèque [Benzeghiba et al., 2007].

### 4.4.1 Discrimination du signal stationnaire

Le choix de la taille de la fenêtre d'analyse de la trame est fixé en considérant que le signal de parole est quasi-stationnaire sur la durée de la trame [Boite et al., 2000]. Par exemple, les coefficients de type MFCC sont basés sur une représentation issue de l'enveloppe spectrale (section 4.3.2, page 84). Cette enveloppe spectrale est estimée généralement sur des fenêtres d'analyse du signal de parole d'une durée autour de 20 à 30 ms [Davis et al., 1980; Rabiner et al., 1993]. Cependant, le signal de parole correspondant à des phonèmes voisés, comme les voyelles par exemple, est quasi-stationnaire sur une durée de 40 à 80 ms alors que la durée du signal de parole correspondant à une occlusive ne dépasse généralement pas les 20 ms [Wang et al., 1996]. En l'absence de méthode d'analyse du signal de parole sur des résolutions multiples [Caraty et al., 1998], le compromis sur le choix de la taille de la trame implique donc certaines limitations. En effet, pour une trame d'une durée de 20 ms, la résolution fréquentielle est relativement faible pour analyser des signaux quasi-stationnaires longs par rapport à des trames de durée plus longue. Par ailleurs, un choix de taille de fenêtre d'analyse trop grand peut diminuer la résolution nécessaire à l'étude d'un signal de parole dont la quasi-

stationnarité est considérée courte [Haykin, 1993]. Il est alors possible de faire appel à un processus gaussien autorégressif qui permet une modélisation des segments quasi-stationnaires du signal de parole [Svendsen et al., 1987; André-Obrecht, 1988; Svendsen et al., 1989; Tyagi et al., 2005b]. Dans ce cas, le critère du maximum de vraisemblance est utilisé pour discriminer les points de transition entre les différents segments successifs.

Une autre approche consiste à représenter la variation continue des paramètres LPC par une technique de décomposition temporelle. Cette décomposition prend alors la forme d'un polynôme à coefficients pondérés [Atal, 1983; Deléglise et al., 1988]. Cependant, les composants du polynôme ne représentent pas forcément les éléments quasi-stationnaires du signal de parole [Montacié et al., 1992]. D'autres méthodes ont ainsi été développées pour améliorer la qualité des paramètres acoustiques choisis dans la représentation de la stationnarité du signal de parole. Il existe ainsi des méthodes de paramétrisation :

- par algorithme de segmentation [Svendsen et al., 1987],
- par algorithme de sélection [Coifman et al., 1992],
- par modélisation statistique [Achan et al., 2004] (Annexe H, page 151).

Par ailleurs, diverses représentations du signal de parole s'efforcent de mieux décrire les caractéristiques de transition entre états considérés stationnaires. Ces représentations permettent une classification de motifs par analyse de la trajectoire d'énergies spectrales en sous-bandes à travers l'axe temporel. Parmi celles-ci, on distingue :

- les paramètres de motifs temporels TRAPs [Hermansky et al., 1998],
- l'usage de perceptrons multicouches [Howard et al., 1988],
- les techniques du spectre de modulation [Haykin, 1994].

Dans l'approche par spectre de modulation, il est possible d'exploiter les informations de modulation de l'amplitude (*Amplitude Modulation*, AM) et de modulation de fréquence (*Frequency Modulation*, FM) [Dimitriadis et al., 2005]. A cet effet, le spectre du signal de parole est décomposé en sous-bandes fréquentielles étroites afin d'extraire la modulation AM, typiquement autour de 4 kHz [Betser et al., 2008]. Cette approche est à comparer à une autre méthode de paramétrisation basée sur l'analyse du spectre de modulation du signal de parole [Tyagi et al., 2005]. L'étude de ce spectre issu de la modulation permet une décomposition du signal de parole en sous-bandes spectrales de plus en plus larges suivant une échelle de type Mel ou Bark [Milner, 1996; Kingsbury et al., 1998; Zhu et al., 2000; Tyagi et al., 2003]. Par ailleurs, une analyse de l'enveloppe spectrale du signal de parole est également possible pour filtrer la modulation AM [Schimmel et al., 2005].

Dans d'autres approches, il est possible de définir des paramètres acoustiques par l'étude de fonctions de transfert pôle-zéro grâce à une modélisation autorégressive du signal de parole [Makhoul, 1975]. Une telle modélisation est utilisée pour représenter la réponse fréquentielle du signal de parole. Cette méthode appliquée dans le domaine temporel permet une modélisation analytique du signal de parole [Kumaresan, 1998; Kumaresan et al., 1999]. L'usage d'une telle prédiction linéaire dans le domaine fréquentiel peut venir en soutien des paramètres TRAPs [Athineos et al., 2003]. D'autre part, l'information de la phase peut être introduite comme paramètre acoustique dans la représentation du signal de parole afin d'améliorer la classification des voyelles [Paliwal et al., 2003]. Cette information de la phase peut aussi permettre de représenter les structures de type formant avec une meilleure résolution que par la simple représentation de l'énergie du spectre [Hedge et al., 2004; Zhu et al., 2004b; Bozkurt et al., 2005]. Par ailleurs, un autre groupe de méthodes de représentation du signal de parole s'attache à rechercher et à étudier les propriétés de l'invariance du signal de parole pour la définition de paramètres acoustiques.

### 4.4.2 Invariance et compensation

Les paramètres de type PLP présentent une robustesse particulière à la variabilité intrinsèque inter-locuteur grâce à une analyse de prédiction linéaire [Hermansky, 1990]. Cette analyse modélise essentiellement les deux principaux pics énergétiques de l'enveloppe spectrale tout le long des trames issues du signal de parole. D'autres approches s'attachent à extraire des paramètres acoustiques invariants à la variation fréquentielle de ces pics énergétiques. Par exemple, il est possible de définir à partir du spectre du signal de parole une transformation d'échelle spectrale permettant ainsi d'obtenir une échelle dans le domaine cepstral ayant la propriété de conserver une magnitude invariante au spectre original [Umesh et al., 1999]. Par ailleurs, une fonction de transformation en ondelette peut être utilisée pour générer des paramètres d'invariants en fréquence dynamique [Favero, 1994]. Ces différents paramètres représentent les invariants du conduit vocal dont la longueur est liée à chaque locuteur. Ils peuvent être combinés aux paramètres MFCCs afin d'y apporter des informations complémentaires pour améliorer les performances de reconnaissance d'un système de RAP [Mertins et al., 2005].

Dans d'autres approches encore, une méthode de normalisation de la longueur du conduit vocal (*Vocal Tract Length Normalization*, VTLN) a pour objectif de modéliser les conduits de résonance en réduisant le plus possible la dépendance aux facteurs physiologiques liés au locuteur [Welling et al., 2002]. Différentes techniques sont issues de cette méthode de normalisation afin de modifier la position des formants lors de l'analyse du signal de parole.

Ces techniques utilisent différents procédés permettant de se rapprocher d'un modèle moyen de locuteur, comme dans :

- la cartographie des formants liée au locuteur [Wakita, 1977; Di Benedetto et al., 1992],
- la transformation de la modélisation de pôle LPC [Slifka et al., 1995],
- l'analyse fréquentielle dynamique linéaire [Tuerk et al., 1993; Eide et al., 1996; Lee et al., 1996; Zhan et al., 1997],
- l'analyse fréquentielle dynamique non-linéaire [Ono et al., 1993],
- l'étude particulière des voix d'enfants [Das et al., 1998].

Il existe par ailleurs des techniques de compensation du canal comme la soustraction de la moyenne cepstrale ou le filtrage des trajectoires des énergies du spectre en sous-bandes fréquentielles issues des paramètres de type RASTA PLP [Hermansky et al., 1994; Westphal, 1997; Kajarekar et al., 1999]. Ces techniques permettent de s'abstraire des composantes acoustiques liées au locuteur sur le spectre à long terme. D'autres techniques de compensation s'appuient quant à elles sur la transformation des paramètres acoustiques comme les paramètres cepstraux par exemple [Chen, 1987; Hunt et al., 1989; Hanson et al., 1990; Hansen, 1996]. Il est également possible de combiner plusieurs types de paramètres acoustiques ensemble en vue d'améliorer les performances d'un système de RAP.

### 4.4.3 Combinaison de paramètres et sélection

Plusieurs flux de paramètres acoustiques peuvent être calculés sur un même signal de parole afin de décrire différentes propriétés acoustiques de ce signal. Ces paramètres peuvent ensuite former des combinaisons entre eux pour améliorer les performances du système de RAP [Zolnay et al., 2005]. En effet, le choix de paramètres acoustiques additionnels à un ensemble de paramètres initiaux peut être considéré comme un élément récupérateur d'erreur lors de l'échec de la reconnaissance du signal de parole analysé [Hirschberg et al., 2004]. Ces paramètres acoustiques sont alors choisis en complément les uns des autres pour être les meilleurs discriminants dans une classification automatique des unités linguistiques définies. En particulier, les algorithmes de RAP sont généralement développés pour répondre aux problèmes de variabilité intrinsèque dans la classification phonétique [Meyer et al., 2010]. En pratique, des solutions sous-optimales sont proposées en sélectionnant un ensemble de paramètres issus de mesures acoustiques. Ces solutions garantissent une forte quantité d'information mutuelle entre cette représentation de mesures acoustiques et les paramètres acoustiques de discrimination phonétique [Omar et al., 2002; Omar et al., 2002b]. Ainsi la recherche d'ensembles de paramètres avec différentes propriétés de représentation acoustique du signal de parole est importante (Annexe I, page 153).

De surcroît, la recherche de transformations optimales pour réduire le nombre de dimensions d'un vecteur acoustique est un domaine d'étude important. En effet, le choix de paramètres acoustiques optimaux est obtenu en maximisant l'information mutuelle entre l'ensemble des paramètres acoustiques définis et les classes d'unités linguistiques correspondantes [Omar et al., 2002b; Padmanabhan et al., 2005] (Annexe J, page 155).

## 4.5 Discussions

La problématique de la gestion de variabilité en RAP est un sujet très vaste. Pour une part, le traitement de la gestion de la variabilité extrinsèque du signal de parole est un sujet commun avec celui de la gestion du bruit dans le domaine du traitement du signal audio [Peeters et al., 2009]. Mais par ailleurs, une variabilité spécifique au signal de parole est définie par la prise en compte des aspects liés au locuteur et aux conditions de production de la parole [Benzeghiba et al., 2006]. Ainsi la problématique considérée de la variabilité en RAP traite également de la gestion de cette variabilité spécifique dite variabilité intrinsèque [Benzeghiba et al., 2007].

Les causes de variabilité intrinsèque du signal de parole sont nombreuses [Benzeguiba et al., 2006b]. Tout d'abord, la structure du signal de parole est affectée par les caractéristiques individuelles du locuteur [Harmegnies et al., 1988]. De plus, la modulation à moyen terme de la voix peut être modifiée involontairement par la transmission d'émotions [Ververidis et al., 2006] ou intentionnellement afin d'émettre des informations linguistiques de haut niveau associées au message transmis [Shafran et al., 2000]. Ces effets font partie intégrante de la communication humaine et sont donc très importants. Enfin, la prononciation idéale ou désirée d'un message linguistique peut être altérée par les différentes possibilités de réalisation acoustique des unités linguistiques constituant le corps du message [Nedel, 2004].

Les paramètres acoustiques adaptés au signal de parole sont définis pour la plupart pour représenter efficacement l'aspect non-stationnaire du signal [Levinson et al., 1983]. De surcroît, l'étape de calcul de ces paramètres peut être appropriée pour la discrimination d'autres types de critères et leur traitement. Ces critères peuvent être par exemple la prise en compte des effets de variation dus à la physiologie du locuteur par méthode de compensation [Welling et al., 2002] ou l'amélioration de la prise en compte de l'invariance [Mertins et al., 2005].

La combinaison de paramètres acoustiques multiples améliore également la robustesse de la RAP au détriment de la complexité de la mise en œuvre d'un système adapté [Hirschberg et al., 2004]. Les approches de réduction du nombre de dimensions des vecteurs acoustiques prennent alors toute leur importance pour la diminution de la complexité du

traitement de ces vecteurs tout en conservant leurs qualités de discrimination [Padmanabhan et al., 2005].

Ainsi, la recherche de nouveaux paramètres acoustiques est motivée par le désir d'amélioration des performances d'un système de RAP pour répondre à une tâche particulière [Meyer et al., 2011]. Dans ce cadre, il est indispensable d'évaluer la robustesse aux différents types de variabilité de paramètres acoustiques issus de l'identification audio par empreinte au sein d'un système de RAP, particulièrement dans la tâche d'identification acoustico-phonétique.

# Chapitre 5. Robustesse de sous-empreintes aux variabilités extrinsèque et intrinsèque

Dans le domaine de la RAP, les systèmes de reconnaissance sont développés dans le but d'extraire l'information linguistique contenue dans un signal de parole malgré la présence de perturbations issues des variabilités extrinsèque et intrinsèque [Benzeghiba et al., 2006].

Le paradigme d'évaluation et les expériences décrites dans ce chapitre visent à mesurer la robustesse de diverses représentations acoustiques du signal de parole face à de telles variabilités. En particulier, la robustesse de différents types de sous-empreinte issus de l'identification audio est évaluée par rapport à celle de paramètres MFCCs (section 4.3.2, page 84). Dans les évaluations précédentes sur le DAP (section 3.5, page 57), en excluant les méthodes supervisées, la méthode QV est celle retournant des sous-empreintes avec les meilleures performances. Cette méthode de calcul de sous-empreinte QV est réutilisée pour les prochaines expériences. Ces types de sous-empreinte sont alors basés sur :

- la méthode de type Philips [Haitsma et al., 2001] (section 1.3, page 17),
- la méthode de type Shazam [Wang, 2003] (section 1.4, page 22),
- la méthode QV basée sur des paramètres MFCCs quantifiés (section 3.3.2, page 50).

Deux paradigmes expérimentaux sont définis pour notre système développé dans une tâche d'identification acoustico-phonétique. Le premier paradigme consiste à évaluer la robustesse à la variabilité extrinsèque des vecteurs définis. Le second paradigme consiste à évaluer la robustesse de ces vecteurs aux variabilités extrinsèque et intrinsèque combinées. Dans un premier temps, le protocole d'évaluation et les bases de données utilisées sont décrites. Puis quatre types de vecteurs sont définis comme paramètres de représentation acoustique du signal de parole. Diverses mesures de distance adaptées leur sont alors associées afin de pouvoir comparer ces vecteurs selon leur critère de similarité respectif. Des expériences sont ensuite effectuées pour évaluer la robustesse de ces vecteurs sur les diverses formes de variabilité en fonction des deux paradigmes considérés. Enfin, la robustesse de ces différents types de vecteurs est discutée selon le type de variabilité considéré.

## 5.1 Protocole d'évaluation

Deux paradigmes expérimentaux sont présentés pour mesurer l'influence de la variabilité extrinsèque puis des variabilités extrinsèque et intrinsèque combinées sur la robustesse de vecteurs à travers la capacité du système à identifier un signal de parole de test.

### 5.1.1 Paradigmes d'évaluation

Le premier paradigme consiste à évaluer la robustesse de vecteurs à la variabilité extrinsèque. Dans ce paradigme, une base de données $A_p$ de parole propre est utilisée pour l'apprentissage. Une version de cette base de données $A_a$ est définie en portant des altérations sur le signal de parole, représentatives de perturbations liées à la variabilité extrinsèque (section 4.2, page 77). Cette base de données altérée $A_a$ est utilisée pour le test. L'évaluation de ce paradigme est basée sur une mesure de distance entre un vecteur de $A_a$ et le vecteur correspondant au même moment d'apparition dans le signal de parole de $A_p$. La mesure de distance est donc effectuée entre deux vecteurs correspondant à une même réalisation acoustique de signal de parole. Cependant, certaines altérations du signal audio peuvent parfois induire une variation temporelle sur ce signal [Junqua, 1997]. Par exemple, selon les contraintes de l'environnement, la chaîne d'acquisition du signal de parole peut être mobile et se situer à différentes distances du locuteur. Dans ce cas, le délai de transmission du signal de parole du locuteur à la source d'enregistrement peut varier durant la prononciation du message. Pour tenir compte de cette contrainte, nous choisissons d'étendre cette mesure de distance entre un vecteur de test de $A_a$ et tous les vecteurs du segment phonétique correspondant dans $A_p$. Parmi toutes les mesures de distance effectuées, si la distance minimale obtenue est en dessous d'un seuil donné, le vecteur de $A_a$ est considéré identifié par l'étiquette phonétique du segment correspondant dans $A_p$ (Figure 5.1).

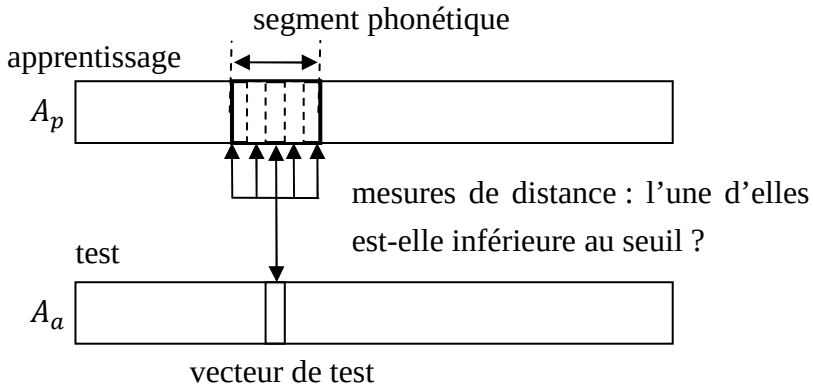



Figure 5.1 : Illustration du premier paradigme (variabilité extrinsèque)

La précision extrinsèque PE est alors obtenue par le ratio du nombre de vecteurs identifiés $V_I$ par rapport au nombre total de vecteurs de test $V_T$ telle que :

$$\mathrm{PE} = \frac{V_I}{V_T} \tag{5.1}$$

Le second paradigme consiste à évaluer la robustesse de vecteurs aux variabilités extrinsèque et intrinsèque combinées. Dans ce paradigme, une base de données $T_p$ de signal de parole propre est utilisée pour le test. Les bases de données $A_p$ et $T_p$ sont disjointes et contiennent un signal de parole prononcé par des locuteurs distincts. Les variabilités rencontrées entre ces deux bases sont d'ordre intrinsèque inter-locuteur (section 4.1, page 72). Une version altérée $T_a$ de $T_p$ est également définie comme test en portant des perturbations similaires à celles appliquées dans $A_a$. L'évaluation de ce paradigme porte sur la mesure de précision moyenne d'une identification acoustico-phonétique au niveau local. Cette évaluation est basée sur la mesure de distance entre un vecteur de test de $T_p$ ou de $T_a$ et tous les vecteurs d'apprentissage de $A_p$. Chaque fois que la distance obtenue entre les vecteurs est inférieure à un seuil donné, l'étiquette phonétique associée au vecteur d'apprentissage est retournée afin d'identifier le vecteur de test (Figure 5.2).

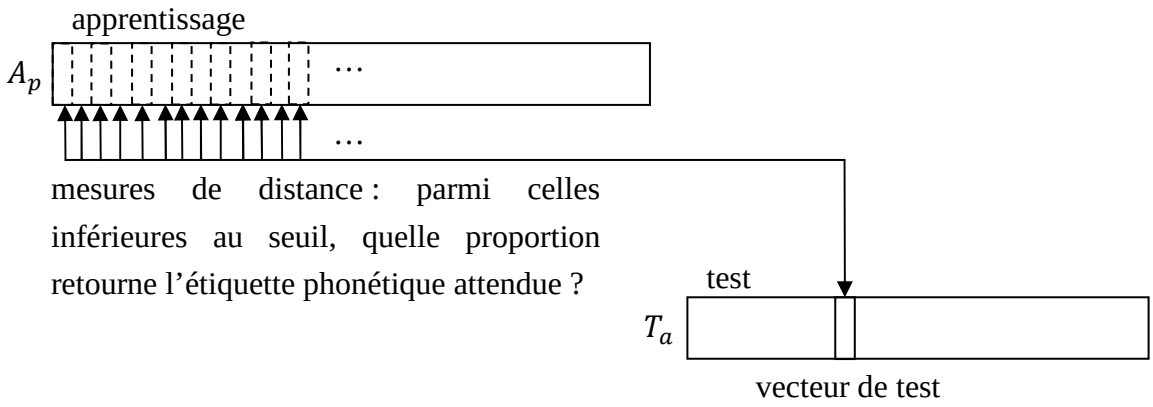



Figure 5.2 : Illustration du second paradigme (variabilités extrinsèque et intrinsèque combinées)

Une mesure de précision locale d'identification acoustico-phonétique est obtenue pour chaque vecteur de test. Cette précision locale est donnée par le ratio du nombre d'étiquettes phonétiques retournées $V_A$ identifiant correctement le vecteur de test par rapport au nombre total d'étiquettes phonétiques retournées $V_R$. Considérons alors $N$ le nombre de vecteurs de test, la précision moyenne PM est retournée par la moyenne de l'ensemble des précisions locales telle que :

$$\mathrm{PM} = \frac{1}{N}\sum_{i=1}^{N}\frac{V_A(\mathrm{i})}{V_R(i)} \qquad (5.2)$$

### 5.1.2 Bases de données

Dans les expériences suivantes, les bases de données utilisées sont TIMIT et NTIMIT reprises de l'expérience précédente (section 3.5.1, page 57) ainsi que CTIMIT [Brown et al., 1995] (Annexe B, page 135). La base TIMIT est considérée comme une base de données de signal de parole propre. La base NTIMIT est définie à partir des données de TIMIT dont le signal de parole a subi une transformation par un filtre convolutionnel. Différents types de filtres sont appliqués afin de représenter la transmission de ce signal de parole à travers divers canaux de réseau téléphonique. La base CTIMIT est composée d'un sous-ensemble de TIMIT dont le signal de parole est réenregistré dans un environnement bruité. Deux autres bases de données sont construites afin de mesurer la robustesse des vecteurs à des perturbations du signal audio similaires à celles rencontrées dans une application d'identification audio par empreinte. La première, appelée MP3 TIMIT, est conçue à partir d'une version de la base TIMIT ayant subie une compression audio avec perte [Le Guyader et al., 2000], sous la forme d'une conversion au format MP3 à 64 kbps [Shlien, 1994]. La seconde, appelée N-MP3 TIMIT, est une version bruitée de la base TIMIT obtenue par ajout d'un bruit blanc gaussien à un rapport signal à bruit de 20 dB sur la plage de fréquences de 0 à 16 kHz, puis par conversion au format MP3 à 64 kbps. L'ensemble phonétique choisi est formé de 39 phonèmes, suivant la définition du CMU/MIT [Lee et al., 1989b].

L'ensemble d'apprentissage de TIMIT prédéfini fait référence à $A_p$ (200 mn, 462 locuteurs). Les bases de données NTIMIT, CTIMIT, MP3 TIMIT et N-MP3 TIMIT sont considérées comme les bases de signal de parole altéré en référence à $A_a$ et $T_a$. Afin de tenir compte de la taille limitée de la base CTIMIT, les ensembles de test sont restreints aux parties communes du signal de parole dans les différentes bases pour permettre une évaluation commune. Les ensembles de données en référence à $A_a$ (105 mn, 462 locuteurs) sont donc définis à partir d'un sous-ensemble de $A_p$. Les ensembles de données en référence à $T_a$ (40 mn, 168 locuteurs) sont formés en suivant la même contrainte. Cette segmentation d'ensemble de données est appliquée à TIMIT pour former un ensemble de test en référence à $T_p$ (40 mn, 168 locuteurs). Nous considérons alors pour cette base TIMIT de test que durant la transmission du signal, son acquisition et sa restitution, aucun facteur extrinsèque ne vient altérer le signal de parole propre initialement émis en sortie du conduit vocal. Ces différents ensembles de bases de données et leur correspondance par rapport aux bases de référence sont représentés dans un tableau récapitulatif (Tableau 5.1).

| usage, référence et durée / base de données | apprentissage | test | | |
|---|---|---|---|---|
| | | extrinsèque | intrinsèque | intrinsèque et extrinsèque |
| | $A_p$ | $A_a$ | $T_p$ | $T_a$ |
| TIMIT | 200 mn | - | 40 mn | - |
| MP3 TIMIT | - | 105 mn | - | 40 mn |
| N-MP3 TIMIT | - | 105 mn | - | 40 mn |
| NTIMIT | - | 105 mn | - | 40 mn |
| CTIMIT | - | 105 mn | - | 40 mn |

$A_p$ et $T_p$ : bases de données disjointes de signal de parole propre
$A_a$ et $T_a$ : versions des bases de données avec un signal de parole altéré

Tableau 5.1 : Ensembles de données utilisés en fonction du type de variabilité pour les expériences

### 5.1.3 Vecteur de représentation acoustique

Dans cette étude, quatre méthodes de calcul de vecteurs sont définies par :

- adaptation de la méthode d'identification audio par empreinte de Philips (AP),
- adaptation de la méthode d'identification audio par empreinte de Shazam (AS),
- calcul de paramètres MFCCs (MFCC),
- calcul de paramètres MFCCs suivi d'une quantification vectorielle (QV).

En se basant sur les techniques d'identification audio par empreinte, les méthodes de calcul de sous-empreinte développées par Philips (section 1.3.1, page 17) et par Shazam (section 1.4.1, page 22) sont adaptées pour correspondre aux caractéristiques acoustiques utiles du signal de parole, générant des trames de 25 ms toutes les 10 ms (section 4.4.1, page 86). Pour cette même raison, dans l'adaptation de la méthode de Philips (AP), les sous-bandes fréquentielles sont calculées sur l'intervalle de 40 Hz à 3700 Hz (section 4.1.1, page 72). Afin de comparer les différents types de sous-empreinte sur des espaces de recherche similaires, une taille de sous-empreinte commune de 20 bits est choisie dans la méthode AP et dans l'adaptation de la méthode de Shazam (AS).

Dans la méthode AS, les paramètres de calcul des sous-empreintes sont définis pour retourner 100 sous-empreintes par seconde en moyenne. Ce nombre moyen est choisi en correspondance avec le nombre de trames calculées par seconde. Cette correspondance doit permettre une meilleure comparaison de robustesse de telles sous-empreintes avec les autres types de vecteurs lors des évaluations. Dans la méthode AS, les sous-empreintes sont définies par le système lorsque les trames du signal de parole contiennent des caractéristiques énergétiques particulières. Donc selon les caractéristiques acoustiques du signal de parole analysé, plusieurs sous-empreintes définies par la méthode AS peuvent être calculées pour une trame donnée. A l'opposé, certaines trames du signal de parole peuvent n'être représentées par aucune sous-empreinte issue de la méthode AS. Par ailleurs, les 20 bits composant une sous-empreinte sont répartis de la manière suivante [Ellis, 2009] :

- 8 bits pour la représentation de la première fréquence,
- 6 bits pour le différentiel entre les deux fréquences,
- 6 bits pour la différence temporelle entre les moments de ces deux fréquences.

Dans la méthode par calcul de paramètres MFCCs (MFCC), les vecteurs sont issus des paramètres MFCCs utilisés comme paramètres acoustiques initiaux dans l'expérience précédente pour calculer les sous-empreintes de la méthode par calcul de paramètres MFCCs suivi par une quantification vectorielle (QV) (section 3.3.2, page 50). La quantification vectorielle est restreinte à la production d'un total de 4096 centroïdes afin de prévenir tout sur-apprentissage. Les sous-empreintes dans la méthode QV sont donc définies sur 12 bits.

### 5.1.4 Distance entre vecteurs acoustiques

Dans les méthodes AP et AS, la mesure de distance choisie entre sous-empreintes est la distance de Hamming. Pourtant dans la méthode AS, des groupes de bits d'une sous-empreinte sont obtenus par quantification scalaire de valeurs réelles. Cependant, afin de conserver une mesure de distance simple et rapide, l'interdépendance des bits correspondant à la quantification scalaire d'une seule et même valeur n'est pas prise en compte.

Dans la méthode MFCC, la mesure de distance entre vecteurs MFCCs est la distance euclidienne. Dans la méthode QV, la sous-empreinte fait référence au centroïde le plus proche du vecteur MFCC de la trame. Dans ce cas, la distance de Hamming n'est pas applicable. Il est alors nécessaire de construire une relation d'équivalence entre les différents critères de similarité selon le type de sous-empreinte. Cette relation d'équivalence doit permettre de comparer une valeur de distance dans la méthode QV avec la distance de Hamming utilisée

dans les méthodes AP et AS. A cet effet, une distance appelée distance de centroïde pondéré est définie dans la méthode QV. Une telle mesure utilise alors la distance de Bhattacharyya pour fournir à chaque centroïde $C$ la liste ordonnée $L(C)$ de ses centroïdes plus proches voisins.

La distance de Bhattacharyya est définie pour mesurer la similarité entre deux distributions de densité de probabilité [Bhattacharyya, 1943]. Elle est fréquemment utilisée comme mesure de distance pour résoudre des problèmes de classification [Djouadi et al., 1990; Mak et al., 1996]. Donc en considérant deux distributions de densité de probabilité $C_1$ et $C_2$, de moyenne et variance respectivement $(\mu_1, \sigma_1^2)$ et $(\mu_2, \sigma_2^2)$, la distance de Bhattacharyya $B(C_1, C_2)$ entre $C_1$ et $C_2$ est définie comme :

$$B(C_1, C_2) = \frac{1}{8}(\mu_2 - \mu_1)^T \left(\frac{\sigma_2^2 + \sigma_1^2}{2}\right)^{-1} (\mu_2 - \mu_1) + \frac{1}{2}\ln\frac{(\sigma_1^2 + \sigma_2^2)/2}{\sqrt{\sigma_1^2.\sigma_2^2}} \qquad (5.3)$$

Une autre mesure de distance aurait pu être utilisée. En effet, issue de la divergence de Kullback-Leibler [Kullback et al., 1951], la distance $KL2$ permet de mesurer un rapport de proximité entre deux distributions de densité de probabilité [Siegler et al., 1997] (Annexe K, page 157). Cependant, cette distance $KL2$ n'a pas été retenue pour les raisons d'asymétrie de la divergence dont elle est issue.

Par ailleurs, le poids $P(C)$ de chaque centroïde $C$ est donné à la fin de la quantification vectorielle par le ratio du nombre de vecteurs MFCCs de $C$ par rapport au nombre total de vecteurs MFCCs de l'apprentissage. Considérant deux vecteurs MFCCs $x_1$ et $x_2$ ayant pour plus proche centroïde respectivement $C_1$ et $C_2$, la distance de centroïde pondéré $D(x_1, x_2)$ entre $x_1$ et $x_2$ est donnée par la somme des poids des centroïdes les séparant dans la liste $L(C_1)$. Cette distance de centroïde pondéré $D(x_1, x_2)$ est définie comme :

$$D(x_1, x_2) = \sum_{[C_j]} P(C_j) \qquad (5.4)$$

avec $[C_j]$ l'ensemble des centroïdes de la liste ordonnée $L(C_1)$ des plus proches voisins de $C_1$ jusqu'à la dernière valeur précédent $C_2$ dans cette liste.

## 5.2 Expériences

Différentes expériences d'identification acoustico-phonétique sont décrites afin d'évaluer la robustesse des divers types de sous-empreinte définis aux variabilités extrinsèque et intrinsèque. Une expérience préliminaire permet de définir une relation d'équivalence entre les différents espaces de recherche obtenus en référence à la mesure de distance associée au type de vecteur. Une seconde expérience est effectuée afin d'évaluer la robustesse des différents types de sous-empreinte sur la variabilité intrinsèque inter-locuteur. Dans la troisième expérience, la robustesse des différents types de sous-empreinte définis est évaluée sur la variabilité extrinsèque. Dans la quatrième expérience, la robustesse des différents types de sous-empreinte est évaluée sur les variabilités extrinsèque et intrinsèque inter-locuteur combinées.

Pour toutes les expériences, les vecteurs dont l'étiquette phonétique correspond au symbole de silence sont exclus de l'apprentissage et des différents tests. Dans les méthodes AP, QV et MFCC, les vecteurs sont calculés trame à trame. Ce n'est pas le cas dans la méthode AS où les sous-empreintes ne sont pas obligatoirement calculées à intervalles de temps réguliers. Donc pour chaque expérience, il est nécessaire de distinguer deux évaluations correspondant à deux groupes de vecteurs. Dans le premier groupe, les méthodes AP, QV et MFCC sont évaluées sur l'ensemble des trames du signal de parole de test. Dans le second groupe, seules les trames communes représentées par les vecteurs calculés sur l'ensemble des méthodes AP, QV, MFCC et AS sont prises en compte pour l'évaluation.

### 5.2.1 Espaces de recherche de taille équivalente

Lors de la recherche des plus proches voisins d'un vecteur, un seuil de validation est défini sur la mesure de distance associée à chacune des méthodes AP, AS, MFCC et QV. Ce seuil restreint l'espace de recherche de ces plus proches voisins à une portion limitée de l'étendue des valeurs possibles (Figure 5.3).

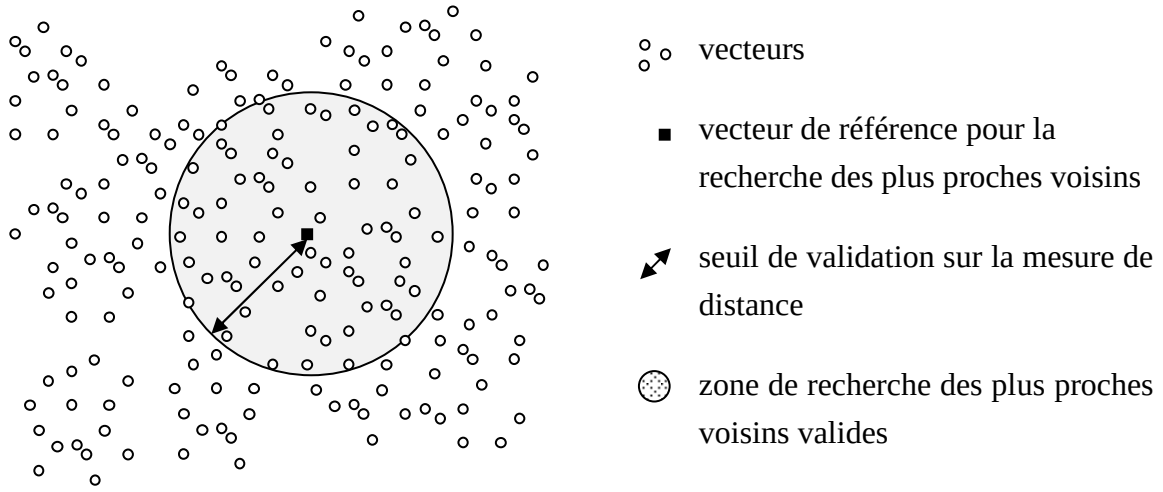



Figure 5.3 : Représentation de l'espace de recherche des plus proches voisins d'un vecteur

Dans les méthodes AP et AS, la valeur de la distance de Hamming entre deux sous-empreintes est limitée par la taille de la sous-empreinte. Ainsi, dans ces méthodes, pour une sous-empreinte définie sur 20 bits, les valeurs de distance de Hamming possibles entre deux sous-empreintes sont les nombres entiers compris entre 0, cas de sous-empreintes identiques, à 20, cas où aucun bit de même poids n'est similaire entre les deux sous-empreintes. Cependant, cette propriété de valeur de distance limitée n'est pas vérifiée dans les méthodes QV et MFCC. En effet, dans ces deux dernières méthodes, la distance utilisée est définie sur des espaces de valeurs réelles non limités.

Afin d'étudier ces différentes méthodes au sein d'évaluation communes, il est nécessaire de définir des espaces de recherche de taille sensiblement équivalente. A cet effet, une relation d'équivalence de la taille des espaces de recherche est définie entre :

- les méthodes AP et AS avec la distance de Hamming,
- la méthode QV avec la distance de centroïde pondéré,
- la méthode MFCC avec la distance euclidienne.

En tenant compte d'un seuil choisi sur la mesure de distance, il est possible de compter pour un vecteur de test le nombre de vecteurs d'apprentissage dont la distance au vecteur du test est inférieure au seuil choisi. A cet effet, le nombre moyen de vecteurs d'apprentissage répondant à cette contrainte est calculé. Ce nombre moyen correspond à la taille moyenne ER

de l'espace de recherche lors de l'étape d'identification. Dans cette expérience préliminaire, les ensembles de test en référence aux bases $A_a$ sont utilisés. Afin d'accélérer le processus d'obtention de cette taille moyenne, seuls 600 vecteurs issus de la méthode AP sont choisis aléatoirement pour chaque ensemble de test. Le tableau suivant présente les résultats de la taille moyenne ER de l'espace de recherche sur la méthode AP en fonction de différentes valeurs de seuil DH sur la distance de Hamming (Tableau 5.2).

| **base de données** / **ER** / **DH** | **MP3 TIMIT** | **N-MP3 TIMIT** | **NTIMIT** | **CTIMIT** |
|---|---|---|---|---|
| 0 | 5 | 5 | 5 | 5 |
| 1 | 60 | 100 | 70 | 35 |
| 2 | 860 | 860 | 730 | 420 |
| 3 | 5080 | 5200 | 4415 | 2690 |
| 4 | 21700 | 22100 | 19300 | 12000 |
| 5 | 73200 | 72700 | 68000 | 44000 |

DH : seuil sur la distance de Hamming
ER : taille moyenne de l'espace de recherche

Tableau 5.2 : Taille moyenne de l'espace de recherche selon le seuil sur la distance de Hamming (méthode AP)

La base d'apprentissage contient environ 1 200 000 sous-empreintes en excluant celles associées à une étiquette phonétique de silence. Par exemple, avec un seuil sur la distance de Hamming à 0 dans la méthode AP, 5 sous-empreintes identiques à la sous-empreinte de test sont atteintes en moyenne sur l'apprentissage. Lorsque le seuil sur la distance de Hamming est à 1 pour MP3 TIMIT, 60 sous-empreintes sont atteintes en moyenne. A partir des résultats donnés en Tableau 5.2 sur la méthode AP, nous choisissons de corréler le seuil sur la distance de Hamming avec la taille des espaces de recherche pour les méthodes QV et MFCC.

Dans la méthode QV, la taille de l'espace de recherche délimitée par une valeur maximale de distance de centroïde pondéré $D$ définie en équation (5.4). Afin d'obtenir un espace de recherche de taille équivalente entre les méthodes QV et AP, une sous-empreinte dans QV est atteinte chaque fois que la mesure $D$ est inférieure à une valeur de seuil équivalent à celui sur la distance de Hamming. Ce seuil équivalent est calculé par le ratio de la taille moyenne de l'espace de recherche ER par rapport au nombre total de sous-empreintes d'apprentissage. A cause du choix de la taille de sous-empreinte à 12 bits dans la méthode QV, un seuil sur une distance de Hamming à 0 n'a pas d'équivalent dans cette méthode.

Dans la méthode MFCC, il est nécessaire d'établir une correspondance entre la distance euclidienne et la distance de Hamming. Dans cette méthode, la taille moyenne de l'espace de recherche ER du Tableau 5.2 est alors utilisée. Cette taille moyenne permet de déterminer la distance euclidienne maximale entre un vecteur de test et les vecteurs de l'apprentissage atteints par l'espace de recherche. Le seuil DE sur la distance euclidienne est alors obtenu par la moyenne de ces distances maximales pour tous les vecteurs de test. Par exemple, pour MP3 TIMIT avec un seuil équivalent sur la distance de Hamming à 0 dans la méthode AP, il s'agit de mesurer dans la méthode MFCC la moyenne des distances euclidiennes maximales entre les vecteurs de test et leurs 5 plus proches voisins dans l'apprentissage. Afin d'accélérer le processus, l'ensemble des 600 vecteurs précédemment choisis aléatoirement est utilisé pour chaque ensemble de test. Ce seuil DE sur la distance euclidienne est ainsi calculé pour les différentes valeurs de seuil DH équivalent sur la distance de Hamming (Tableau 5.3).

| base de données / DE / DH | MP3 TIMIT | N-MP3 TIMIT | NTIMIT | CTIMIT |
|---|---|---|---|---|
| 0 | 29,9 | 25,3 | 27,6 | 30,2 |
| 1 | 33,3 | 29,2 | 31,7 | 31,9 |
| 2 | 38,2 | 32,6 | 35,1 | 35,0 |
| 3 | 42,2 | 36,2 | 38,4 | 37,9 |
| 4 | 46,3 | 39,8 | 41,8 | 40,8 |
| 5 | 50,6 | 43,7 | 45,7 | 44,0 |

DH : seuil sur la distance de Hamming
DE : moyenne des distances euclidiennes maximales

Tableau 5.3 : Moyenne des distances euclidiennes maximales selon le seuil équivalent sur la distance de Hamming (méthode MFCC)

Pour la méthode MFCC, durant l'identification, l'espace de recherche explore les vecteurs d'apprentissage dont la distance euclidienne avec le vecteur de test est inférieure à ce seuil DE. En tenant compte des distances de Hamming comme référence pour les méthodes AP et AS, des relations d'équivalence de la taille des espaces de recherche sont ainsi obtenues pour les méthodes QV et MFCC. Ces relations d'équivalence permettent alors de comparer toutes ces méthodes au sein d'évaluations communes.

### 5.2.2 Robustesse à la variabilité intrinsèque

Dans cette expérience, les différents types de sous-empreinte sont évalués sur leur robustesse à la variabilité intrinsèque inter-locuteur. La mesure de précision PM est utilisée pour l'analyse des résultats. Dans ce cas, les ensembles de TIMIT sont utilisés comme suit :

- l'ensemble de TIMIT en référence à la base $A_p$ comme base d'apprentissage,
- l'ensemble de TIMIT en référence à la base $T_p$ comme base de test.

Les seuils de distance pour la taille équivalente des espaces de recherche sont définis pour répondre aux deux paradigmes. Ces seuils sont déterminés à partir des ensembles de test en référence aux bases $A_a$. Hors la base TIMIT n'est pas définie parmi ces ensembles de test. Donc les seuils de distance sont repris de ceux obtenus avec la base de données MP3 TIMIT. L'expérience est effectuée dans un premier temps sur chacune des trames de la base de test avec les méthodes AP, QV et MFCC. Puis dans un second temps, l'expérience est reproduite sur les trames communes représentées par tous les types de vecteurs issus des méthodes AP, AS, QV et MFCC. Rappelons que dans la méthode QV, le seuil équivalent sur la distance de Hamming à 0 n'est pas défini. La figure suivante présente les résultats de précision PM tout en variant la valeur de seuil équivalent sur la distance de Hamming (Figure 5.4).

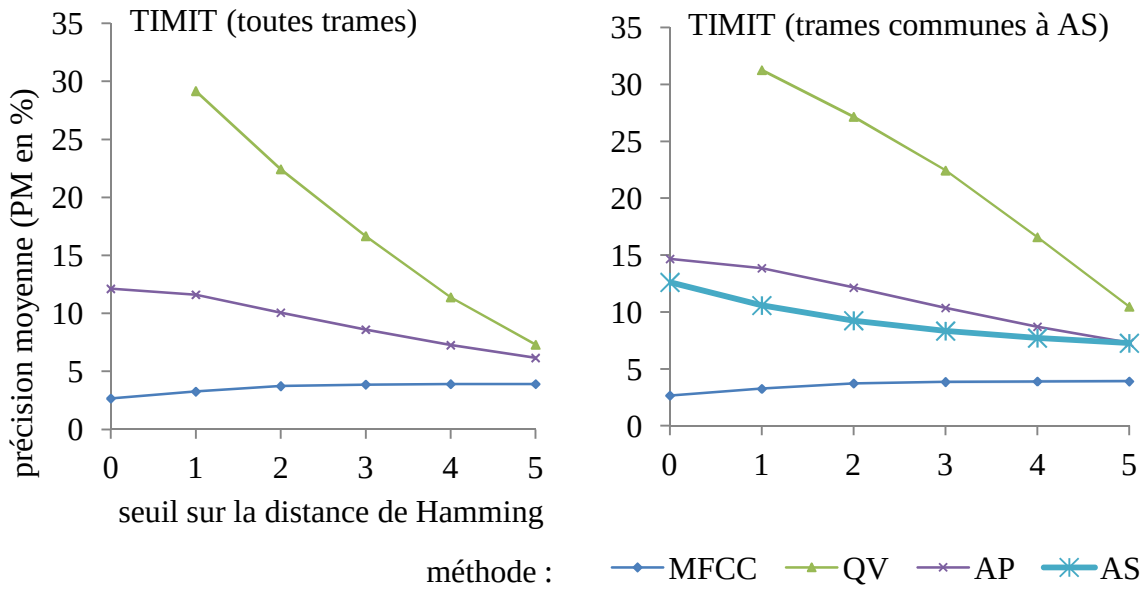



Figure 5.4 : Résultats de robustesse des vecteurs à la variabilité intrinsèque inter-locuteur

Pour cette évaluation, la méthode QV définie à partir de paramètres acoustiques MFCCs quantifiés est bien plus robuste que les méthodes AP et AS issues des techniques de l'identification audio par empreinte. Par rapport à l'usage direct de paramètres MFCCs, la quantification vectorielle apportée par la méthode QV permet de ne pas prendre en compte les vecteurs situés dans des zones de l'espace de recherche faiblement représentées. De surcroît, le lissage par quantification des vecteurs MFCCs assure dans la méthode QV un niveau de détail moins élevé de la représentation acoustique du signal de parole. Grâce à cette perte de résolution, les sous-empreintes issues de la méthode QV sont représentatives des caractéristiques les plus discriminantes des paramètres acoustiques MFCCs.

Par ailleurs, la robustesse des sous-empreintes des méthodes QV, AP et AS se dégrade lorsque la contrainte sur le seuil de distance est relâchée. Dans ce cas, l'espace de recherche déterminé contient alors une proportion moindre de sous-empreintes avec une étiquette phonétique correcte. Dans la méthode MFCC, les résultats sont proches du hasard ($1/39 \approx 2{,}6$ %). Leur évolution en fonction du seuil sur la distance n'est alors pas pertinente. L'expérience suivante étudie la robustesse des sous-empreintes de toutes ces méthodes à la variabilité extrinsèque.

### 5.2.3 Robustesse à la variabilité extrinsèque

Dans cette expérience, les différents types de vecteurs sont évalués sur leur robustesse à la variabilité extrinsèque. La mesure de précision extrinsèque PE est alors utilisée pour l'analyse des résultats. Les ensembles de bases de données sont utilisés comme suit :

- l'ensemble de TIMIT en référence à la base $A_p$ comme base d'apprentissage,
- les ensembles des autres bases en référence aux bases $A_a$ comme bases de test.

Dans un premier temps, les méthodes AP, QV et MFCC sont comparées entre elles sur chacune des trames de chaque base de test. La figure suivante présente les résultats de précision PE tout en variant la valeur de seuil équivalent sur la distance de Hamming (Figure 5.5).

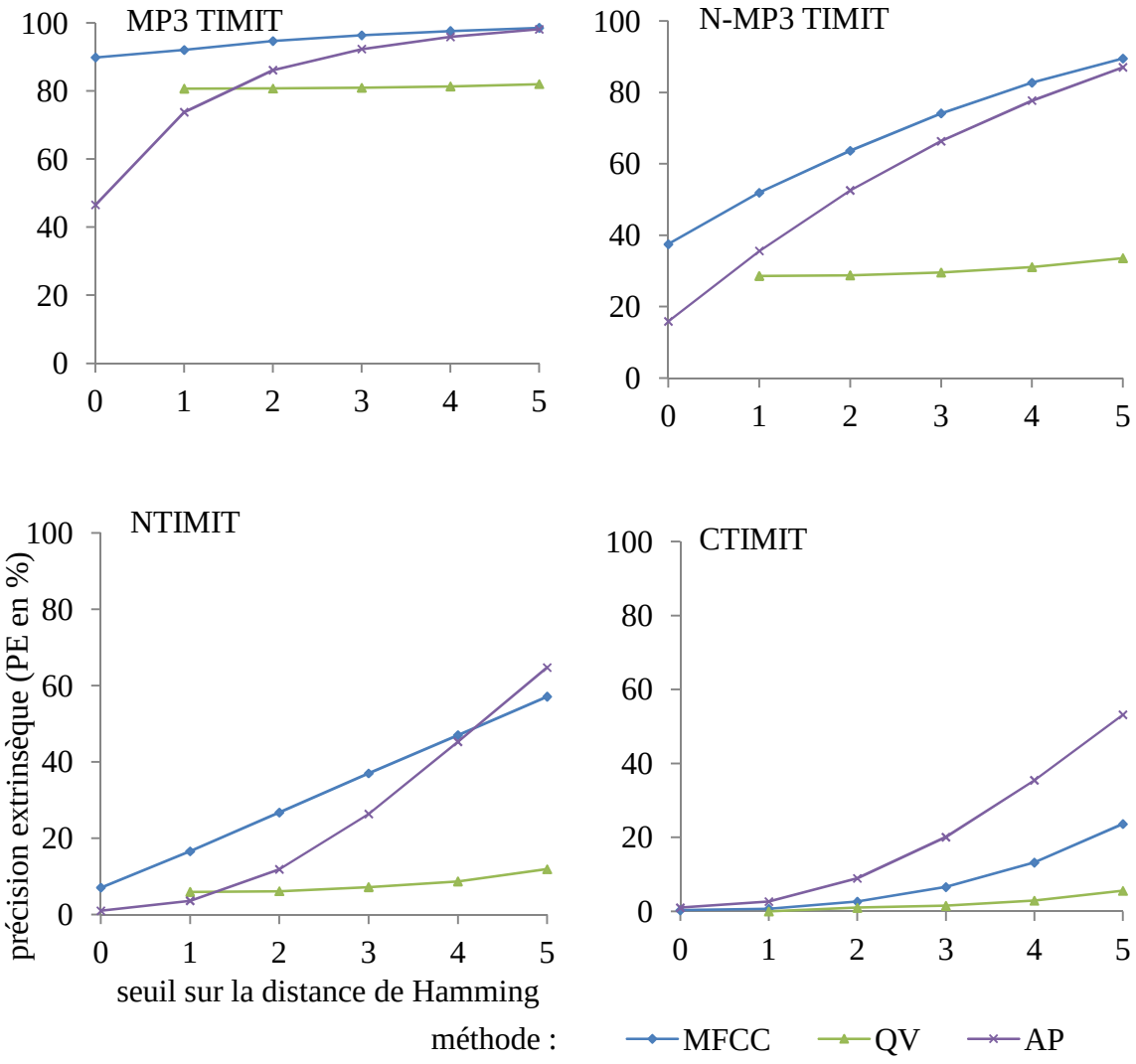



Figure 5.5 : Résultats de robustesse des vecteurs à la variabilité extrinsèque (toutes trames)

Les meilleurs résultats de précision extrinsèque PE sont obtenus par la méthode MFCC dans les évaluations sur MP3 TIMIT, N-MP3 TIMIT et NTIMIT, excepté pour une grande valeur de seuil sur NTIMIT. Dans le cas d'une transformation du signal de parole par des facteurs de variabilité extrinsèque de type convolutif ou bruit blanc, la représentation acoustique du signal de parole sous la forme de vecteurs MFCCs est donc la plus robuste de toutes les méthodes évaluées. Cependant, ce constat n'est pas vérifié dans l'évaluation sur CTIMIT. Dans ce cas, la méthode AP est le mode de représentation acoustique du signal de parole retournant les sous-empreintes les plus robustes. Ce dernier résultat est relativement inattendu pour cette expérience et n'est observable que sur le signal de parole altéré issu de la base de données CTIMIT. CTIMIT diffère des autres bases de données par la présence de bruits additionnels non connus issus du monde réel, en plus d'un filtre sur le canal de transmission.

Dans la méthode QV, la précision PE n'augmente pas autant que dans les autres méthodes suivant le relâchement de la contrainte sur le seuil de distance. Dans cette méthode QV, lorsque la valeur d'une sous-empreinte de test est modifiée par rapport à celle de la sous-empreinte initiale issue de l'apprentissage, leur centroïde respectif correspondant ne se situe pas dans un voisinage proche. Les sous-empreintes de la méthode AP sont plus robustes à la variabilité extrinsèque que celles de la méthode QV à partir d'un seuil équivalent sur la distance de Hamming à 2. Par exemple, la précision PE dans la méthode AP augmente avec l'évolution du seuil sur la distance de Hamming jusqu'à atteindre 65 % sur NTIMIT et 55 % sur CTIMIT pour un seuil à 5. Ces résultats hétérogènes suggèrent que la différence de topologie des sous-empreintes entre les méthodes AP et QV produisent un partitionnement de l'espace de recherche très différent.

La même expérience est reproduite en tenant compte uniquement des trames communes représentées par tous les types de vecteurs issus des méthodes AP, AS, QV et MFCC (Figure 5.6).

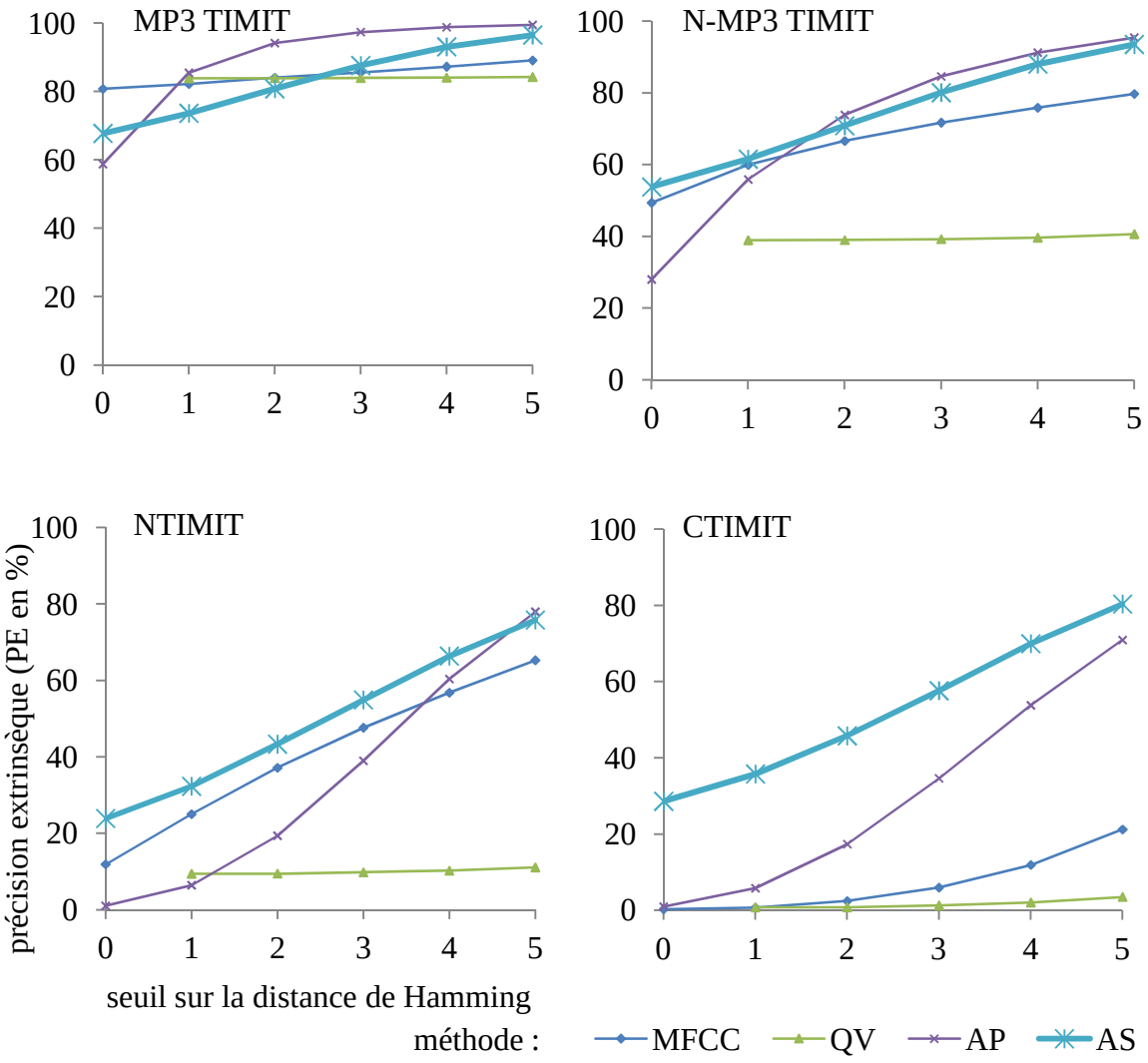



Figure 5.6 : Résultats de robustesse des vecteurs à la variabilité extrinsèque (trames communes à la méthode AS)

En ne tenant compte que des trames communes générant des vecteurs sur l'ensemble des méthodes définies, les résultats sont plus contrastés. La méthode AS retourne les sous-empreintes les plus robustes à la variabilité extrinsèque dans les tests sur NTIMIT et CTIMIT. Concernant les évaluations sur N-MP3 TIMIT (ajout d'un bruit blanc) et CTIMIT (environnement bruité), les méthodes AP et AS basées sur les techniques d'identification audio par empreinte s'avèrent les plus robustes à la variabilité extrinsèque à partir d'un seuil sur la distance de Hamming à 2. Ce comportement suggère que ces méthodes de calcul de sous-empreinte sont plus robustes aux bruits additifs pour ce type d'évaluation. En effet, les dégradations appliquées sur le signal de parole pour les bases de données MP3 TIMIT (compression avec perte) et NTIMIT (filtre sur le canal de transmission) sont essentiellement liées à un bruit convolutif [Kamper et al., 2009]. Le résultat le plus significatif est la bonne robustesse à la variabilité extrinsèque des sous-empreintes issues de la méthode AS sur CTIMIT. L'expérience suivante étudie la robustesse des sous-empreintes de toutes ces méthodes aux variabilités extrinsèque et intrinsèque.

### 5.2.4 Robustesse aux variabilités extrinsèque et intrinsèque

Dans cette expérience, les différents types de sous-empreinte sont évalués sur leur robustesse aux variabilités extrinsèque et intrinsèque inter-locuteur combinées. La mesure de précision moyenne PM est alors utilisée pour l'analyse des résultats. Les ensembles de bases de données sont utilisés comme suit :

- l'ensemble de TIMIT en référence à la base $A_p$ comme base d'apprentissage,
- les ensembles des autres bases en référence aux bases $T_a$ comme bases de test.

Les méthodes AP, QV et MFCC sont tout d'abord comparées entre elles sur chacune des trames de chaque base de test. La figure suivante présente les résultats de précision PM tout en variant la valeur de seuil équivalent sur la distance de Hamming (Figure 5.7).

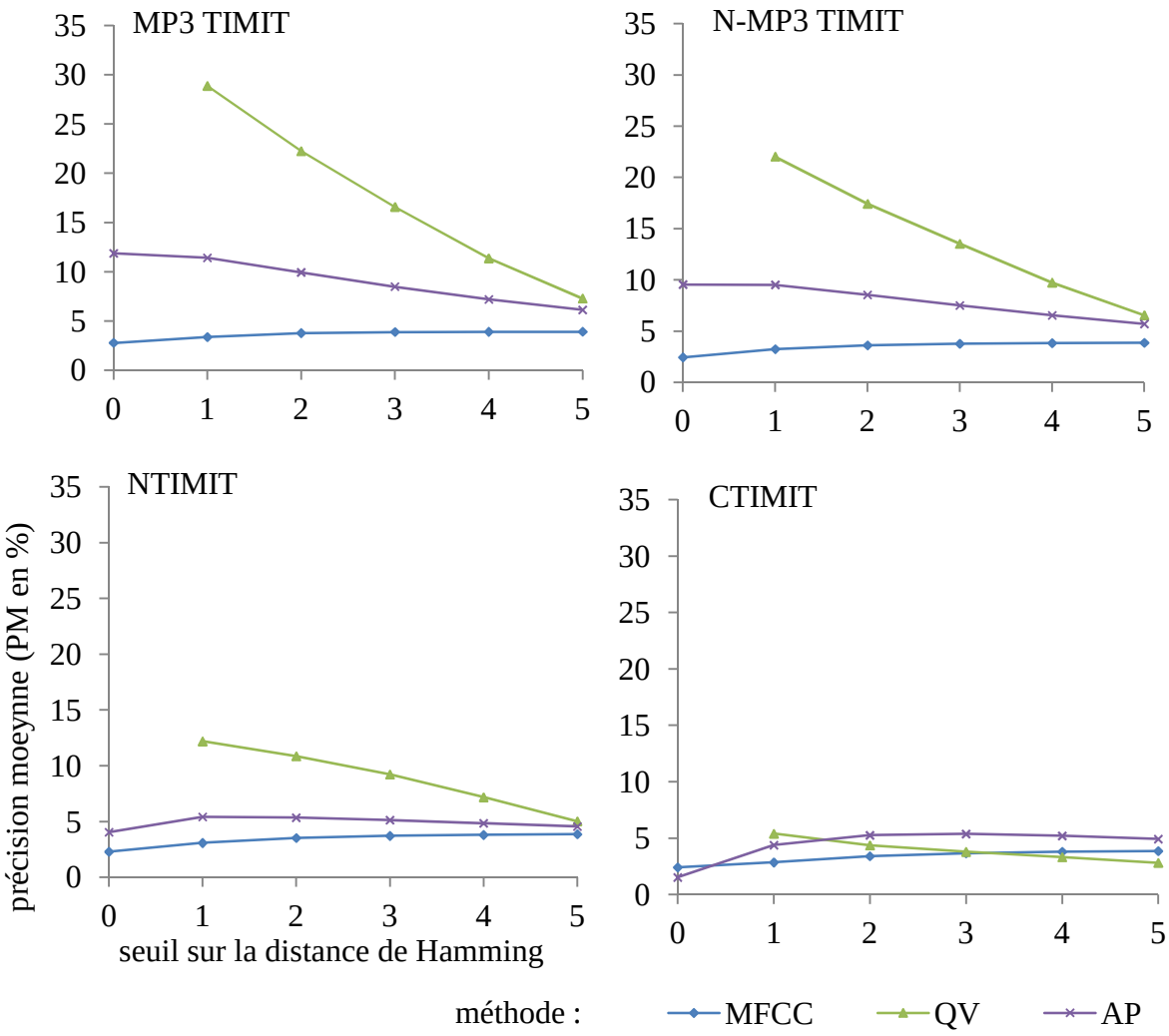



Figure 5.7 : Résultats de robustesse des vecteurs aux variabilités extrinsèque et intrinsèque (toutes trames)

Pour cette évaluation de la précision moyenne PM, les vecteurs obtenus par la méthode MFCC sont très peu robustes, quelle que soit la base de données de test. Une raison possible à cette faible performance par rapport à l'évaluation précédente de robustesse à la variabilité extrinsèque seule en section 5.2.3 est le manque de discrimination de vecteurs MFCCs qui s'avèrent proches en distance euclidienne mais avec une étiquette phonétique différente.

Hormis l'évaluation sur CTIMIT, les sous-empreintes issues de la méthode QV sont plus robustes aux variabilités extrinsèque et intrinsèque inter-locuteur combinées que celles issues des méthodes AP et AS. Pour l'évaluation sur CTIMIT, tous les types de sous-empreinte ont une robustesse très faible à de telles variabilités. Cependant, sur CTIMIT, les sous-empreintes issues de la méthode AP s'avèrent sensiblement plus robustes que celles issues de la méthode QV à partir d'un seuil équivalent sur la distance de Hamming à 2.

Afin de tenir compte de la contrainte de calcul de la sous-empreinte selon la méthode AS, la même expérience est reproduite en tenant compte uniquement des trames communes représentées par tous les types de vecteurs issus des méthodes AP, AS, QV et MFCC (Figure 5.8).

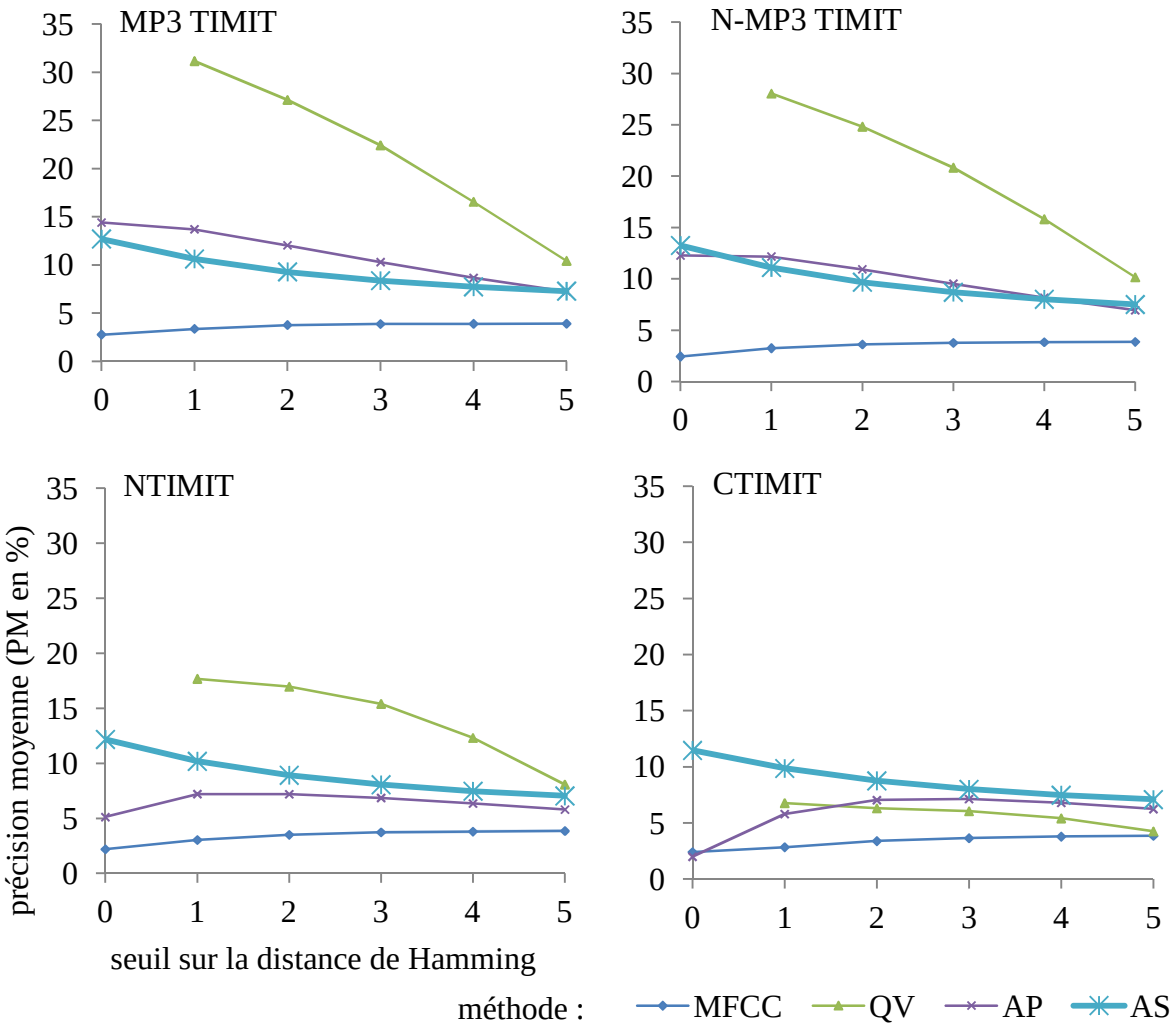



Figure 5.8 : Résultats de robustesse des vecteurs aux variabilités extrinsèque et intrinsèque (trames communes à la méthode AS)

Les commentaires formulés précédemment lors de la mesure de la précision moyenne PM sur toutes les trames (Figure 5.7, page 112) sont de nouveau vérifiés. Hormis l'évaluation sur CTIMIT, les sous-empreintes issues de la méthode QV sont plus robustes aux variabilités extrinsèque et intrinsèque inter-locuteur combinées que celles issues des méthodes AP et AS. En effet, la robustesse des sous-empreintes issues de cette méthode AS est médiocre dans les tests sur MP3 TIMIT, N-MP3 TIMIT et NTIMIT en comparaison à la robustesse des sous-empreintes issues de la méthode QV. Cependant, les sous-empreintes issues de la méthode AS

montrent une robustesse sensiblement équivalente, quel que soit l'ensemble de test utilisé. Sur CTIMIT, la robustesse des sous-empreintes issues de la méthode QV est fortement dégradée. Dans ce cas, les sous-empreintes issues de la méthode AS sont les plus robustes. Les différents ensembles de test ont tous pour origine le signal de parole de la base de données TIMIT ayant subi des altérations. Les variations entre ces différents ensembles de test sont dues à la présence de déformations du signal de parole liées à la variabilité extrinsèque. Les résultats stables sur la méthode AS quel que soit l'ensemble de test montrent la robustesse des sous-empreintes issues de cette méthode à la variabilité extrinsèque.

Par ailleurs, les résultats obtenus sur MP3 TIMIT (MP3 TIMIT, Figure 5.7 et Figure 5.8) sont très proches de ceux obtenus lors de l'évaluation à la variabilité intrinsèque inter-locuteur sur la version originale de la base de données TIMIT (section 5.2.2, page 106). Cette observation confirme la conservation des caractéristiques acoustiques utiles du signal de parole lors de la compression avec perte de ce signal audio au format MP3.

## 5.3 Discussions

Dans la présente étude, divers types de sous-empreinte sont définis à partir de méthodes issues de l'identification audio par empreinte tout comme à partir de méthodes issues de l'étude des paramètres MFCCs. Ces différents types de sous-empreinte ont été évalués sur leur robustesse aux variabilités extrinsèque et intrinsèque.

Les résultats de ces évaluations sont fortement contrastés en fonction du type de variabilité étudié et de la méthode de représentation acoustique choisie. Dans un cas, les sous-empreintes adaptées de l'identification audio par empreinte (méthodes AP et AS) sont plus robustes à la variabilité extrinsèque. Dans l'autre cas, l'exploitation des paramètres MFCCs suivie d'une quantification vectorielle (méthode QV) permet le calcul de sous-empreintes plus robustes à la prise en compte de variabilités extrinsèque et intrinsèque interlocuteurs combinées. Par ailleurs, l'évolution des précisions extrinsèque PE et moyenne PM en fonction du seuil admis sur la mesure de distance varie fortement selon le type de sous-empreinte. Ce résultat suggère que le partitionnement de l'espace de recherche diffère fortement selon la méthode de calcul de sous-empreinte issue de l'identification audio ou de MFCCs quantifiés. Concernant l'expérience sur la robustesse à la variabilité extrinsèque, le résultat le plus remarquable celui sur les sous-empreintes adaptées de la méthode de Shazam. Ces sous-empreintes sont en effet robustes à la variabilité extrinsèque lorsque le signal de parole est perturbé par l'environnement et les conditions de transmission du signal acoustique.

Ces résultats montrent un intérêt particulier dans la recherche de propriétés complémentaires de robustesse sur ces différentes représentations acoustiques du signal de parole sous forme de sous-empreinte. Ainsi, en perspective d'une possible poursuite de cette étude, une forme de combinaison pourrait se traduire par une composition de plusieurs types de vecteurs. Dans un système de RAP utilisant une telle combinaison, les résultats partiels d'une identification par un vecteur acoustique à base de paramètres MFCCs pourraient retourner les plus proches voisins du vecteur dans la base d'apprentissage. Puis, une identification acoustico-phonétique pourrait être appliquée sur ces résultats partiels en utilisant des sous-empreintes adaptées des méthodes d'identification audio par empreinte.

Une autre approche possible serait la prise en compte de ces nouveaux paramètres acoustiques dans un système de référence. Ces paramètres acoustiques issus de l'identification audio par empreinte seraient alors évalués dans le cadre d'une tâche spécifique à la RAP.

# Conclusion de la seconde partie

La recherche de paramètres acoustiques robustes pour la représentation du signal de parole tient compte de différents types de variabilité. D'un coté, la gestion de la variabilité extrinsèque du signal de parole est un sujet commun avec celui de la gestion du bruit dans le domaine du traitement du signal audio [Peeters et al., 2009] (section 4.2, page 77). D'autre part, une variabilité intrinsèque spécifique au signal de parole est définie par la prise en compte des multiples aspects de sa production [Benzeghiba et al., 2007] (section 4.1, page 72). Dans cette étude, nous avons évalué la robustesse de différents types de sous-empreinte à ces variabilités extrinsèque et intrinsèque. A cet effet, nous avons développé un système d'identification acoustico-phonétique au niveau local. Les types de sous-empreinte évalués sont basés sur :

- la méthode de type Philips [Haitsma et al., 2001] (section 1.3, page 17),
- la méthode de type Shazam [Wang, 2003] (section 1.4, page 22),
- une méthode exploitant des paramètres MFCCs quantifiés (section 3.3.2, page 50).

Les résultats de ces évaluations sont contrastés en fonction du type de sous-empreinte et de la variabilité considérée. Dans un cas, les sous-empreintes issues de l'identification audio par empreinte sont plus robustes à la variabilité extrinsèque que celles issues de paramètres MFCCs quantifiés (section 5.2.3, page 107). Dans l'autre cas, les paramètres MFCCs quantifiés sont des sous-empreintes plus robustes aux variabilités extrinsèque et intrinsèque inter-locuteur combinées (section 5.2.4, page 111). Il apparaît que le partitionnement de l'espace de recherche diffère fortement en fonction du type de sous-empreinte choisi et de son critère de similarité associé. Toutefois, en présence de perturbations liées à l'environnement et aux conditions de transmission du signal de parole, les sous-empreintes issues de la méthode de type Shazam sont plus robustes à la variabilité extrinsèque que celles issues de paramètres MFCCs quantifiés (cas de CTIMIT dans la section 5.2.4, page 111).

Les résultats de ces diverses expériences mettent en évidence des propriétés complémentaires de robustesse de ces sous-empreintes. La combinaison de différentes représentations acoustiques du signal de parole peut se traduire par une composition de plusieurs types de vecteurs au sein d'un même système de RAP.

# Troisième partie : Perspectives et conclusion générale

Dans les deux premières parties de cette étude, nous avons développé de nouveaux types de représentation acoustique du signal de parole en se basant sur les caractéristiques de méthodes d'identification audio par empreinte.

La première partie de cette étude a montré la possibilité d'adapter la méthode d'identification audio par empreinte pour développer un système dédié à certaines tâches particulières de la RAP, comme l'identification d'enregistrement bruité, la reconnaissance de mots isolés ou le décodage acoustico-phonétique (DAP). A cet effet, différents types de sous-empreinte ont été définis, adaptés autant de la méthode d'identification audio développée par Philips (section 1.3, page 17) que de paramètres MFCCs (section 4.3.2, page 84). Ces derniers paramètres ont permis de définir de nouveaux types de sous-empreinte grâce à l'usage d'une quantification vectorielle en mode supervisé ou non-supervisé (section 3.3, page 49). Ces nouveaux types de sous-empreinte améliorent l'utilisation d'un système d'identification audio par empreinte pour répondre à la tâche de DAP. Cependant, la performance d'un tel système reste en retrait par rapport à l'usage d'un système de référence à base de modèles de Markov cachés.

La seconde partie de cette étude a montré un intérêt particulier pour des sous-empreintes adaptées de la méthode d'identification audio développée par Shazam (section 1.4, page 22). La robustesse de ce type de sous-empreinte a été évaluée dans diverses conditions de dégradation du signal de parole en tenant compte des variabilités extrinsèque et intrinsèque inter-locuteur. A cet effet, une application d'identification acoustico-phonétique a été développée pour une évaluation à l'échelle de la trame. Dans certaines conditions, les sous-empreintes issues de l'adaptation de la méthode de Shazam ont alors montré une meilleure robustesse à ces variabilités par rapport aux sous-empreintes des autres méthodes définies.

Ces conditions de reconnaissance correspondent à un signal de parole de test dégradé par rapport à celui de l'apprentissage (apprentissage sur une base de données de signal de parole propre, test sur une base de données de signal de parole bruité en conditions réelles).

Cette étude pose de nouvelles problématiques pour de futurs travaux de recherche. D'une part, un système d'identification audio par empreinte a été développé pour répondre à la tâche de DAP (Chapitre 3, page 43). Par ailleurs, des sous-empreintes adaptées de la méthode de Shazam s'avèrent robustes dans des conditions difficiles de reconnaissance de la parole (Chapitre 5, page 93). Il serait alors intéressant d'évaluer la performance du système développé pour le DAP en utilisant ce type de sous-empreinte dans des conditions difficiles de reconnaissance.

Cependant, les systèmes d'identification audio par empreintes sont reconnus pour être robustes dans la détection d'évènements ponctuels dans un continuum sonore. Une tâche usuelle d'un tel système est la détection d'un extrait de musique dans un flux audio. Une tâche similaire appliquée au domaine de la RAP est alors la détection de mots-clés. Compte-tenu des similarités de leur application et des résultats de robustesse obtenus par la présente adaptation de la méthode de type Shazam, la prochaine étape de ce travail de recherche est l'évaluation de la robustesse de telles sous-empreintes au sein d'un système de détection de mots-clés. La performance d'un tel système serait alors comparée à celle obtenue par un système de référence.

Au cours d'une expérience préliminaire, nous désirons évaluer la robustesse de paramètres MFCCs au sein d'un système de détection de mots-clés de référence. Dans un premier temps, les différentes applications d'un système de détection de mots-clés sont décrites. Puis dans un second temps, le système de détection de mots-clés de référence [Rose et al., 1990] est présenté. Dans un troisième temps, une expérience préliminaire de détection de mots-clés est effectuée sur un signal de parole de test présentant des variabilités intrinsèque (section 4.1, page 72) et extrinsèque (section 4.2, page 77) par rapport au signal de parole d'apprentissage. La robustesse des paramètres acoustiques MFCCs utilisés au sein de ce système de détection de mots-clés de référence est alors évaluée en fonction du type de variabilité. Enfin, plusieurs propositions sont évoquées pour une adaptation du système d'identification audio initialement développé pour le DAP (Chapitre 3, page 43) vers une tâche de détection de mots-clés.

# Chapitre 6. Expérience préliminaire pour la détection de mots-clés

Etudiée depuis les années 1970 [Christiansen et al., 1976], la détection de mots-clés dans la parole continue consiste à reconnaître et à localiser toutes les occurrences des mots d'une liste de mots-clés dans un continuum de parole donné [Medress et al., 1978]. L'objectif principal d'un système de détection de mots-clés est la discrimination des segments du signal de parole contenant le mot-clé recherché lors de l'analyse d'un continuum de parole [Medress et al., 1979]. Ainsi, les segments contenant les mots-clés sont extraits sans une reconnaissance détaillée de tous les mots prononcés [Wilpon et al., 1989]. En effet, dans le cas d'un système de détection de mots-clés, seul un nombre limité de mots-clés utiles est détecté à partir du signal de parole analysé [Hofstetter et al., 1992]. Certaines applications interactives peuvent alors ne nécessiter que la prononciation d'un mot-clé donné pour déclencher une réponse appropriée [Nakamura et al., 1993]. La détection du seul mot-clé est dans ce cas suffisante, sans nécessiter la reconnaissance complète du signal de parole [Zue et al., 1997]. Le système peut donc fournir une réponse immédiate suite à cette détection du mot-clé [Gorin et al., 1997]. L'utilisation d'un système de détection de mots-clés permet donc de réduire la complexité et les défauts associés à un système de RAP en parole continue [Zue et al., 1997; Tabibian et al., 2011]. De surcroît, les systèmes de détection de mots-clés ne rencontrent pas les mêmes difficultés que la RAP en parole continue pour la discrimination des hésitations, des répétitions liées aux faux départs, des pauses ou des phrases grammaticalement incorrectes [Smídl et al., 2006]. Des travaux existants décrivent de manière exhaustive l'évolution des systèmes de détection de mots-clés [Gelin, 1997; Ben Ayed, 2003; Leblouch, 2009]. Une description des principales applications d'un système de détection de mot-clé est disponible (Annexe L, page 159).

Dans le contexte de la détection de mots-clés, différentes stratégies basées sur des modèles de Markov cachés (*Hidden Markov Model*, HMM) continus sont proposées (Annexe F, page 145). Dans la plupart des cas, un système construit autour d'un ensemble de HMMs à base d'unités linguistiques est entraîné par une importante base de données d'apprentissage constituée du signal de parole associé aux transcriptions correspondantes [Young et al., 2006]. Lors du développement des systèmes de détection de mots-clés, les efforts se sont portés vers la recherche de la minimisation des fausses alarmes, c'est-à-dire minimiser le nombre de mots-clés détectés à tord. Ces efforts ont contribué à définir des méthodes de modélisation des séquences de signal de parole en dehors des segments formés par la prononciation des mots-clés [Chigier, 1992]. Ainsi, la modélisation des mots hors vocabulaire permet de décrire le signal de parole en dehors des mots-clés. L'étendue de ces mots hors vocabulaire est représentée sous la forme de modèles de rejet définis au sein du système de détection de mots-clés. Ces modèles de rejet, dits modèles poubelles (*garbage model*), peuvent compléter le système et être utilisés comme modèles discriminants pour la détection des mots-clés (Annexe M, page 163). En fonction du type d'application défini, la performance d'un système de détection de mots-clés peut être exprimée par différentes mesures d'évaluation (Annexe N, page 167).

Un système de détection de mots-clés de référence [Rose et al., 1990] est présenté. Ce système de référence doit être en mesure de détecter et de situer des mots-clés prononcés dans un continuum de parole. Une expérience préliminaire est alors effectuée afin d'évaluer la robustesse d'un tel système de référence sur le signal de parole de test ayant subi diverses variations par rapport à celui de l'apprentissage. Ces variations du signal de parole sont issues autant de la variabilité intrinsèque (section 4.1, page 72) que de la variabilité extrinsèque (section 4.2, page 77). Enfin, certaines propositions sont indiquées pour adapter le système d'identification audio initialement développé pour le DAP (Chapitre 3, page 43) vers une tâche de détection de mots-clés.

## 6.1 Présentation du système de référence

Le système de détection de mots-clés de référence [Rose et al., 1990] présenté utilise les propriétés des HMMs (Annexe F, page 145) et des modèles poubelles (Annexe M, page 163). En particulier, l'architecture de ce système est basée sur la mise en concurrence des résultats partiels de HMMs de mots-clés avec modèle poubelle et de HMMs phonétiques. Etant donné un mot-clé à détecter et un signal de parole de test, le système de détection de mots-clés prédit le meilleur intervalle temporel d'apparition de ce mot-clé dans le test. Cette prédiction est associée à une mesure de confiance. Si cette mesure de confiance est supérieure à un certain

seuil fixé, alors le mot-clé est déclaré prononcé dans l'intervalle temporel. Si cette mesure de confiance est inférieure au seuil, alors le mot-clé est considéré non prononcé.

Compte-tenu de l'architecture du système présenté, la construction des modèles s'effectue en plusieurs étapes. Tout d'abord, un ensemble de HMMs phonétiques est appris sur l'ensemble des données d'apprentissage. Puis, un nouveau HMM est alors construit pour chacun des mots-clés définis à partir de ces HMMs phonétiques (Annexe F, page 145). Compte-tenu de la faible représentation de la prononciation des mots-clés dans l'ensemble de la base d'apprentissage, on considère que ces prononciations particulières ont peu d'influence sur la construction des HMMs phonétiques. Le modèle poubelle utilisé pour ce système est alors représenté par l'ensemble des HMMs phonétiques, complètement connectés.

Etant donné une telle modélisation par mot-clé intégrant un modèle poubelle, la détection des mots-clés est effectuée en recherchant la séquence des états des modèles maximisant la vraisemblance du signal de parole de test. Une pénalité sur les valeurs de vraisemblance sous la forme d'un coefficient pondérateur est affectée au modèle poubelle en faveur des modèles de mots-clés. La valeur optimale de cette pénalité est déterminée de manière empirique afin de maximiser le compromis entre le nombre de mots-clés correctement détectés et celui de fausses alarmes. La détection de mots-clés est alors déterminée en vérifiant si le meilleur chemin fourni par l'algorithme de Viterbi retourne les modèles de mots-clés considérés ou non.

Pour chacun des mots-clés détectés, un résultat intermédiaire est obtenu par la valeur de vraisemblance retournée par l'algorithme de Viterbi, normalisée par la durée du mot-clé. Considérant un mot-clé $W$ détecté sur l'intervalle temporel $(t_1, t_n)$ avec pour état final $q_f$, ce résultat intermédiaire est donné sous la forme d'un score $S_W$ tel que :

$$S_W = \frac{\log P(q_f | x_{t_n} \dots x_{t_1})}{t_n - t_1} \tag{6.1}$$

avec $x_t$ le vecteur d'observation à l'instant $t$ et $P(q_f | x_{t_n} \dots x_{t_1})$ la probabilité d'obtenir l'état $q_f$ à partir de la séquence de vecteurs d'observation $(x_{t_n} \dots x_{t_1})$.

Ce résultat intermédiaire $S_W$ est fortement dépendant de la variabilité temporelle entre les différentes prononciations d'un mot (section 4.1.3, page 74). Cette dépendance réduit la fiabilité de la détection sur un segment donné et rend difficile la distinction entre bonne détection et fausse alarme [Rose et al., 1990]. Il est donc nécessaire d'effectuer un traitement supplémentaire sur ce résultat intermédiaire afin d'améliorer la fiabilité de la détection. Pour chacun des résultats intermédiaires, un décodage acoustico-phonétique (DAP) exploitant les

mêmes HMMs phonétiques est alors effectué sur l'intervalle temporel de détection du mot-clé. Un nouveau score $S_{DAP}$ est alors calculé à partir de ce DAP, similairement au calcul du score sur le mot-clé $S_W$ obtenu selon l'équation (6.1). En résultat final, une mesure de confiance est alors retournée par la différence des deux scores précédents sous la forme du score résultant $S_R$ tel que :

$$S_R = S_W - S_{DAP} \tag{6.2}$$

Un seuil sur cette mesure de confiance est ajouté afin de valider ou non la détection du mot-clé. La valeur de ce seuil, tout comme la valeur de pénalité sur le modèle poubelle, est déterminée durant une étape de développement afin de maximiser le compromis entre bonnes détections et fausses alarmes.

## 6.2 Expérience

L'application présentée au sein de cette étude est dédiée à la tâche de détection de mots-clés classique. Cette application répond alors à une requête de type *Keyword Spotting* (KWS) (Annexe L, page 159). Le système est développé à l'aide de l'ensemble de programme HTK [Young et al., 2006]. L'expérience préliminaire consiste en l'évaluation de la robustesse de ce système de détection de mots-clés de référence à différents types de variabilité. En référence aux bases de données utilisées précédemment pour l'évaluation de robustesse des sous-empreintes aux variabilités extrinsèque et intrinsèque (section 5.2, page 101), cette expérience utilise alors les bases de données TIMIT, NTIMIT et CTIMIT ainsi que l'ensemble phonétique de 39 phonèmes.

Le sous-ensemble de la base d'apprentissage prédéfini de TIMIT (200 mn, 462 locuteurs) est utilisé pour l'apprentissage des HMMs phonétiques. Afin de tenir compte de la taille limitée de la base CTIMIT, les ensembles de données d'évaluation sont restreints aux parties communes du signal de parole dans ces différentes bases de données. Ainsi, deux ensembles disjoints issus du sous-ensemble prédéfini du test de chacune des bases de données TIMIT, NTIMIT et CTIMIT sont utilisés pour i) le développement (14 mn, 56 locuteurs) et ii) le test (26 mn, 112 locuteurs). Les sous-ensembles d'apprentissage, de développement et de test contiennent un signal de parole prononcé par des différents locuteurs. Le sous-ensemble de développement permet de déterminer les valeurs optimales de pénalité entre modèle de mot-clé et modèle poubelle ainsi que de seuil sur la mesure de confiance. Le sous-ensemble de test permet de valider les résultats de cette expérience préliminaire.

L'ensemble des mots-clés est formé de 80 mots choisis aléatoirement parmi les mots de 4 phonèmes ou plus disponibles dans TIMIT [Grangier et al., 2009] (Annexe O, page 171). Ce choix aléatoire est motivé par l'évaluation des performances du système sur la détection de mots-clés quelque soit le mot-clé sélectionné, présent ou non dans les sous-ensembles de développement et de test. Les séquences phonétiques relatives aux mots-clés sont obtenues par la correspondance temporelle entre la réalisation acoustique du mot-clé lors de son apparition dans le signal de parole et les phonèmes issus de la transcription phonétique. Par ailleurs, une correction manuelle de la segmentation temporelle des mots-clés est effectuée afin de porter les effets de coarticulation (absorption, insertion) en dehors des mots-clés. Par exemple, lors de certaines prononciations de la séquence de mots « *controlled the* », le dernier phonème du mot « *controlled »* est parfois absorbé par le premier phonème du mot suivant.

Les HMMs phonétiques sont construits dans les mêmes conditions que lors de la mise en place du système de référence pour l'évaluation précédente sur le DAP (section 3.5.2, page 59). Chaque HMM phonétique est composé de 3 états contenant un modèle de mélange de 128 gaussiennes par état.

Les résultats de cette évaluation sont exprimés en terme de figure de mérite (*Figure Of Merit*, FOM) à partir de la mesure d'aire sous la courbe ROC (Annexe N, page 167). Le FOM est défini comme la valeur moyenne du ratio du nombre de bonnes détections par rapport au nombre de fausses alarmes, dans l'intervalle de 0 à 10 fausses alarmes en moyenne par mot-clé et par heure. Ainsi, considérant cette aire sous la courbe ROC définie en fonction du taux de détection $T_d$ et du nombre moyen $s$ de fausses alarmes par mot-clé et par heure, la mesure de FOM s'exprime comme :

$$FOM = \frac{1}{10}\int_{s=0}^{10} T_d \, d(s)$$

Les résultats sont présentés sur les sous-ensembles de développement et de test des bases TIMIT, NTIMIT et CTIMIT pour des valeurs optimales de pénalité et de seuil sur la mesure de confiance (Figure 6.1).

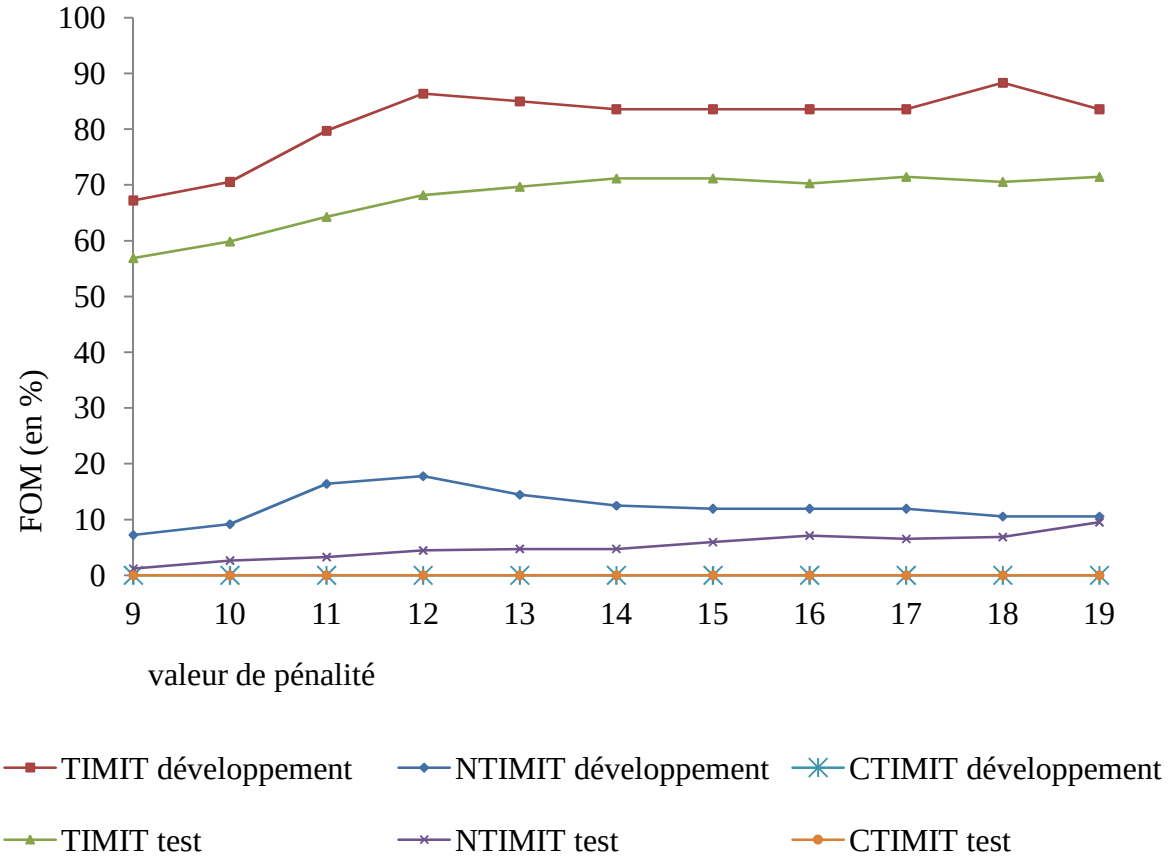



Figure 6.1 : Résultats de l'expérience préliminaire pour la détection de mots-clés (TIMIT, NTIMIT, CTIMIT)

Sur l'ensemble des mots-clés retournés, quelques fausses alarmes sont dues à la réalisation phonétique du mot-clé au sein d'un autre mot prononcé. Ainsi, par exemple durant l'évaluation sur TIMIT, la prononciation du mot-clé « *solid* » peut être réalisée dans le mot prononcé « *solitary* ». De la même manière, la prononciation du mot-clé « *street* » peut être réalisée dans la séquence des mots prononcés « *as treats* » ou dans le mot au pluriel « *streets* ».

Par ailleurs, la baisse de performance entre l'évaluation sur le développement et celle sur le test pour TIMIT et NTIMIT montre une adaptation aux données du développement des deux paramètres variables de pénalité et de seuil sur la mesure de confiance. Cette adaptation est liée au faible nombre de mots-clés présents sur cet ensemble de développement, soit 30 mots-clés présents. Cette faible représentation des mots-clés prive le système d'un ajustement efficace de ces deux paramètres variables. Toutefois, les résultats obtenus permettent de distinguer certaines caractéristiques sur la capacité de robustesse du système évalué en fonction du test.

En effet, ce système de détection de mots-clés de référence retourne de bons résultats lorsque les conditions de test sont proches de celles de l'apprentissage. Dans le cas présent, ces conditions proches correspondent à un signal de parole propre. Lors du test sur TIMIT, l'évaluation porte alors sur une robustesse à la variabilité intrinsèque (section 4.1, page 72). Cependant, lors du test sur NTIMIT, les performances de ce système sont fortement dégradées par l'apparition d'une variabilité extrinsèque (section 4.2, page 77) sur le signal de parole de test. Lors du test sur CTIMIT, le système de détection de mots-clés est même complètement inefficace lorsque le signal de parole de test subit des contraintes de bruit additionnel lié à son environnement acoustique. La dégradation des résultats en fonction du test utilisé est similaire à celle rencontrée par les sous-empreintes issues de MFCCs quantifiés dans les précédentes expériences sur la variabilité intrinsèque (section 5.2.2, page 106) et les variabilités extrinsèque et intrinsèque combinées (section 5.2.4, page 111).

## 6.3 Discussions

Le système de référence développé est basé sur l'apprentissage de HMMs phonétiques et l'ajout d'un modèle poubelle en plus des modèles de mots-clés. D'autres méthodes de détection de mots-clés sont reconnues dans l'état de l'art comme étant plus performantes (Annexe P, page 173). Cependant, ces méthodes sont également plus complexes à mettre en œuvre. Le système de référence choisi s'appuie sur l'optimisation de deux paramètres variables :

- une pénalité pour favoriser le rapport de vraisemblance en faveur du modèle de mot-clé par rapport au modèle poubelle,
- un seuil sur la mesure de confiance. La mesure de confiance est la différence des rapports de vraisemblance entre le modèle de mot-clé et la séquence de modèles phonétiques issue d'un DAP sur le segment temporel du mot-clé.

Compte-tenu de la robustesse rencontrée par les sous-empreintes adaptées de la méthode de Shazam lors de la précédente expérience sur l'évaluation de la robustesse des sous-empreintes (Chapitre 5, page 93), il serait intéressant de connaître la robustesse de telles sous-empreintes lorsqu'elles sont utilisées au sein d'un système d'identification audio adapté pour la détection de mots-clés. A partir du système d'identification audio pour le DAP développé précédemment (Chapitre 3, page 43), il est possible de reprendre les caractéristiques de ces deux paramètres variables pour construire un système d'identification

audio par empreinte adapté à la détection de mots-clés. Dans un tel système d'identification audio, la base de référence générée par les données d'apprentissage pourrait alors contenir (Figure 6.2) :

- une représentation du signal de parole sous la forme de sous-empreintes conservées linéairement suivant leur ordre d'apparition,
- la transcription phonétique correspondante,
- une segmentation complémentaire indiquant les positions de mot-clé dans le continuum de parole.

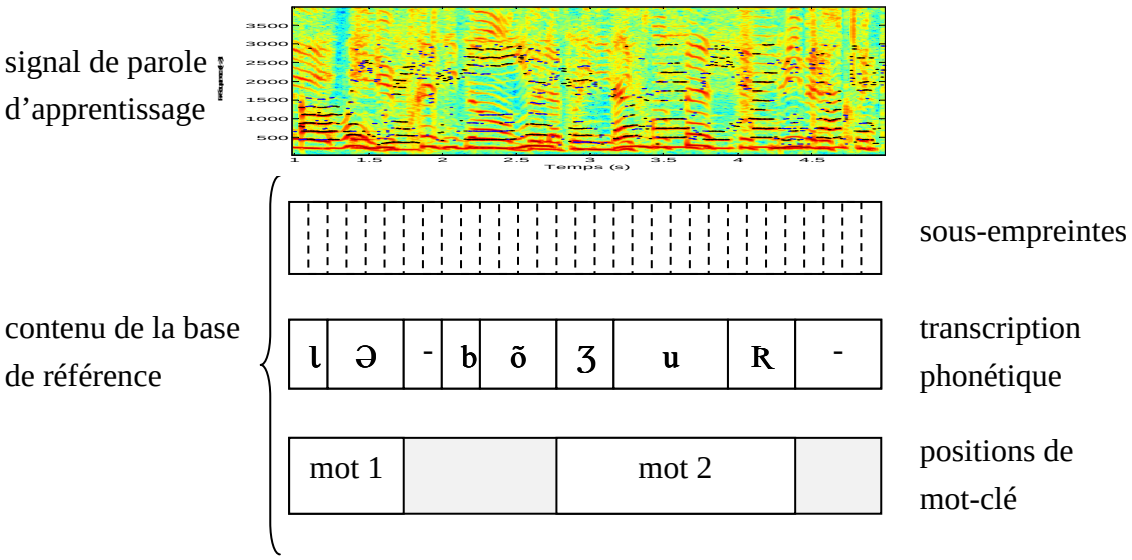



Figure 6.2: Contenu de la base de référence d'un système d'identification audio pour la détection de mots-clés

Au sein de la base de référence, il serait alors possible de distinguer une empreinte constituée par l'étiquette du mot-clé et le segment de sous-empreintes correspondant à la position de ce mot-clé. Lors de l'identification, dans un premier temps, des scores de vraisemblance seraient calculés entre un segment de sous-empreintes issues du signal de parole de test et des segments de sous-empreintes de la base de référence. Ces derniers segments correspondraient tout autant à des empreintes de mot-clé qu'à des séquences de phonèmes prononcés en dehors des positions de mot-clé. Les scores de vraisemblance en provenance des séquences de phonèmes subiraient alors une pénalité en faveur de ceux d'une empreinte de mot-clé. Suite à cette pénalisation, toute empreinte de mot-clé serait conservée comme candidat si son score de vraisemblance est supérieur au meilleur score de vraisemblance obtenu par les séquences de phonèmes.

Puis dans un second temps, une mesure de confiance serait calculée sur chaque candidat. A cet effet, pour chaque candidat, un DAP tel celui proposé dans le précédent système développé (Chapitre 3, page 43) serait alors effectué sur le segment de sous-empreintes du candidat uniquement. Un score de vraisemblance serait alors retourné par ce DAP local. La mesure de confiance serait donc obtenue par la différence des scores de vraisemblance entre le candidat et le DAP local. Si cette mesure de confiance était inférieure à un seuil donné, alors l'étiquette du candidat serait considérée comme un mot-clé détecté. Sinon le mot-clé de l'empreinte candidate serait considéré non prononcé.

Ainsi, les deux paramètres variables présents dans le système de référence à base de HMMs trouvent une correspondance pour le futur développement d'un système d'identification audio par empreinte adapté à la détection de mots-clés. Grâce à un tel système, les sous-empreintes robustes précédemment obtenues par adaptation de la méthode de Shazam (section 5.1.3, page 98) pourraient être évaluées pour leur robustesse à une tâche de détection de mots-clés dans des conditions difficiles (signal de parole de test dégradé par différents types de variabilité par rapport au signal de parole d'apprentissage).

# Conclusion générale

Au cours de ce travail de recherche, nous avons étudié différentes méthodes d'identification audio par empreinte dédiées à l'identification d'extraits de musique. Nous avons évalué dans quelles conditions ces méthodes peuvent être adaptées au domaine de la reconnaissance automatique de la parole (RAP). Puis différentes méthodes de calcul de sous-empreintes ont été évaluées pour leur robustesse à des variabilités extrinsèque et intrinsèque. Nous désirons élargir le cadre de cette évaluation de robustesse de sous-empreintes à la détection de mots-clés dans des conditions difficiles de reconnaissance.

Dans la première partie de ce travail de recherche, nous avons décrit les méthodes d'identification audio par empreinte reconnues pour leurs performances dans la tâche de détection d'extraits de musique (Chapitre 1, page 9). Au sein d'un système d'identification audio par empreinte, trois modules sont présents pour :

- le calcul des sous-empreintes issues de l'analyse acoustique du signal,
- la conservation des empreintes et métadonnées associées dans une base de référence,
- la comparaison des empreintes avec un signal audio de test.

Cependant, les techniques développées par ces méthodes ne sont pas adaptées à la reconnaissance de la parole [Ogle et al., 2007]. Une première adaptation a précédemment été proposée par Vasiloglou [Vasiloglou et al., 2004] pour l'identification de mots isolés dans un environnement mono-locuteur. Des expériences complémentaires ont été effectuées afin d'évaluer ce système dans les tâches de reconnaissance de mots isolés en environnement mono-locuteur (section 2.3.2, page 35) et de reconnaissance de phonèmes isolés en environnement multi-locuteur (section 2.3.3, page 36). Puis nous avons poursuivi cette voie de recherche en construisant un système d'identification audio par empreinte adapté au décodage acoustico-phonétique (DAP). Nous avons développé au cours de ces travaux une technique de relâchement de la contrainte d'identification (section 3.2, page 47) afin d'augmenter l'espace de recherche des sous-empreintes à comparer pour l'identification. De

plus, une étape de programmation dynamique est ajoutée en adaptant certaines considérations de travaux de recherche en synthèse de la parole par concaténation d'unités phonétiques [Sigasaka, 1988]. Cette étape permet d'assembler des séquences de phonèmes afin de réduire les effets de perturbation liés à la coarticulation. De nouvelles méthodes de calcul de sous-empreinte sont également définies sur des MFCCs quantifiés (section 3.3, page 49). L'usage de ces nouvelles sous-empreintes retourne de meilleurs résultats qu'avec celles adaptées de Vasiloglou (section 3.5, page 57). Cependant, ces nouveaux vecteurs sont plus complexes et le système développé reste moins performant qu'un système de référence construit autour de HMMs phonétiques.

Dans la seconde partie de ce travail de recherche, les variabilités intrinsèque (section 4.1, page 72) et extrinsèque (section 4.2, page 77) d'un signal de parole sont décrits ainsi que les différents paramètres acoustiques mis en œuvre pour être robuste à ces variabilités. Puis, nous avons évalué la robustesse de différents types de sous-empreinte à ces variabilités. Les résultats de ces évaluations sont contrastés. Dans un cas, les sous-empreintes issues de l'identification audio par empreinte sont plus robustes à la variabilité extrinsèque que celles issues de paramètres MFCCs quantifiés (section 5.2.3, page 107). Dans l'autre cas, une quantification vectorielle appliquée à des paramètres MFCCs permet le calcul de sous-empreintes plus robustes aux variabilités extrinsèque et intrinsèque inter-locuteur combinées dans la plupart des situations de test (section 5.2.4, page 111). Toutefois, en présence de perturbations liées à l'environnement et aux conditions de transmission du signal de parole, les sous-empreintes issues de l'adaptation de la méthode de type Shazam (section 5.1.3, page 98) sont plus robustes à la variabilité extrinsèque que celles issues de paramètres MFCCs quantifiés (cas de CTIMIT, Figure 5.8, page 113).

Enfin, en perspective, nous avons évalué la robustesse de paramètres MFCCs dans un système de détection de mots-clés de référence [Rose et al., 1990]. Ce système de référence est basé sur l'utilisation de HMMs phonétiques pour la construction de modèles de mots-clés et d'un modèle poubelle. Les variations subies par le signal de parole de test sont autant liées à la variabilité intrinsèque qu'à la variabilité extrinsèque. Ce système s'avère particulièrement inefficace lorsque les conditions du signal de parole de test diffèrent de celles de l'apprentissage par la présence de variabilité extrinsèque (section 6.2, page 124). Hors, dans les précédentes expériences sur l'évaluation de la robustesse des sous-empreintes, les sous-empreintes adaptées de la méthode de Shazam (section 1.4, page 22) sont apparues robustes à ce type de variabilité (Chapitre 5, page 93). Nous proposons alors pour de futurs travaux de recherche d'intégrer des stratégies similaires à celles du système de référence au sein d'un système d'identification audio par empreintes, tel celui développé pour le DAP (Chapitre 3, page 43), afin d'évaluer la robustesse des sous-empreintes dans une détection de mots-clés.

# Annexe A. Empreinte audio

L'emploi du terme « empreinte audio » (*audio fingerprint*) [Haitsma et al., 2002] utilisé dans notre étude ne fait pas référence aux techniques d'identification basées sur l'« empreinte vocale » (*voiceprint*) connue dans le domaine de l'authentification du locuteur [Doddington, 1985]. Le terme empreinte vocale a été introduit dans les années 1960 [Kersta, 1962]. Cependant, l'usage des techniques d'empreinte vocale a suscité une grande polémique. De nombreux scientifiques ont été très critiques envers cette méthode de reconnaissance du locuteur [Bolt et al., 1970]. Depuis lors, le terme « empreinte vocale » est déprécié dans la communauté de la parole. La principale raison d'une telle désapprobation est que les techniques d'empreinte vocale sont liées à l'analyse des composantes dynamiques du signal de parole. Cependant, le signal issu de la production de parole peut être perçu comme un mélange de caractéristiques qui dépendent autant d'aspects physiques que d'aspects liés à l'apprentissage [Faundez-Zanuy et al., 2005]. Donc l'empreinte vocale doit être distinguée des caractéristiques physiques statiques telles que la géométrie de la main ou le motif rétinien.

Depuis une dizaine d'années, l'expression d'empreinte acoustique, également connue sous le terme d'empreinte audio, est apparue de manière régulière dans la communauté du traitement du signal audio [Cano et al., 2002]. Dans ce cas, une empreinte audio est une signature compacte permettant la représentation d'un évènement audio [Cano et al., 2005]. Donc dans notre étude, le terme « empreinte » fera uniquement référence à une empreinte acoustique calculée à partir d'une analyse du signal audio. Si nécessaire, on pourra utiliser l'expression « empreinte de parole » pour définir une telle empreinte audio adaptée à la reconnaissance automatique de la parole.

# Annexe B. Bases de données TIMIT

La base de données TIMIT de parole lue est conçue pour fournir les données acoustico-phonétiques nécessaires au développement et à l'évaluation de systèmes de RAP anglophones [Fischer et al., 1986]. Le succès de l'utilisation de cette base de données de signal de parole propre a donné lieu à la création de nombreuses bases dérivées. Ces bases de données sont formées à partir de l'intégralité ou d'un sous-ensemble particulier de TIMIT. Elles sont adaptées pour répondre à des conditions de transmission spécifiques du signal de parole.

Chaque fichier audio contenant la prononciation d'un message linguistique est associé à un fichier texte contenant la transcription en mots du signal de parole. Cette transcription fournit les marqueurs de segmentation temporelle pour le début et la fin de chaque mot. Ces fichiers de transcription contiennent également une représentation phonétique afin de fournir les marqueurs de segmentation temporelle pour le début et la fin de chaque phonème. L'ensemble de ces bases de données exploitent les mêmes fichiers de transcription phonétique. L'ensemble phonétique choisi est formé de 39 phonèmes, suivant la définition du CMU/MIT [Lee et al., 1989b].

Les bases de données issues de TIMIT et utilisées dans le présent document sont :

- TIMIT [Fischer et al., 1986]

La base de données TIMIT est constituée de phrases de langue anglophone phonétiquement segmentées et prononcées par 630 locuteurs, masculins à 70 %. Compte-tenu des protocoles et contraintes appliqués lors de son enregistrement, la base TIMIT est considérée comme une base de données de signal de parole propre [Fischer et al., 1986].

- NTIMIT [Jankowski et al., 1990]

La base de données NTIMIT est la transmission de la base TIMIT à travers des filtres acoustiques représentant divers canaux de réseaux téléphoniques filaires. La principale différence entre les bases de données TIMIT et NTIMIT est la présence de bruits additifs et convolutifs additionnels [Kamper et al., 2009].

- CTIMIT [Brown et al., 1995]

La base de données CTIMIT est un réenregistrement de la moitié de la base TIMIT par une transmission du signal de parole à travers divers environnements de réseaux téléphoniques cellulaires. La transmission du signal de parole est souvent perturbée par les conditions extérieures et indépendantes du locuteur.

- KED-TIMIT [CSTR, 2001]

La base de données KED-TIMIT est un sous-ensemble de TIMIT constitué de 453 phrases de langue anglophone prononcées par un seul et même locuteur mâle américain.

# Annexe C. Base de données BREF80

La base de données BREF de parole lue est conçue pour fournir les données acoustico-phonétiques nécessaires au développement et à l'évaluation de systèmes de RAP francophones [Gauvain et al., 1990]. La base de données BREF80 est un sous-ensemble de la base de données BREF constituée originellement de 100 heures d'enregistrement. Cette base de données BREF80 est constituée de 5330 phrases de langue francophone phonétiquement segmentées et prononcées par 80 locuteurs, masculins à 45 % [Lamel et al., 1991]. La base de données BREF80 est considérée comme une base de données de signal de parole propre grâce à l'enregistrement du signal audio en environnement contrôlé [Lamel et al., 1991; Besacier et al., 2001].

# Annexe D. Calcul du SNR pondéré

Considérant les énergies moyennes d'un signal de parole $E(S)$ et d'un signal de bruit $E(B)$, le rapport signal à bruit (*Signal to Noise Ratio*, SNR) désiré en dB est donné par :

$$SNR = 10\log_{10}\left(\frac{E(S)}{E(B)}\right)$$

Étant donné que l'énergie d'un signal $E(S)$ pour $n$ échantillons $x$ peut être définie comme :

$$E(S) = \frac{1}{n}\sum_{i=1}^{n} x_i^2$$

Alors, en tenant compte d'un coefficient de pondération $a$ appliqué sur les $n$ échantillons $x$ du signal de bruit $E(B)$, l'énergie d'un signal de bruit pondéré $E(B_a)$ est défini comme :

$$E(B_a) = \frac{1}{n}\sum_{i=1}^{n}(a.x_i)^2 \quad \Leftrightarrow \quad E(B_a) = a^2.\frac{1}{n}\sum_{i=1}^{n} x_i^2$$

En appliquant le coefficient de pondération $a$ sur les échantillons du signal de bruit $E(B)$, l'énergie du signal de bruit pondéré $E(B_a)$ peut s'écrire comme :

$$E(B_a) = a^2.E(B)$$

Donc en tenant compte du SNR désiré, le coefficient de pondération $a$ s'exprime comme :

$$SNR = 10\log_{10}\left(\frac{E(S)}{a^2E(B)}\right) \quad \Leftrightarrow \quad a = \sqrt{\frac{E(S)}{10^{\frac{SNR}{10}}E(B)}}$$

# Annexe E. Programmation dynamique

Dans un système de RAP appliqué à la parole continue, l'objectif est de détecter à la fois le mot prononcé et sa position dans le continuum de parole [Wilpon et al., 1989]. La principale difficulté dans le processus de décision de la reconnaissance du mot est la prise en compte de la distorsion temporelle non linéaire observable entre deux occurrences de prononciation d'un même mot (section 4.1.3, page 74). Cette distorsion est autant présente pour plusieurs locuteurs différents que pour un seul et même locuteur (section 4.1.4, page 76). A cet effet, un algorithme de programmation dynamique permet de comparer des séquences de donnée de taille différente dont les débuts et fins sont similaires [Bellman, 1957]. Dans les années 1970, une première approche pour la détection de mots-clés est fondée sur une extension de la reconnaissance de mots isolés [Bridle, 1973; Christiansen et al., 1976]. Par la suite, une amélioration de cet algorithme permet de relâcher les contraintes sur les régions de début et de fin de la séquence analysée afin de déterminer le meilleur alignement entre les séquences [Myers et al., 1980; Rabiner et al., 1980; Myers et al., 1981].

Dans l'approche par programmation, il s'agit de calculer la distance d'alignement entre un modèle de référence représentant le mot recherché et tous les segments possibles du signal de parole à étudier [Christiansen et al., 1977]. Dans ce contexte, le mot est considéré reconnu dans le segment étudié lorsque la distance d'alignement entre les deux segments est inférieure à un seuil déterminé [Tuffelli et al., 1977]. Dans un premier temps, un score de vraisemblance est calculé pour chaque mot associé à un segment du signal de parole analysé. Chacun des chemins proposés par la programmation dynamique est considéré comme une possible reconnaissance du mot recherché. Dans un second temps, les recouvrements entre segments conservés de mots reconnus sont filtrés. Puis les valeurs de probabilité d'apparition sont alors normalisées afin de déterminer les bonnes reconnaissances par rapport aux fausses alarmes.

Une technique de comparaison par un algorithme d'alignement temporel dynamique (*Dynamic Time Warping*, DTW) permet l'usage d'une fenêtre de séquence d'analyse

acoustique du signal de parole de taille fixe suivant le continuum de parole pour mesurer la distance entre les mots à comparer [Vintsyuk, 1968; Sakoe et al., 1978]. Cet algorithme permet un alignement acoustique non linéaire entre une référence et un signal test à identifier afin de prendre en compte la variabilité du rythme d'élocution [Myers et al., 1981b]. Dans ce cas, lorsque la distance mesurée entre les segments comparés sur plusieurs fenêtres contigües est inférieure à un seuil déterminé, le mot est considéré reconnu. De manière précise, l'algorithme DTW est basé sur un dictionnaire $C$ contenant l'ensemble des $n$ séquences de références $R$ tel que :

$$C = \{R_x\}_{1 \leq x \leq n}$$

Considérant une mesure de distance $D_\propto(S_1, S_2)$ entre deux séquences $S_1$ et $S_2$ selon un critère de similarité $\propto$, il est alors possible de mesurer l'écart, c'est-à-dire le coût de déformation, entre la séquence évaluée $S$ et une séquence de référence $R$. L'objectif de cette mesure est de déterminer le meilleur alignement possible entre la séquence évaluée et la séquence de référence en correspondance. La séquence optimale $\bar{S}$ est donc désignée par :

$$\bar{S} = \underset{R \in C}{\operatorname{argmin}}\, D_\propto(S, R)$$

Différentes mesures de distance peuvent être utilisées selon le type de paramètres acoustiques et le critère de similarité choisis [Myers et al., 1981c]. Lorsque des vecteurs acoustiques à base d'analyse cepstrale sont choisis tels les paramètres MFCCs (section 4.3.2, page 84), l'usage d'une distance $D_n$ utilisant les normes $L_n$ est préféré [Levy, 2006] :

$$D_n(S, R) = \left( \sum_{k=0}^{p} |R(k) - S(k)|^n \right)^{\frac{1}{n}}, \forall n \in [1; +\infty]$$

Dans de nombreux systèmes exploitant ces paramètres acoustiques MFCCs, la distance utilisant la norme $L_2$ est souvent utilisée [Levy, 2006]. Cette distance de norme $L_2$ correspond à la distance euclidienne. Par ailleurs, dans le cas de l'utilisation de paramètres acoustiques à base de prédiction linéaire, la distance d'Itakura-Saito $D_{it}$, également appelée distance d'Itakura, est préférée [Rabiner et al., 1977] :

$$D_{it}(S, R) = \ln \left[ \frac{R^t . R_b . R}{S^t . R_b . S} \right]$$

avec $R_b$ la matrice des coefficients d'auto-corrélation de la séquence de référence $R$ évaluée sur la séquence $S$.

Différents critères d'application permettent de relâcher les contraintes de similarité afin d'accepter diverses variations entre le contenu des deux séquences [Myers et al., 1981c]. En particulier, la prise en compte des contraintes locales de déplacement entre les vecteurs acoustiques d'une séquence donnée permet de mieux représenter les contraintes physiques du mécanisme de production phonatoire [Rabiner et al., 1993]. Ainsi, le coût de déformation cumulé $D(i,j)$ entre deux séquences $S_1$ et $S_2$ aux positions $i$ dans $S_1$ et $j$ dans $S_2$ peut être défini comme par exemple :

$$D(i,j) = \min \begin{pmatrix} D(i-1,j) + d(i,j) \\ D(i,j-1) + d(i,j) \\ D(i-1,j-1) + 2.\,d(i,j) \end{pmatrix}$$

avec $d(i,j)$ la distance locale entre les vecteurs acoustiques aux positions $i$ dans $S_1$ et $j$ dans $S_2$.

Considérant ce coût de déformation cumulé, le meilleur parcours des contraintes locales de déplacement est donc celui qui minimise la distance aux éléments contenus dans la séquence mesurée par rapport à ceux contenus dans la séquence de référence [Vintsyuk, 1968]. Cependant, une valeur locale aberrante d'un signal de parole détérioré à l'intérieur de la séquence mesurée va fortement influer sur la valeur de distance globale pour le choix du parcours [Holmes et al., 1986]. De surcroît, la contrainte de similarité de début et de fin pour l'application de cet algorithme limite les possibilités de comparaison [Kwong et al., 1996]. Il s'agit alors de choisir les contraintes locales les plus adéquates à la représentation des séquences en fonction de l'application envisagée [Ariyaeeinia et al., 1997; Park et al., 2005; Li et al., 2011]. Bien que le coût d'exécution de la DTW classique est de l'ordre de $O(N^2)$, l'ajout de critères d'approximation sur la distance locale permet de considérablement réduire le nombre d'éléments à mesurer [Demuynck et al., 2011].

De telles approches en programmation dynamique dépendent fortement des variations acoustiques dues aux erreurs de locution, aux changements de locuteur ainsi qu'aux perturbations liées aux conditions d'enregistrement et de restitution du signal de parole (Chapitre 4, page 71). Afin d'augmenter la robustesse de ces approches, les distances d'alignement peuvent être calculées non seulement par rapport au mot recherché mais également en discrimination par rapport à d'autres mots de référence [Higgins et al., 1985]. En l'occurrence, un tel système retourne une liste constituée de la concaténation des mots de référence minimisant la distance avec le segment de signal de parole à analyser [Chang et al., 1993]. Dans ce cas, le mot recherché est considéré reconnu s'il est contenu dans cette liste de mots concaténés. Donc la distance d'alignement du mot recherché n'est pas considérée comme un nombre absolu. Au contraire, cette distance d'alignement dépend alors également

des distances par rapport aux autres mots de référence de la liste retournée. Ce calcul de distance relative permet ainsi d'augmenter la robustesse du système par rapport aux variations liées à la variabilité extrinsèque (section 4.2, page 77).

Cependant, les approches basées sur une DTW ont montré certaines limites lors de l'augmentation de la taille des ensembles de bases de données de signal de parole disponibles pour l'apprentissage des systèmes [Wong et al., 1998].

# Annexe F. Modèle de Markov caché

Des systèmes à base de modèles de Markov cachés (*Hidden Markov Model*, HMM) ont été développés pour la RAP en général [Bahl et al., 1983; Bahl et al., 1986; Montacié et al., 1996] et pour la détection de mots-clés en particulier [Kawabata et al., 1988; Wilpon et al., 1990; Wilcox et al., 1991; Foote et al., 1995; Rose, 1995]. Ces HMMs sont capables de modéliser simultanément les caractéristiques fréquentielles et temporelles du signal de parole [Young et al., 2006]. Par rapport aux approches basées sur une DTW, ce type de modélisation améliore la robustesse aux changements de locuteur et aux changements de canal de transmission lorsque plusieurs prononciations du mot-clé considéré sont disponibles dans la base d'apprentissage [Class et al., 1990]. De précédents travaux décrivent de manière précise la construction de tels HMMs et leur utilisation au sein d'un système de RAP [Lefèvre, 2000].

Un modèle de Markov caché (HMM) est un modèle statistique contenant des variables cachées [Jelinek, 1976]. Il s'agit d'un automate à états finis qui permet de modéliser les aspects stochastiques du signal de parole [Paul, 1985]. Ce modèle est constitué d'un ensemble d'états liés entre eux par un certain nombre de transitions permises [Schwartz et al., 1984]. Dans ce cas, chaque fois qu'une observation est émise, le système procède au passage d'un état à l'autre ou au bouclage dans le même état selon les transitions permises. De manière générale, les HMM utilisés en RAP sont d'ordre 1 compte-tenu de l'aspect séquentiel du signal de parole [Bakis, 1976]. Cet ordre 1 signifie que la possibilité de se trouver dans un état donné d'un HMM à un instant $(t+1)$ ne dépend que de l'état dans lequel le système se trouvait à l'instant $t$. D'autres possibilités de modélisation, comme les HMMs d'ordre 2, existent mais rendent les systèmes de RAP plus complexes [Mari et al., 1994]. Pour chaque HMM, un état de début et un état de fin sont ajoutés à ces états d'observation pour assurer la transition lors de l'enchaînement des HMMs les uns à la suite des autres durant le processus de reconnaissance. De surcroît, un HMM contient pour chacun de ses états une probabilité d'émission [Schwartz et al., 1984; Bourlard et al., 1985]. Cette probabilité est souvent

représentée par une distribution statistique qui retourne un taux de vraisemblance pour chaque vecteur observé [Rabiner, 1989].

Au sein d'un HMM à base de modèle discret, les vecteurs de paramètres acoustiques représentant le signal de parole sont quantifiés indépendamment des variables liées à l'état caché du modèle [Schwartz et al., 1984; Bourlard et al., 1985; Paul, 1985]. L'information acoustique au sein d'un HMM peut également être représentée sous la forme de densités continues [Haeb-Umbach et al., 1993; Rabiner et al., 1993]. Ainsi, l'utilisation de HMMs dits continus supprime tout besoin de quantification des vecteurs acoustiques [Rabiner, 1989]. Cette approche de modélisation continue fait preuve d'une grande adaptabilité aux variations du locuteur et du canal de transmission [Jarre et al., 1987]. Les distributions associées alors aux HMMs sont des distributions à base de densités de probabilité, souvent représentées sous la forme de modèles de mélange de gaussiennes (*Gaussian Mixture Model*, GMM) [Korkmazskiy et al., 1997]. Dans ce cas, le système effectue en même temps l'apprentissage des GMMs et celui des probabilités de transition d'un état au suivant. Lors de l'émission d'un vecteur d'observation $x_n$ à l'instant $t$, les différentes probabilités de transition $P$ de l'état $q$ à l'instant $(t-1)$ vers l'état suivant sont alors déterminées par les GMMs associées à ces états (Figure F.1).

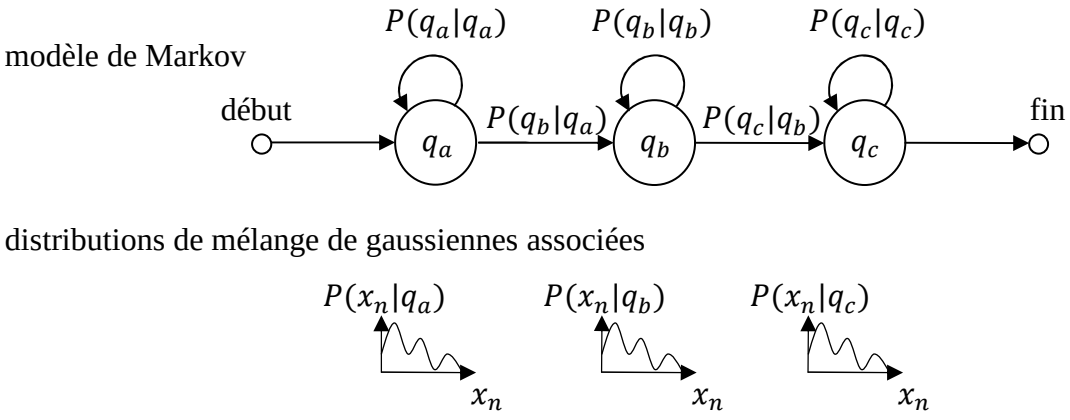



Figure F.1 : Représentation d'un HMM gauche-droit à 3 états avec GMMs [Young et al., 2006]

Le décodage de parole est concrétisé par l'usage de l'algorithme de Viterbi [Viterbi, 1967] afin de trouver la meilleure concaténation de HMMs entraînés individuellement [Bourlard et al., 1985]. L'algorithme EM (*Expectation-Maximization*) [Dempster et al., 1977] permet alors de maximiser la vraisemblance des données d'apprentissage à partir de leur transcription en unités linguistiques qui leur est associée [Ephraim et al., 1989]. Cet algorithme permet de déterminer l'alignement optimal de la forme acoustique sur le HMM considéré, en retournant la plus forte probabilité d'émission de cette forme acoustique [Rose et al., 1990]. Il existe des systèmes de RAP à base de HMM modélisant des mots dans leur intégralité [Lee et al., 1989]. Toutefois, en règle générale, chaque HMM défini par le système représente une unité linguistique (e.g, phonème, syllabe, diphone ou autre unité de sous-mot) [Lee et al., 1988; Wood et al., 1991; Hofstetter et al., 1992; Marcus, 1992].

Une des principales évolutions introduites par l'usage de HMM dans la RAP est l'adoption d'une modélisation par unité linguistique, souvent par phonème ou l'un de ses dérivés [Kawabata et al., 1988; Lee et al., 1988b; Rose et al., 1990]. En effet, dans de tels systèmes de RAP, le modèle HMM d'un mot est composé d'une séquence de plusieurs HMMs d'unité linguistique partagés entre tous les mots représentés. Les systèmes de reconnaissance de mots à base de phonèmes, dédiés par exemple pour la détection de mots-clés, sont fréquemment construits autour de HMMs à 3 états par phonème [Bourlard et al., 1985]. Dans ce cas, l'hypothèse généralement retenue est que l'état central du HMM représente la partie stationnaire du phonème. Les états périphériques d'un tel HMM représentent alors les effets de coarticulation dus à la liaison du phonème modélisé avec les phonèmes voisins [Rabiner, 1989]. Dans un système de RAP à base de HMM, les bases de données de signal de parole fournies pour l'apprentissage des modèles sont souvent phonétiquement annotées [Wagner, 1981; Ljolje et al., 1991]. Dans ce cas, un dictionnaire des phonèmes aux mots les accompagne. Cette représentation sous forme de séquences de phonèmes est alors utilisée pour la modélisation acoustique des mots [Riley, 1991]. Un même mot peut être représenté par différentes séquences suivant les réalisations phonétiques possibles lors de la prononciation de ce mot [Riley, 1991]. Les différentes transcriptions fournies en mots et en phonèmes garantissent alors une indépendance de l'interprétation du transcripteur humain lors de l'utilisation d'un système de RAP [Carney, 1994]. Donc dans le cas d'une approche par représentation phonétique, la fonction de reconnaissance de mots contient chacun des mots sous la forme de séquence(s) de phonèmes représentant le mot [Knill et al., 1996].

Grâce à l'usage de HMMs pour définir une représentation de l'espace phonétique, le modèle d'un mot donné bénéficie donc non seulement de l'apprentissage liée à la présence de ce mot dans la base d'apprentissage mais également de l'apprentissage de l'ensemble des phonèmes qui le composent [Chen, 1986]. Un autre avantage important à ce type de modélisation phonétique est la capacité à identifier des mots qui sont absents de la base d'apprentissage. Dans ce cas, un nouveau modèle de mot est fabriqué à partir d'une concaténation des modèles de phonèmes déjà appris qui le composent [Knill et al., 1995]. Cette possibilité permet désormais d'élargir les applications de reconnaissance de mots à des ensembles de mots non connus lors de la phase d'apprentissage du système.

# Annexe G. Classes phonétiques

Les réalisations acoustiques élémentaires que représentent les phonèmes [Nolan, 1983] peuvent être regroupées en classes phonétiques selon des caractéristiques communes facilement discriminables. Ainsi parmi toutes les réalisations acoustiques possibles, les principales classes phonétiques présentes en français et en anglais sont [Adamczewski et al., 1977] :

- les voyelles. Il s'agit principalement des voyelles de l'écrit. Ces phonèmes se caractérisent par l'effet de voisement qui produit des formants. Les formants sont des zones fréquentielles à forte densité d'énergie. Ils correspondent à l'intérieur du conduit vocal à une résonance de la fréquence fondamentale produite par la vibration des cordes vocales. Cette caractéristique formantique permet par ailleurs de distinguer les voyelles les unes des autres par l'analyse des valeurs de leur premier et second formant.

- les occlusives. Appelées également plosives, ces phonèmes se caractérisent par un instant de fermeture du conduit vocal, suivi d'un brusque relâchement. Les occlusives sont donc constituées de deux parties : une première partie de silence lors de la fermeture et une seconde partie d'explosion au moment du relâchement. Les occlusives peuvent être soit voisées, on les appelle occlusives sonores, soit non-voisées, on les appelle occlusives sourdes.

- les fricatives. Ces phonèmes sont produits lors de la friction du flux d'air traversant le conduit vocal et arrivant au niveau des lèvres, des dents ou de la langue. Un bruit de hautes fréquences apparaît alors. Les fricatives peuvent être sonores ou sourdes.

- les semi-consonnes. Appelées également semi-voyelles ou glissantes, ces phonèmes voisés permettent de représenter les caractéristiques acoustiques rencontrées lors de la transition d'une voyelle à une autre.

- les liquides. Ces phonèmes voisés possèdent les mêmes caractéristiques que les semi-consonnes mais avec des durées et des énergies plus faibles.

- les nasales. Ces phonèmes voisés sont obtenus par le passage de l'air dans le conduit vocal restreint au conduit nasal à partir des cordes vocales.

- les diphtongues (anglo-américain). Ces phonèmes sont caractérisés par deux états stables formantiques et leur transition.

- les affriquées (anglo-américain). Ces phonèmes sont caractérisés par une occlusive immédiatement suivie d'une fricative courte.

# Annexe H. Modélisation statistique

Les systèmes de RAP en parole continue s'appuient majoritairement sur une approche statistique issue de la théorie de l'information [Boite et al., 2000]. En RAP, l'approche par modélisation statistique permet de modéliser la référence acoustique d'un mot par un modèle de référence [Jelinek, 1976]. Le continuum de parole est souvent représenté par une suite de vecteurs d'observation calculés sur des portions du signal de courte durée [Young et al., 2006] (section 4.4.1, page 86). Les vecteurs d'observation peuvent représenter le signal de parole sous la forme de différents types de paramètres acoustiques [Davis et al., 1980; Mokbel, 1992; Eyben et al., 2010] (section 4.3, page 80).

A partir des vecteurs d'observation $X$ formés par les paramètres acoustiques, le module de reconnaissance de la parole s'applique à trouver la séquence de mots prononcés $\bar{W}$ la plus probable parmi les séquences de mots possibles. Ces mots possibles sont issus d'un vocabulaire $V$ défini. Soit $P(W|X)$ la probabilité d'obtenir le mot $W$ à partir des vecteurs d'observation $X$, il s'agit donc de rechercher la séquence maximisant l'équation suivante [Bahl et al., 1983] :

$$\bar{W} = \arg\max_{W \in V} P(W|X)$$

En appliquant la théorie de Bayes [Lin et al., 1977] à l'équation ci-dessus et en tenant compte de l'indépendance entre vecteurs d'observation $X$ et mots du vocabulaire $V$, on obtient :

$$\bar{W} = \arg\max_{W \in V} P(X|W)P(W)$$

La probabilité $P(X|W)$ d'apparition des vecteurs d'observation $X$ connaissant le mot $W$ est issue de la probabilité des modèles acoustiques correspondants entraînés pendant la phase d'apprentissage. La probabilité $P(W)$ d'apparition du mot $W$ est estimée par le modèle de

langage associé au module de reconnaissance. Ainsi, la prise de décision intègre à la fois les informations acoustiques et linguistiques [Young et al., 2006] (Figure H.1).

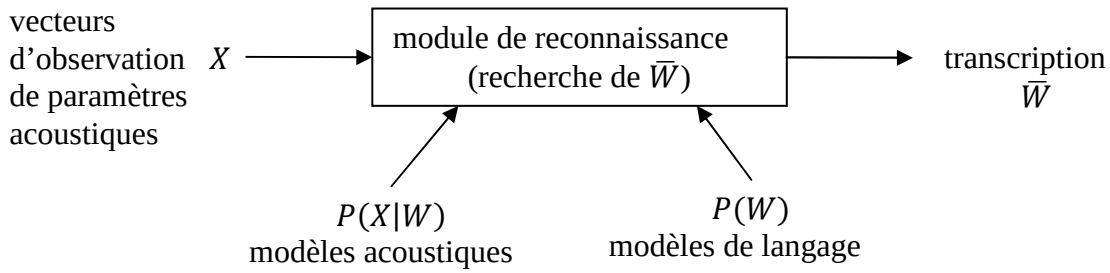



Figure H.1 : Principe général du module de reconnaissance de la parole [Young et al., 2006]

Afin de contraindre le nombre de modèles acoustiques nécessaires à la représentation de l'ensemble des mots du vocabulaire, il est possible de modéliser le signal acoustique de la parole sous la forme d'un nombre limité d'unités linguistiques. Ces unités linguistiques sont alors considérées comme des sons unitaires de la langue (section 4.1.2, page 74). Ainsi, le module de reconnaissance de la parole ne contient qu'un nombre fini de modèles acoustiques d'unités linguistiques [Mergel et al., 1985]. Le signal de parole est alors considéré comme la succession de ces modèles d'unités linguistiques. Par ailleurs, en l'absence de connaissance sur le langage, le module de reconnaissance de la parole est simplifié à la recherche de la séquence $\bar{W}$ telle que :

$$\bar{W} = \arg\max_{W \in V} P(X|W)$$

Compte-tenu de la variabilité intrinsèque de la parole (section 4.1, page 72), la prononciation d'un mot génère un segment de vecteurs d'observation dont la taille et les caractéristiques acoustiques sont la plupart du temps uniques.

# Annexe I. Combinaison de paramètres acoustiques

La recherche d'ensembles de paramètres avec différentes propriétés de représentation acoustique du signal de parole est importante. En effet, la combinaison de tels paramètres acoustiques permet la représentation d'informations complémentaires présentes dans chacun de ces ensembles. Dans ce cadre, le calcul de plusieurs flux isolés de paramètres acoustiques à partir d'une analyse du spectre en sous-bandes de fréquence permet de tenir compte d'une évolution différente de ces paramètres dans chacune des sous-bandes de fréquence à travers le temps [Bourlard et al., 1997; Tibrewala et al., 1997; Tomlinson et al., 1997]. Par ailleurs, à partir de cette approche multi-flux de vecteurs acoustiques, de nombreuses techniques de mesure acoustique du signal de parole sont apparues. Ces techniques à base de combinaisons de paramètres acoustiques peuvent être constituées par :

- les corrélations temps/fréquence à résolution multiple [Vaseghi et al., 1997; Hariharan et al., 2001],
- les paramètres acoustiques à base de segments et de trames [Hon et al., 1999],
- les paramètres de types MFCC, PLP et paramètres auditifs [Jiang et al., 1999],
- les paramètres discriminants et à base d'analyse spectrale [Benitez et al., 2001],
- les paramètres acoustiques et articulatoires [Kirchhoff, 1998; Tolba et al., 2002],
- les paramètres issus du cepstre à base de LPC, de coefficients MFCC, de coefficients PLP [Zolnay et al., 2005] et encore ceux issus de moyennes temporelles d'énergies spectrales [Omar et al., 2002; Omar et al., 2002b],
- les coefficients cepstraux et les paramètres PLP avec un usage de sous-bandes fréquentielles [Kingsbury et al., 2002],
- les paramètres de types MFCC, PLP et en ondelettes [Gemello et al., 2006],
- les paramètres liés issus d'une fonction de délai modifié [Hedge et al., 2005],

- les paramètres de types MFCC et RASTA PLP combinés à des filtres de fréquence [Pujol et al., 2005].

Par ailleurs, d'autres approches exploitent l'information obtenue par des paramètres acoustiques spécifiques au sein d'un même flux d'analyse du signal de parole. Ces paramètres spécifiques peuvent être issus de l'étude de :

- la périodicité de motifs acoustiques et les vibrations dans le signal de parole à moyen terme [Thomson et al., 1998],
- l'expression dans la voix [Zolnay et al., 2002; Gracianera et al., 2004],
- la qualité de la parole et l'analyse du pitch [Stephenson et al., 2004].

Enfin, d'autres méthodes font appel à une modélisation statistique. Par exemple, il est possible d'utiliser la combinaison des techniques de perceptrons multicouches au sein de chaînes de Markov cachées (*MultiLayer Perceptron - Hidden Markov Model*, MLP-HMM) d'une part avec celles de mélanges de gaussiennes au sein de chaînes de Markov cachées (*Gaussian Mixture Model - Hidden Markov Model*, GMM-HMM) d'autre part [Ellis et al., 2001]. Une telle combinaison exploite alors l'information de probabilité d'apparition en sortie des MLP-HMMs afin d'intégrer par la suite cette information au sein du vecteur d'observation en entrée des GMM-HMMs. D'autres combinaisons sont possibles pour exploiter les propriétés de paramètres acoustiques issus de MLPs [Zhu et al., 2004]. Dans d'autres cas, ces paramètres acoustiques issus de MLPs peuvent être combinés avec des paramètres TRAPs [Morgan et al., 2004] ou en conjonction avec des filtres de Gabor [Kleinschmidt et al., 2002]. Par ailleurs, la représentation statistique par GMM peut également être utilisée pour représenter la présence ou l'absence de paramètre acoustique discriminant [Eide, 2001].

Cependant, dans toutes ces méthodes, l'utilisation de combinaisons de paramètres acoustiques génère des vecteurs complexes de représentation acoustique du signal de parole. Ces vecteurs acoustiques complexes peuvent alors contenir de l'information redondante. Une étape de réduction du nombre des dimensions de tels vecteurs acoustiques est alors applicable afin de simplifier les vecteurs en diminuant leur taille tout en limitant la perte d'information utile.

# Annexe J. Optimisation du vecteur acoustique

La recherche de transformations optimales pour réduire le nombre de dimensions d'un vecteur acoustique est un domaine d'étude important. En effet, le choix de paramètres acoustiques optimaux est obtenu en maximisant l'information mutuelle entre l'ensemble des paramètres acoustiques définis et les classes d'unités linguistiques correspondantes [Omar et al., 2002b; Padmanabhan et al., 2005]. A cet effet, il est possible de retourner une analyse par composante principale (*Principal Component Analysis*, PCA) afin de déterminer les coefficients les plus discriminants du vecteur de paramètres acoustiques [Tokuhira et al., 1999]. Cependant, une analyse discriminante linéaire (*Linear Discriminant Analysis*, LDA) peut être préférée afin de résoudre le problème de discrimination sous-optimale parfois rencontré lors de l'application d'une PCA [Fukunaga, 1972; Duda et al., 1973]. De bons résultats avec la méthode par LDA sont obtenus pour une tâche de RAP à petit vocabulaire [Haeb-Umbach et al., 1992; Haeb-Umbach et al., 1993; Siohan, 1995]. Cette méthode a été améliorée par l'usage d'une analyse discriminante non linéaire (*NonLinear Discriminant Analysis*, NLDA) [Fontaine et al., 1997].

Pour les tâches de reconnaissance avec un vocabulaire plus important, une analyse discriminante hétéroscédastique (*Heteroscedastic Discriminant Analysis*, HDA) permet de tenir compte de matrices de covariance à l'intérieur des classes d'unités linguistiques définies [Kumar et al., 1998]. En outre, une mesure de distance de Chernoff peut également être utilisée pour évaluer la distance entre les classes d'unités linguistiques [Loog et al., 2004]. Cependant, une telle distance statistique ne répond pas aux critères de définition d'une métrique de distance de densité de probabilités. Les analyses LDA et HDA peuvent être combinées grâce à une transformation linéaire par maximum de vraisemblance [Gopinath, 1998; Saon et al., 2000]. Dans ce cas, le traitement effectué est similaire à l'usage de matrices

de covariance [Gales, 1999]. D'autres combinaisons sont possibles. Par exemple, l'analyse HDA peut être combinée avec le critère d'erreur phonétique minimum [Zhang et al., 2005].

Cependant, le choix de ces méthodes est coûteux car il nécessite le calcul du vecteur acoustique de représentation du signal de parole au complet avant la réduction du nombre de ses dimensions. Il est alors possible de définir d'autres stratégies pour choisir le meilleur espace de représentation comme par exemple par l'utilisation du principe du maximum d'entropie [Abdel-Haleem et al., 2004].

# Annexe K. Distance KL2

La divergence de Kullback-Leibler, dite divergence $KL$ ou entropie relative, est une mesure de dissimilarité entre deux distributions de densité de probabilités [Kullback et al., 1951]. La métrique de distance dite $KL2$ entre deux distributions de densité de probabilité $C_1$ et $C_2$ est alors définie comme [Siegler et al., 1997] :

$$KL2(C_1, C_2) = KL(C_1, C_2) + KL(C_2, C_1)$$

Comme $C_1$ et $C_2$ sont des distributions de densité de probabilité de moyenne et variance respectivement $(\mu_1, \sigma_1^2)$ et $(\mu_2, \sigma_2^2)$, on obtient alors la distance :

$$KL2(C_1, C_2) = \frac{\sigma_1^2}{\sigma_2^2} + \frac{\sigma_2^2}{\sigma_1^2} + (\mu_2 - \mu_1)^2 . \left(\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}\right)$$

La distance $KL2$ peut être donc utilisée pour mesurer un rapport de proximité entre deux distributions de densité de probabilité. Pourtant à l'origine, la divergence $KL$ ne remplit cependant pas tous les axiomes de la définition d'une distance. En effet, elle n'est pas symétrique et ne respecte pas le critère de l'inégalité triangulaire. Donc lors de la comparaison de fonctions issues de distributions de densité de probabilités, d'autres mesures de distance peuvent être préférées comme par exemple la distance de Bhattacharyya [Bhattacharyya, 1943].

# Annexe L. Applications d'un système de détection de mots-clés

La détection de mots-clés permet d'utiliser les techniques de recherche d'information développées initialement pour le texte (formules logiques ou recherches vectorielles) [Salton, 1962]. A cet effet, les systèmes de détection de mots-clés ont dans un premier temps été utilisés pour le développement d'applications analysant le signal de parole comme :

- la mise en œuvre de serveurs vocaux interactifs par téléphone [Millar et al., 1988; Wilpon et al., 1990; Matsu'ura et al., 1994],
- la classification de messages vocaux [Rose et al., 1991],
- la numérotation téléphonique par ordre vocal [Nakamura et al., 1993].

Depuis quelques années, les systèmes de détection de mots-clés sont utilisés dans le cadre de nouvelles applications manipulant des contenus multimédia pour :

- la structuration de documents [Bohac et al., 2011],
- le filtrage de contenus [Rouvier et al., 2008],
- la recherche de thématique [Hazen et al., 2008],
- le résumé automatique [Nakazawa et al., 2001],
- l'indexation audio [Peng et al., 2008].

Au sein de toutes ces applications, il est nécessaire de distinguer les différentes tâches de détection de mots-clés, de recherche de termes parlés et de recherche de document audio :

- la détection de mots-clés (*KeyWord Spotting*, KWS) est appliquée pour la recherche de mots particuliers lors de l'analyse d'un continuum de parole [Medress et al., 1979]. Le flux audio du signal de parole est analysé en continu afin d'indexer les mots identifiés au fur et à mesure de leur apparition [Wilpon et al., 1989].

- la recherche de termes parlés (*Spoken Term Detection*, STD) est la continuité de la détection de mots-clés. Il s'agit de trouver toutes les occurrences d'un terme parlé, constitué par la séquence d'un ou plusieurs mots adjacents, dans un ensemble d'archives audio [NIST STD, 2006]. Cette tâche est constituée de deux processus distincts d'indexation des archives dans un premier temps puis de recherche des termes à détecter au sein de ces archives dans un second temps [Vergyri et al., 2007]. Au cours de la campagne NIST STD [NIST STD, 2006], l'évaluation de la tâche de recherche de termes parlés a consisté à détecter 1000 mots-clés référencés sur trois types de bases de données orales (3 heures de journaux télévisés, 3 heures de conversation téléphonique et 2 heures de transcription de réunions). Les performances observées des systèmes participants varient considérablement en fonction du type de base de données [Vergyri et al., 2007]. Par exemple, pour un taux fixé à une fausse alarme pour 25 000 mots-clés détectés, les taux de détection des différents systèmes sont de l'ordre de 90 % pour la base de données de journaux télévisés, de 85 % pour la base de données de conversation téléphonique et de 30 % pour la base de données de transcription de réunions [NIST STD, 2006].

- la recherche documentaire audio (*Spoken Document Retrieval*, SDR) s'attache à répondre à une requête textuelle dans de grandes archives de données audio [Garofolo et al., 2000]. Cette tâche s'apparente à une application de type moteur de recherche [Johnson et al., 1999]. A la différence de la recherche de termes parlés qui localise le moment d'apparition du terme à détecter, la recherche documentaire audio retourne la liste des documents audio contenant la requête textuelle [Ng et al., 1998]. Un ensemble de campagnes d'évaluation dédiées à cette tâche est mis en place par le NIST : ce sont les campagnes TREC SDR (*Text REtrieval Conferences*) [NIST SDR, 1997]. Certaines de ces évaluations ont montré que la détérioration de la reconnaissance peut être compensée par la recherche de mots redondants [Garofolo et al., 2000]. Par ailleurs, une classification par groupes de mots appartenant à un même champ sémantique peut améliorer les performances pour la tâche de SDR [Allan, 2002].

D'autres campagnes d'évaluation comme par exemple les campagnes ESTER (*Evaluation des Systèmes de Transcription enrichie d'Emissions Radiophoniques*) ont permis une évaluation générale des systèmes de RAP dédiés au traitement de la langue française [Gravier et al., 2004; Deléglise et al., 2005; Galliano et al., 2005]. Dans ces évaluations, différentes tâches de RAP sont dédiées à la transcription et à l'indexation de journaux radiophoniques. Parmi les analyses résultantes, il apparaît que la segmentation par locuteur permet une amélioration des performances des systèmes de RAP [Galliano et al., 2009].

Enfin, d’autres types de campagnes encore visent à évaluer des systèmes multilingues et inter-langues pour la recherche d’information dans plusieurs langues comme les campagnes annuelles CLEF (*Cross Language Evaluation Forum*) [CLEF, 2010]. Les campagnes CLEF permettent notamment la mise en œuvre de méthodes d’évaluation, par exemple, pour l’étude des systèmes de résumé automatique de documents multiples dans un cadre multilingue [Turchi et al., 2010].

# Annexe M. Modèle poubelle

Des modèles de rejet, dits modèles poubelles (*garbage model*), peuvent compléter un système de détection de mots-clés et être utilisés comme modèles discriminants pour la détection des mots-clés [Chigier, 1992]. L'objectif de tels modèles poubelles est d'absorber les portions du signal de parole extérieures aux mots-clés lors de l'étape de reconnaissance. Ces modèles poubelles viennent alors en concurrence des modèles de mots-clés lors de l'étape de reconnaissance pour la détection des mots-clés [Sukkar, 1994]. Dans ce cas, un tel système de modélisation est alors constitué de deux parties (Figure M.1) :

- un ensemble de modèles représentant les mots-clés,
- un ensemble de modèles représentant les parties du signal de parole en dehors des mots-clés, ce sont les modèles poubelles.

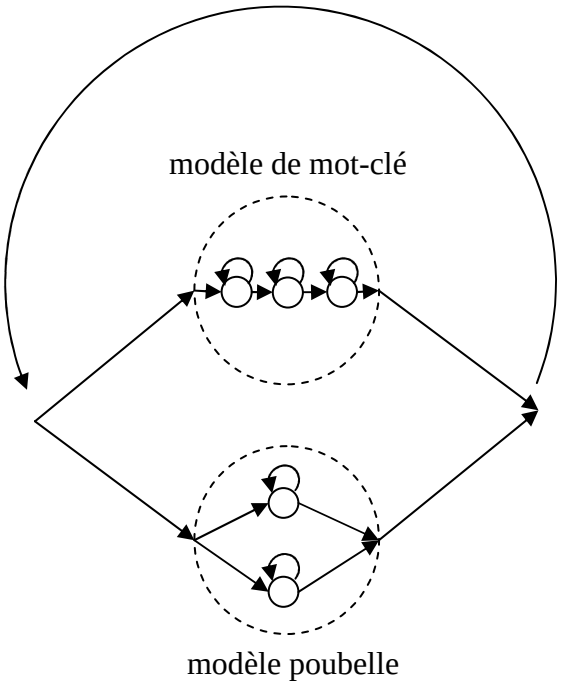



Figure M.1 : Topologie d'un système de détection de mots-clés avec modèle poubelle [Grangier et al., 2009]

Etant donné un système conçu autour d'une telle modélisation en deux parties distinctes, la détection de mots-clés est effectuée par la recherche de la séquence d'états retournant le maximum de vraisemblance sur le signal de parole analysé. Cette opération est effectuée à travers un décodage de Viterbi [Rose, 1992]. Dans ce cas, la détection d'un mot-clé donné est déterminée en analysant si le meilleur chemin retourné par le décodage de Viterbi traverse le modèle du mot-clé ou non. Dans une telle modélisation, le choix des probabilités de transition entre les états constituant le modèle du mot-clé établit le compromis entre taux de fausses alarmes et taux de faux rejets. Une fausse alarme correspond ici à une détection du mot-clé à tord. Un faux rejet correspond, quant à lui, à la non-détection d'un mot-clé présent.

La modélisation des parties du signal de parole en dehors des mots-clés est un élément important dans le choix de la définition des modèles poubelles [Rose, 1995]. D'un coté, la modélisation la plus simple pour un modèle poubelle est la connexion complète de tous les modèles d'unités linguistiques les uns avec les autres [Rohlicek et al., 1989; Rose et al., 1990; Wilpon et al., 1990]. D'un autre coté, la modélisation la plus complexe pour un modèle poubelle est la connexion de HMMs représentant les mots issus d'un grand vocabulaire tout en excluant les mots-clés à détecter [Rose, 1992; Rohlicek et al., 1993; Weintraub, 1993]. Cette dernière approche permet une meilleure représentation du modèle poubelle grâce à l'utilisation d'une information linguistique [Fetter et al., 1996]. Cette approche augmente cependant le coût de calcul nécessaire au décodage du signal de parole. Elle nécessite également une plus grande quantité de données d'apprentissage, en particulier pour la définition du modèle de langage associé. Par ailleurs, d'autres approches s'attachent à modéliser sous forme d'anti-modèles les différents types d'évènements acoustiques considérés perturbateurs pour le modèle correspondant. Ainsi, dans de telles approches, à chaque modèle de phonème correspond un anti-modèle modélisant les erreurs de substitutions et de fausses acceptations de ce phonème [Sukkar et al., 1996]. Diverses variations de construction de modèle poubelle sont possibles en fonction du compromis choisi entre complexité et performance [Boite et al., 1993; Bourlard et al., 1994; Jitsuhiro et al., 1998].

Le résultat du décodage de Viterbi dépend de la séquence des décisions locales pour le choix du meilleur parcours. Ce choix peut s'avérer fragile et faiblement robuste selon les variations entre les modèles locaux appris lors de l'apprentissage et le signal de parole analysé. Dans le cadre présent de la détection de mots-clés à base de modélisation HMM, un mot-clé peut ne pas être détecté si le HMM de la première unité linguistique qui le constitue n'est pas représentatif de l'étendue des possibilités de prononciation de cette unité [Kanazawa et al., 1995]. Afin d'améliorer la robustesse à de telles variations, des approches par rapport de vraisemblance ont été proposées [Rose et al., 1990; Weintraub, 1995]. Dans ce cas, le

score de confiance en sortie du système de détection de mots-clés correspond au rapport entre la vraisemblance estimée par un HMM incluant l'occurrence du modèle du mot-clé détecté (HMM A) et celle estimée par un HMM excluant ce modèle de mot-clé (HMM B) (Figure M.2). La détection de mots-clés est alors effectuée en comparant les scores de sortie par rapport à un seuil prédéfini.

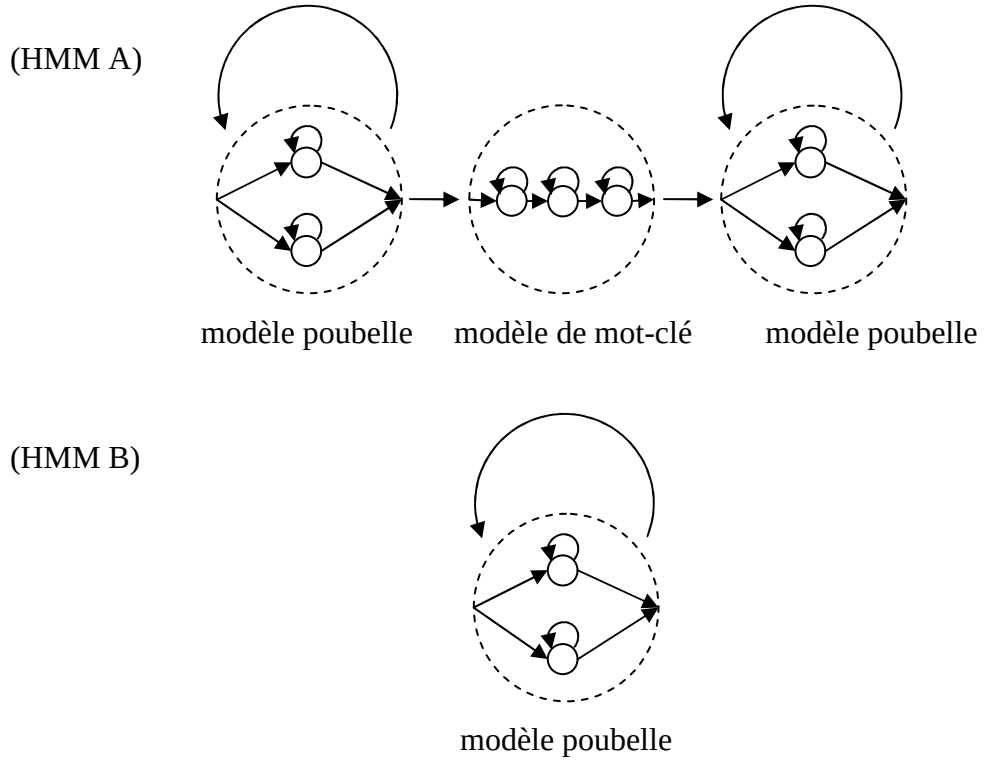



Figure M.2 : Topologie d'un système de détection de mots-clés avec stratégie de rapport de vraisemblance [Grangier et al., 2009]

Diverses approches utilisant HMMs et modèles poubelles ont été développées. Par exemple, d'autres unités linguistiques que les simples phonèmes peuvent être utilisées, comme les syllabes [Lleida et al., 1993; Klemm et al., 1995; El Meliani et al., 1998] ou les triphones [Rose et al., 1993]. Dans ce dernier cas, un arbre de décision permet de réduire le nombre de classes d'apprentissage définies dans le système. Une autre approche consiste en la construction de modèles poubelles à partir de la taille des mots en fonction de leur nombre de phonèmes [Zhang et al., 2001]. Ainsi dans ce cas, différents modèles poubelles permettent de modéliser les mots à deux, trois, quatre, cinq ou six phonèmes et plus. Il peut par ailleurs être possible de ne calculer le rapport de vraisemblance entre modèles que sur la portion de signal de parole où le mot-clé a été considéré comme détecté [Junkawitsch et al., 1997].

# Annexe N. Mesures d'évaluation pour la détection de mots-clés

Contrairement à la RAP en parole continue, l'évaluation de la détection de mots-clés porte sur la capacité du système à minimiser :

- le nombre de fausses alarmes qui sont les mots-clés détectés à tord,
- le nombre de faux rejets qui sont les mots-clés présents mais non détectés.

La mesure du taux moyen d'erreur de mot-clé (*Word Error Rate*, WER) est peu utilisée pour l'évaluation de systèmes de détection de mots-clés. Dans le cas de la détection de mots-clés, il est nécessaire de pondérer l'importance de l'apparition de fausses alarmes par rapport à celle de faux rejets. En effet, les mots-clés à détecter apparaissent généralement peu souvent dans la grande quantité de signal de parole à analyser. Pour cette raison, le nombre de fausses alarmes peut être beaucoup plus grand que le nombre de faux rejets qui sont les détections potentielles rejetées à tord. Etant donnée une requête $q$ parmi $Q$ requêtes, le système de détection de mots-clés retourne alors :

- $R(q)$ le nombre d'occurrences de mots-clés répondant à la requête $q$ dans les transcriptions de références,
- $A(q)$ le nombre total d'éléments retournés lors de la requête $q$,
- $C(q)$ le nombre d'éléments correctement retournés lors de la requête $q$.

Dans ce cas, on peut alors exprimer les mesures de précision et de rappel telles que :

$$précision = \frac{1}{Q}\sum_{q=1}^{Q}\frac{C(q)}{A(q)} \qquad rappel = \frac{1}{Q}\sum_{q=1}^{Q}\frac{C(q)}{R(q)}$$

La F-mesure permet de combiner précision et rappel au sein d'une même mesure :

$$F\text{-}mesure = \frac{2 \times précision \times rappel}{précision + rappel}$$

Cette F-mesure est également appelée mesure $F_1$ car les mesures de précision et de rappel sont considérés avec le même coefficient pondérateur. Toutefois, il est possible de pondérer l'importance de la précision par rapport au rappel au sein d'une mesure commune $F_\beta$-mesure telle que :

$$F_\beta\text{-}mesure = \frac{(1+\beta) \times précision \times rappel}{\beta \times précision + rappel}, \forall \beta \in [1; +\infty]$$

Dans le cas d'une mesure de type rappel-précision, la mesure d'aire sous la courbe ROC (*Area Under Curve*, AUC) est plus informative que la F-mesure [Rohlicek et al., 1989]. Cette mesure consiste à exprimer le taux de bonne détection comme une fonction du taux de fausses alarmes. Elle est souvent accompagnée de la mesure de la figure de mérite (*Figure Of Merit*, FOM). Le FOM est défini comme la valeur moyenne du ratio du nombre de bonnes détections par rapport au nombre de fausses alarmes, dans l'intervalle de 0 à 10 fausses alarmes en moyenne par mot-clé et par heure (Figure N.1). Elle est donc usuelle pour évaluer la performance d'un système de détection de mots-clés en fonction de la gestion des paramètres variables internes au système.

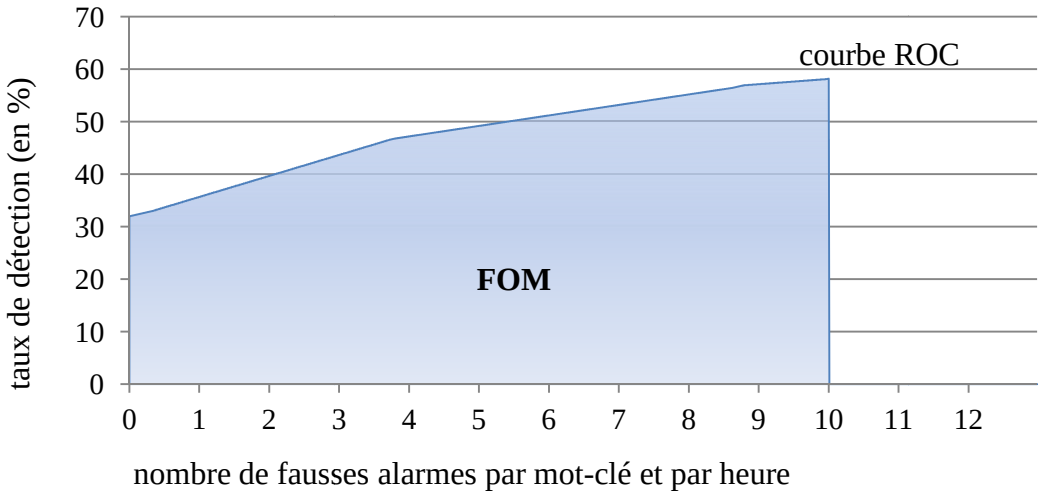



Figure N.1 : Exemple de représentation de la figure de mérite FOM (aire sous la courbe ROC)

Ainsi, considérant cette aire AUC sous la courbe ROC définie en fonction du taux de détection $T_d$ et du nombre moyen $s$ de fausses alarmes par mot-clé et par heure, la mesure de la figure de mérite FOM s'exprime comme :

$$FOM = \frac{1}{10}\int_{s=0}^{10} T_d \, d(s)$$

Par ailleurs, la campagne d'évaluation NIST STD définit deux autres mesures spécifiques à la détection de mots-clés sous les termes OCC (*OCCurrence-weighted value*) et ATWV (*Actual Term-Weighted Value*) [NIST STD, 2006]. Cette dernière mesure d'évaluation ATWV a deux caractéristiques importantes :

- une fausse alarme sur un mot-clé est plus fortement pénalisée qu'un faux rejet sur ce mot-clé,
- les résultats de détection sont moyennés sur l'ensemble des mots-clés recherchés plutôt que sur leur nombre d'apparitions.

Cette mesure ATWV considère alors la contribution de chaque terme de manière équivalente. En reprenant les résultats retournés par un système de détection de mots-clés sur une requête $q$ parmi $Q$ requêtes, on peut alors exprimer les taux de faux rejets $P_{faux\,rejet}$ et de fausses alarmes $P_{fausse\,alarme}$ comme :

$$P_{faux\,rejet}(q) = 1 - \frac{C(q)}{R(q)} \qquad\qquad P_{fausse\,alarme}(q) = \frac{A(q) - C(q)}{K - C(q)}$$

avec $K$ le nombre total de mots-clés présents dans les transcriptions de référence. En fonction de ces deux termes, la mesure ATWV s'exprime alors comme :

$$ATWV = 1 - \frac{1}{Q}\sum_{q=1}^{Q}\left(P_{faux\,rejet}(q) + \beta . P_{fausse\,alarme}(q)\right)$$

où $\beta$ est un paramètre utilisateur défini par défaut à la valeur 1000.

# Annexe O. Liste de mots-clés

Un ensemble de 80 mots-clés est choisi parmi les mots présents dans le sous-ensemble de test de la base de données TIMIT (Annexe B, page 135). Ces mots-clés sont choisis au hasard parmi la liste des mots de 4 phonèmes ou plus [Grangier et al., 2009]. Ces mots-clés sont :

absolute
admitted
aligning
anxiety
apartments
apparently
argued
bedrooms
brand
camera
characters
cleaning
climates
controlled
creeping
crossings
crushed
decaying
demands
depicts
dominant
dressy
drunk
efficient
episode
everything
excellent
experience
family
firing
followed
forgiveness
freedom
fulfillment
functional
grazing
henceforth
ignored
illnesses
imitate
increasing
inevitable
introduced
January
materials
millionaires
mutineer
needed
obvious
package
paramagnetic
patiently
pleasant
possessed
pressure
radiation
recriminations
redecorating
rejected
secularist
shampooed
solid
spilled
spreader
story
strained
streamlined
street
stripped
stupid
superb
surface
swimming
sympathetically
unenthusiastic
unlined
urethane
usual
walking
weekday

# Annexe P. Détection de mots-clés par approche discriminante

Dans un système de détection de mot-clé avec modèle poubelle, il est possible de maximiser le rapport de vraisemblance entre un mot-clé donné et les modèles poubelles associés à ce mot-clé [Rose et al., 1990]. Ce rapport de vraisemblance est obtenu tout en le minimisant sur un ensemble de fausses alarmes générées par le système lors d'une première détection des mots-clés effectuée au préalable [Sukkar et al., 1996]. D'autre part, il est également possible d'appliquer un paramètre d'erreur de classification minimale (*Minimum Classification Error*, MCE) au problème de détection de mots-clés [Juang et al., 1992]. Les modèles acoustiques peuvent alors être recalculés lors de l'apprentissage afin de minimiser le score de vraisemblance des modèles en dehors des mots-clés dans les zones du signal de parole où le mot-clé apparaît [Sandness et al., 2000]. Cependant, cette adaptation ne tient pas compte des cas de fausses alarmes. Cette adaptation ne permet donc pas de diminuer le score de vraisemblance des modèles issus des mots-clés dans les zones du signal de parole où le mot-clé n'apparaît pas. Dans tous ces cas, la notion de détection de mots-clés n'apparaît qu'une fois l'apprentissage des modèles d'unités linguistiques effectué. Les approches de systèmes de détection de mots-clés avec stratégie de rapport de vraisemblance ne permettent alors pas un apprentissage des modèles maximisant les performances de détection de mots-clés [Keshet et al., 2009]. Afin de répondre à cette contrainte, des approches par apprentissage de paramètres discriminants sont préférées [Grangier et al., 2009].

Certaines approches discriminantes sont développées autour de la recherche de combinaisons de différents systèmes de détection de mots-clés à base de HMM [Lin et al., 2009]. Par exemple, un réseau de neurones artificiels (*Artificial Neural Network*, ANN) peut servir durant l'apprentissage à combiner les rapports de vraisemblance à partir de différents modèles [Morgan et al., 1990; Cercadillo et al., 1993; Weintraub et al., 1997; Wöllmer et al.,

2009]. Un réseau de neurones artificiels est un modèle de calcul dont la conception est schématiquement adaptée du fonctionnement de vrais neurones. Les réseaux de neurones sont généralement optimisés par des méthodes d'apprentissage statistique [Bourlard et al., 1987]. Diverses approches existent dont les plus connues sont les perceptrons multi-couches [Bourlard et al., 1990; Mathan et al., 1991] et les TDNNs (*Time Delay Neural Network*) [Waibel et al., 1990; Zeppenfeld et al., 1992].

Par ailleurs, l'usage de machines à vecteurs supports (*Support Vector Machine*, SVM) est également possible afin de combiner différentes moyennes calculées pour les valeurs de vraisemblance au niveau phonétique [Ben Ayed et al., 2003]. L'approche par SVM se base sur la création de classes séparatrices afin de regrouper les séquences similaires. La modélisation par SVM permet alors de représenter une séquence par une approximation de la fonction d'évolution des éléments constituant cette séquence [Tavenard et al., 2007]. Dans le cas d'éléments à plusieurs dimensions, un SVM distinct modélise alors chacune de ces dimensions . Pour chaque SVM, un certain nombre de points caractéristiques sont conservés : ce sont les vecteurs supports [Ganapathiraju et al., 2004]. Dans ce cas, un ensemble de données d'apprentissage est utilisé pour déterminer les hyperplans séparateurs, générant des vecteurs supports représentatifs du groupe de données. Deux séquences sont alors considérées similaires si on peut prédire l'une grâce au modèle issu de l'autre. La mise place d'une mesure de similarité permet ensuite de s'assurer de la fiabilité de cette reconstruction. La complexité d'un tel SVM est alors faible, en $O(N)$ [He et al., 2008].

# Bibliographie

[Abdel-Haleem et al., 2004] Y.H. Abdel-Haleem, S. Renals, and N.D. Lawrence, "Acoustic space dimensionality selection and combination using the maximum entropy principle," *Proc. of IEEE ICASSP*, pp. 637-640, 2004.

[Achan et al., 2004] K. Achan, S. Roweis, A. Hertzmann, and B. Frey, "A segmental HMM for speech waveforms," University of Toronto, Toronto, Canada, Technical Report UTML, 2004-001, 2004.

[Adamczewski et al., 1977] H. Adamczewski and D. Keen, *Phonétique et phonologie de l'anglais contemporain*, A. Colin, Ed. Orléans, France, 1977.

[Adda-Decker et al., 2005] M. Adda-Decker, P. Boula de Mareuil, Adda G., and L. Lamel, "Investigating syllabic structures and their variation in spontaneous French," *Speech Communication*, no. 46, pp. 119-139, 2005.

[Ahmed et al., 1974] N. Ahmed, T. Natarajan, and K.R. Rao, "Discrete Cosine Transform," *IEEE Trans. on Computers*, pp. 90-93, 1974.

[Ahmed et al., 1985] W. Ahmed and G. Bekey, "A new algorithm for isolated word recognition with large within-class variability," *Proc. of IEEE ICASSP*, vol. 10, pp. 878-881, 1985.

[Allamanche et al., 2001] E. Allamanche, J. Herre, O. Hellmuth, B. Fröba, and M. Cremer, "Audioid: towards content-based identification of audio material," *Audio Engineering Society 110th convention*, 2001.

[Allan, 2002] J. Allan, "Perspectives on information retrieval and speech," *Proc. of Information Retrieval Techniques for Speech Applications*, pp. 323-326, 2002.

[Allauzen, 2003] A. Allauzen, "Modélisation linguistique pour l'indexation automatique de documents audiovisuels," LIMSI-CNRS, Orsay, France, Thèse de doctorat, 2003.

[André-Obrecht, 1988] R. André-Obrecht, "A new statistical approach for the automatic segmentation of continuous speech signals," *IEEE Trans. on ASSP*, vol. 1, no. 36, pp. 29-40, 1988.

[Ariyaeeinia et al., 1997] A.M. Ariyaeeinia and P. Sivakumaran, "Comparison of VQ and DTW classifiers for speaker verification," *European Conference on Security and Detection*, pp. 142-146, 1997.

[Aronowitz, 2010] H. Aronowitz, "Phoneme lattice construction and its application to speech recognition," Software Patent 7725319, 2010.

[Atal, 1983] B.S. Atal, "Efficient coding of LPC parameters by temporal decomposition," *Proc. of IEEE ICASSP*, pp. 81-84, 1983.

[Atal et al., 1974] B.S. Atal and M.R. Schroeder, "Recent advances in predictive coding - applications to speech synthesis," *Speech Communication*, vol. 1, pp. 27-31, 1974.

[Athineos et al., 2003] M. Athineos and D. Ellis, "Frequency domain linear prediction for temporal features," *Proc. of IEEE Workshop on ASRU*, pp. 261-266, 2003.

[Bahi et al., 2009] H. Bahi and N. Benati, "A new keyword spotting approach," *International Conference on Multimedia Computing and Systems*, pp. 77-80, 2009.

[Bahl et al., 1983] L.R. Bahl, F. Jelinek, and R.L. Mercer, "A maximum likelihood approach to continuous speech recognition," *IEEE Trans. on PAMI*, no. 5, pp. 179-190, 1983.

[Bahl et al., 1986] L. Bahl, P. Brown, P. De Souza, and R. Mercer, "Maximum mutual information estimation of Hidden Markov Model parameters for speech recognition," *Proc. of IEEE ICASSP*, pp. 49-52, 1986.

[Bahl et al., 1989] L.R. Bahl, R. Bakis, J. Bellegarda, P.F. Brown, D. Burshtein, S.K. Das, P.V. de Souza, P.S. Gopalakrishnan, F. Jelinek, D. Kanevsky, R.L. Mercer, A.J. Nadas, D. Nahamoo, and M.A. Picheny, "Large vocabulary natural language continuous speech recognition," *Proc. of IEEE ICASSP*, vol. 1, pp. 465-467, 1989.

[Baker et al., 1986] J. Baker and D. Pinto, "Optimal and suboptimal training strategies for automatic speech recognition in noise, and the effects of adaptation on performance," *Proc. of IEEE ICASSP*, no. 11, pp. 745-748, 1986.

[Bakis, 1976] R. Bakis, "Continuous speech recognition via centisecond acoustic states," *Proc. of JASA*, vol. 59, no. 1, p. 97, 1976.

[Balado et al., 2007] F. Balado, N.J. Hurley, E.P. McCarthy, and G.C.M. Silvestre, "Performance analysis of robust audio hashing," *IEEE Trans. on Information Forensics and Security*, vol. 2, no. 2, pp. 254-266, 2007.

[Baluja et al., 2008] S. Baluja and M. Covell, "Waveprint: Efficient wavelet-based audio fingerprinting," *Pattern Recognition*, vol. 11, no. 41, pp. 3467-3480, 2008.

[Bellettini et al., 2010] C. Bellettini and G. Mazzini, "A framework for robust audio fingerprinting," *Journal of Communications*, vol. 5, no. 5, pp. 409-424, 2010.

[Bellman, 1957] R. Bellman, *Dynamic Programming*.: Princeton University Press, 1957.

[Ben Ayed, 2003] Y. Ben Ayed, "Détection de mots-clés dans un flux de parole," ENST, Nancy, France, Thèse de doctorat, 2003.

[Ben Ayed et al., 2003] Y. Ben Ayed, D. Fohr, J.P. Haton, and G. Chollet, "Confidence measures for keyword spotting using support vector machines," *Proc. of IEEE ICASSP*, vol. 1, pp. 588-591, 2003.

[Benitez et al., 2001] C. Benitez, L. Burget, B. Chen, S. Dupont, H. Garugadri, H. Hermansky, P. Jain, S. Kajarekar, and S. Sivadas, "Robust ASR front-end using spectral based and discriminant features: experiments on the Aurora task," *Proc. of Eurospeech*, pp. 429-432, 2001.

[Benzeghiba et al., 2006] M. Benzeghiba, R. De Mori, O. Deroo, S. Dupont, D. Jouvet, L. Fissore, P. Laface, A. Mertins, C. Ris, R. Rose, V. Tyagi, and C.J. Wellekens, "Impact of variabilities on speech recognition," *Proc. of SPECOM*, 2006.

[Benzeghiba et al., 2007] M. Benzeghiba, R. De Mori, O. Deroo, S. Dupont, T. Erbes, D. Jouvet, L. Fissore, P. Laface, A. Mertins, C. Ris, R. Rose, V. Tyagi, and C. Wellekens, "Automatic speech recognition and speech variability: A review," *Speech Communication*, pp. 763-786, 2007.

[Benzeguiba et al., 2006b] M. Benzeguiba, R. De Mori, O. Deroo, S. Dupont, T. Erbes, D. Jouvet, L. Fissore, P. Laface, A. Mertins, C. Ris, R. Rose, V. Tyagi, and C. Wellekens, "Automatic speech recognition and intrinsic speech variation," *Proc. of IEEE ICASSP*, vol. 5, pp. 1021-1024, 2006b.

[Besacier et al., 2001] L. Besacier, C. Bergamini, D. Vaufreydaz, and E. Castelli, "The effect of speech and audio compression on speech recognition performance," *Proc. on IEEE Multimedia Signal Processing Workshop*, pp. 301-306, 2001.

[Betser, 2008] M. Betser, "Modélisation sinusoïdale et applications à l'indexation audio," Télécom ParisTech, Paris, France, Thèse de doctorat, 2008.

[Betser et al., 2008] M. Betser, P. Collen, G. Richard, and B. David, "Estimation of frequency for AM/FM models using the phase vocoder framework," *IEEE Trans. on Signal Processing*, vol. 5, no. 56, pp. 505-517, 2008.

[Bhattacharyya, 1943] A. Bhattacharyya, "On a measure of divergence between two statistical populations defined by their probability distributions," *Bulletin of the Calcutta Mathematical Society*, vol. 35, pp. 99-110, 1943.

[Blackman et al., 1959] R.B. Blackman and J. Tukey, "Particular pairs of windows," *The Measurement of Power Spectra from The Point of View of Communications Engineering*, pp. 98-99, 1959.

[Blanche-Benveniste, 1999] C. Blanche-Benveniste, "Constitution et exploitation d'un grand corpus," *Revue française de linguistique appliquée*, vol. 4, no. 1, pp. 65-74, 1999.

[Blumstein, 1986] S.E. Blumstein, "On acoustic invariance in speech," in *Invariance and variability in speech processes*, Perkell and Klatt, Ed.: MIT Press, 1986.

[Bogert et al., 1963] B. Bogert, M. Healy, and J. Tukey, "The quefrency alanysis of time series for echoes: cepstrum, pseudo-autocovariance, cross-cepstrum and saphe cracking," *Time Series Analysis*, pp. 209-243, 1963.

[Bohac et al., 2011] M. Bohac and K. Blavka, "Automatic segmentation and annotation of audio archive documents," *International Workshop on ECMS*, pp. 1-6, 2011.

[Boite et al., 1993] J.M. Boite, H. Bourlard, B. D'hoore, and M. Haesen, "A new approach towards keyword spotting," *Proc. of Eurospeech*, pp. 1273-1276, 1993.

[Boite et al., 2000] R. Boite, H. Bourlard, T. Dutoit, J. Hancq, and Leich H., *Traitement de la parole*. Lausanne, Suisses: Presses Polytechniques et Universitaire Romandes, 2000.

[Bolt et al., 1970] R.H. Bolt, F.S. Cooper, Jr David E.E., P.B. Denes, J.M. Pickett, and K.N. Stevens, "Speaker identification by speech spectrograms: a scientist's view of its reliability for legal purposes," *Proc. of JASA*, vol. 47, no. 2, pp. 597-612, 1970.

[Boney et al., 1996] L. Boney, A.H. Teufik, and K.N. Hamdy, "Digital watermarks for audio signals," *Proc. of IEEE ICMCS*, pp. 473-480, 1996.

[Boula de Mareuil et al., 2002] P. Boula de Mareuil and M. Adda-Decker, "Studying pronunciation variants in French by using alignment techniques," *Proc. of Interspeech*, pp. 2273-2276, 2002.

[Bourlard et al., 1985] H. Bourlard, Y. Kamp, and C. Wellekens, "Speaker dependent connected speech recognition via phonetic Markov models," *Proc. of IEEE ICASSP*, no. 10, pp. 1213-1216, 1985.

[Bourlard et al., 1987] H. Bourlard and C.J. Wellekens, "Multilayer perceptrons and automatic speech recognition," *Proc. of International Conference on Neural Networks*, pp. 407-416, 1987.

[Bourlard et al., 1990] H. Bourlard and C.J. Wellekens, "Links between Markov Models and Multilayer Perceptrons," *Proc. of IEEE PAMI*, vol. 12, no. 12, pp. 1167-1178, 1990.

[Bourlard et al., 1994] H. Bourlard, B. D'hoore, and J.M. Boite, "Optimizing recognition and rejection performance in wordspotting systems," *Proc. of IEEE ICASSP*, vol. 1, pp. 373-376, 1994.

[Bourlard et al., 1997] H. Bourlard and D. Dupont, "Sub-band based speech recognition," *Proc. of IEEE ICASSP*, pp. 1251-1254, 1997.

[Bozkurt et al., 2005] B. Bozkurt and L. Couvreur, "On the use of phase information for speech recognition," *Proc. of Eusipco*, 2005.

[Bridle, 1973] J.S. Bridle, "An efficient elastic template method for detecing given words in running speech," *Proc. of British Acoustical Society Meeting*, vol. 73SHC3, 1973.

[Brown et al., 1995] K.L. Brown and E.B. George, "CTIMIT: A speech corpus for the cellular environment with applications to automatic speech recognition," *Proc. of IEEE ICASSP*, pp. 105-108, 1995.

[Brück et al., 2004] J.M. Brück, S. Bres, and D. Pellerin, "Construction d'une signature audio pour l'indexation de documents audiovisuels," *Proc. of CORESA*, 2004.

[Burges et al., 2003] C.J.C Burges, J.C Platt, and S. Jana, "Distortion discriminant analysis for audio fingerprinting," *Proc. of IEEE Speech and Audio Processing*, vol. 11, no. 3, pp. 165-174, 2003.

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

[Cano, 2007] P. Cano, "Content-based audio search: from fingerprinting to semantic audio retrieval," Pompeu Fabra University, Barcelona, España, PhD thesis, 2007.

[Cano et al., 2002] P. Cano, E. Battle, T. Kalker, and J. Haitsma, "A review of algorithms for audio fingerprinting," *IEEE Workshop on MSP*, pp. 169-173, 2002.

[Cano et al., 2005] P. Cano, E. Battle, E. Gaomez, L. C.T. Gomes, and M. Bonnet, "Audiofingerprinting: concepts and applications," *Studies in Computational Intelligence*, vol. 2, pp. 233-245, 2005.

[Caraty et al., 1997] M.J. Caraty, C. Montacié, and F. Lefèvre, "Dynamic lexicon for a very large vocabulary vocal dictation," *Proc. of Eurospeech*, pp. 2691-2694, 1997.

[Caraty et al., 1998] M.J. Caraty and C. Montacié, "Multi-resolution for speech analysis," *Proc. of ICSLP*, pp. 1142-1144, 1998.

[Caraty et al., 2010] M.J. Caraty and C. Montacié, "Multivariate analysis of vocal fatigue in continuous reading ," *Proc. of Interspeech*, pp. 470-473, 2010.

[Carney, 1994] E. Carney, *A survey of English spelling*. London, United Kingdom: Routledge, 1994.

[Cercadillo et al., 1993] J.A. Cercadillo and A.H. Gomez, "Grammar learning and word spotting using recurrent neural networks," *Proc. of Eurospeech*, pp. 21-23, 1993.

[Chang et al., 1993] P.C. Chang and B.H. Juang, "Discriminative training of dynamic programming based speech recognizers," *IEEE Trans. on SAP*, vol. 1, pp. 135-143, 1993.

[Chatterjee et al., 2011] S. Chatterjee and W.B. Kleijn, "Auditory Model-Based Design and Optimization of Feature Vectors for Automatic Speech Recognition," *IEEE Trans. on ASLP*, vol. 6, no. 19, pp. 1813-1825, 2011.

[Chen, 1986] F. Chen, "Lexical access and verification in a broad phonetic approach to continuous digit recognition," *Proc. of IEEE ICASSP*, no. 11, pp. 1089-1092, 1986.

[Chen, 1987] Y. Chen, "Cepstral domain stress compensation for robust speech recognition," *Proc. of IEEE ICASSP*, pp. 717-720, 1987.

[Chen, 1988] C.H. Chen, *Signal processing handbook*. New York: Dekker, 1988, p. 531.

[Chen et al., 2001] J. Chen, K.K. Paliwal, and Nakamura S., "Sub-band based additive noise removal for robust speech recognition," *Proc. of Eurospeech*, pp. 571-574, 2001.

[Chigier, 1992] B. Chigier, "Rejection and keyword spotting algorithms for a directory assistance city name recognition application," *Proc. of IEEE ICASSP*, vol. 2, pp. 93-96, 1992.

[Chiu et al., 2010] C.Y. Chiu, D. Bountouridis, J.C. Wang, and H.M. Wang, "Background music identification through content filtering and min-hash matching," *Proc. of IEEE ICASSP*, pp. 2414-2417, 2010.

[Christiansen et al., 1976] R.W. Christiansen and C.K. Rushforth, "Word spotting in continuous speech using Linear Predictive Coding," *Proc. of IEEE ICASSP*, 1976.

[Christiansen et al., 1977] R. Christiansen and C. Rushforth, "Detecting and locating key words in continuous speech using linear predictive coding," *IEEE Trans. on ASSP*, vol. 5, no. 25, pp. 361-367, 1977.

[Claes et al., 1996] T. Claes and D. Van Compernolle, "SNR-normalisation for robust speech recognition," *Proc. of IEEE ICASSP*, vol. 1, pp. 331-334, 1996.

[Class et al., 1990] F. Class, A. Kaltenmeier, P. Regel, and K. Trottler, "Fast speaker adaptation for speech recognition systems," *Proc. of IEEE ICASSP*, vol. 1, pp. 133-136, 1990.

[CLEF, 2010] CLEF. (2010) Cross-Language Evaluation Forum. [Online]. http://www.clef-campaign.org

[Coifman et al., 1992] R.R. Coifman and M.V. Wickerhauser, "Entropy based algorithms for best basis selection," *IEEE Trans. on Information Theory*, vol. 2, no. 38, pp. 713-718, 1992.

[Collen et al., 2007] P. Collen, J.B. Rault, and M. Betser, "Phase estimating method for a digital signal sinusoidal simulation," Software Patent PCT/FR2006/051361, 2007.

[Collet et al., 2005] M. Collet, D. Charlet, and F. Bimbot, "A correlation metric for speaker tracking using anchor models," *Proc. of IEEE ICASSP*, vol. 1, pp. 713-716, 2005.

[Cook et al., 2006] R. Cook and M. Cremer, "A tunable, efficient, specialized multidimensional range query algorithm ," *IEEE International Symposium on Signal Processing and Information Technology*, pp. 397-402, 2006.

[Cooley et al., 1965] J.W. Cooley and J.W. Tukey, "An algorithm for the machine calculation of complex Fourier series," *Mathematics of Computation*, vol. 19, no. 90, pp. 297-301, 1965.

[Cormen et al., 2001] T.H. Cormen, C.E. Leiserson, R.L. Rivest, and C. Stein, *Introduction to Algorithms*.: MIT Press, 2001.

[Cotton et al., 2010] C.V. Cotton and D.P.W. Ellis, "Audio fingerprinting to identify multiple videos of an event," *Proc. of IEEE ICASSP*, pp. 2386-2389, 2010.

[Cox et al., 1996] I.J. Cox, J. Kilian, T. Leighton, and T. Shamoon, "A secure, robust watermark for multimedia," *Proc. of First International Workshop on Information Hiding*, pp. 185-206, 1996.

[Crouzet, 2000] O. Crouzet, "Segmentation de la parole en mots et régularités phonotactiques : Effets phonologiques, probabilistes ou lexicaux ?," Laboratoire de linguistique de Nantes, Nantes, Thèse de doctorat, 2000.

[CSTR, 2001] CSTR. (2001) Centre for Speech Technology Research, Carnegie Mellon University. [Online]. http://festvox.org/dbs/dbs_kdt.html

[Culling et al., 1994] J. Culling and Q., Marshall, D. Summerfield, "Effects of simulated reverberation on the use of binaural cues and fundamental frequency differences for separating concurrent vowels," *Speech Communication*, no. 14, pp. 71-95, 1994.

[Dallos, 1973] P. Dallos, *The Auditory Periphery: Biophysics and Physiology*. New York, USA: Academic Press, 1973.

[Das et al., 1998] S. Das, D. Nix, and M. Picheny, "Improvements in children speech recognition performance," *Proc. of IEEE ICASSP*, vol. 1, pp. 433-436, 1998.

[Davis et al., 1980] S. Davis and P. Melmerstein, "Comparison of parametric representation for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. on ASSP*, vol. 28, pp. 357-366, 1980.

[Deléglise et al., 1988] P. Deléglise, F. Bimbot, C. Montacié, and G. Chollet, "Temporal decomposition and acoustic-phonetic decoding for the automatic recognition of continuous speech," *Proc. of ICPR*, vol. 2, pp. 839-841, 1988.

[Deléglise et al., 2005] P. Deléglise, Y. Estève, S. Meignier, and T. Merlin, "The LIUM speech transcription system: a CMU Sphinx III-based system for French broadcast news," *Proc. of Interspeech*, pp. 1653-1656, 2005.

[Dempster et al., 1977] A.P. Dempster, N.M. Laird, and D.B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society, Series B*, no. 39, pp. 1-38, 1977.

[Demuynck et al., 2004] K. Demuynck, O. Garcia, and D. Van Compernolle, "Synthesizing speech from speech recognition parameters," *Proc. of ICSLP*, 2004.

[Demuynck et al., 2011] K. Demuynck, D. Seppi, H. Van hamme, and D. Van Compernolle, "Progress in example based automatic speech recognition," *Proc. of IEEE ICASSP*, pp. 4692-4695, 2011.

[Denbigh et al., 1994] P.N. Denbigh and H.Y. Luo, "An algorithm for separating overlapping voices," *IEE Colloquium on Techniques for Speech Processing and their Application*, vol. 9, pp. 1-6, 1994.

[Deviren, 2004] M. Deviren, "Systèmes de reconnaissance de la parole revisités : Réseaux Bayésiens dynamiques et nouveaux paradigmes," Université de Nancy, Nancy, Thèse de doctorat, 2004.

[Di Benedetto et al., 1992] M.G. Di Benedetto and J.S. Liénard, "Extrinsic normalization of vowel formant values based on cardinal vowels mapping," *Proc. of ICSLP*, pp. 579-582, 1992.

[Diller, 1979] T. Diller, "Phonetic word verification," *Proc. of IEEE ICASSP*, vol. 4, pp. 256-261, 1979.

[Dimitriadis et al., 2005] D.V. Dimitriadis, P. Maragos, and A. Potamianos, "Robust AM-FM Features for Speech Recognition," *IEEE Signal Processing Letters*, vol. 9, no. 12, pp. 621-624, 2005.

[Divoux et al., 1990] P. Divoux, J. Di Martino, A. Boyer, J.F. Mari, K. Smaili, and J.P. Haton, "Statistical methods in multi-speaker automatic speech recognition," *Applied stochastic models and data analysis*, vol. 3, no. 6, pp. 143-155, 1990.

[Djouadi et al., 1990] A. Djouadi, O. Snorrason, and F. Garber, "The quality of Training-Sample estimates of the Bhattacharyya coefficient," *Proc. of IEEE Trans. on PAMI*, vol. 12, no. 1, pp. 92-97, 1990.

[Doddington, 1985] G. Doddington, "Speaker recognition - Identifying people by their voices," *Proc. of IEEE*, vol. 73, no. 11, pp. 1651-1664, 1985.

[Duda et al., 1973] R.O. Duda and P.E. Hart, *Pattern classification and scene analysis*. New York: Wiley, 1973.

[Duez, 2003] D. Duez, "Modelling aspects of reduction and assimilation in spontaneous French speech," *Proc. of IEEE-ISCA Workshop on Spontaneous Speech Processing and Recognition*, pp. 120-124, 2003.

[Ehlers et al., 1997] F. Ehlers and H.G. Schuster, "Blind separation of convolutive mixtures and an application in automatic speech recognition in a noisy environment," *IEEE Trans. on Signal Processing*, pp. 2608-2612, 1997.

[Eide, 2001] E. Eide, "Distinctive features for use in automatic speech recognition," *Proc. of Eurospeech*, pp. 1613-1616, 2001.

[Eide et al., 1996] E. Eide and H. Gish, "A parametric approach to vocal tract length normalization," *Proc. of IEEE ICASSP*, pp. 346-349, 1996.

[El Meliani et al., 1998] R. El Meliani and D. O'Shaughnessy, "Specific language modelling for new-word detection in continuous-speech recognition," *Proc. of IEEE ICASSP*, vol. 1, pp. 321-324, 1998.

[Ellis, 2009] D.P.W. Ellis. (2009) Robust Landmark-Based Audio Fingerprinting. [Online]. http://labrosa.ee.columbia.edu/matlab/fingerprint

[Ellis et al., 2001] D.P.W. Ellis, R. Singh, and S. Sivadas, "Tandem acoustic modeling in large-vocabulary recognition," *Proc. of IEEE ICASSP*, pp. 517-520, 2001.

[Elman et al., 1986] J. Elman and J. Mc Clelland, "Exploiting lawful variability in the speech wave," in *Invariance and variability in speech processes*, Perkel and Klatt, Ed.: MIT Press, 1986.

[Enochson et al., 1968] L.D. Enochson and R.K. Otnes, *Programming and Analysis for Digital Time Series Data*. USA: U.S. Dept. of Defense, Shock & Vibration Information Analysis Center, 1968, p. 142.

[Ephraim et al., 1989] Y. Ephraim, D. Malah, and B.H. Juang, "On the application of hidden Markov models for enhancing noisy speech," *IEEE Trans. on ASSP*, vol. 12, no. 37, pp. 1846-1856, 1989.

[Eyben et al., 2010] F. Eyben, M. Wöllmer, and B. Schuller, "openSMILE - The Munich Versatile and Fast Open-Source Audio Feature Extractor," *Proc. of ACM Multimedia*, pp. 1459-1462, 2010.

[Fant, 1960] G. Fant, *Acoustic theory of speech production*. The Hague: Mouton, 1960.

[Farnetani, 1997] E. Farnetani, "Coarticulation and connected speech," in *The Handbook of Phonetic Sciences*, Blackwell, Ed., 1997.

[Faundez-Zanuy et al., 2005] M. Faundez-Zanuy and E. Monte-Moreno, "State-of-the-art in speaker recognition," *IEEE Aerospace and Electronic Systems Magazine*, vol. 20, no. 5, pp. 7-12, 2005.

[Favero, 1994] R.F. Favero, "Compound wavelets: wavelets for speech recognition," *Proc. of IEEE-SP Int. Symp. on Time-Frequency and Time-Scale Analysis*, pp. 600-603, 1994.

[Fetter et al., 1996] P. Fetter, A. Kaltenmeier, T. Kuhn, and P. Regel-Brietzmann, "Improved modeling of OOV words in spontaneous speech," *Proc. of IEEE ICASSP*, vol. 1, pp. 534-537, 1996.

[Fineberg et al., 1996] A.B. Fineberg and K.C. Yu, "Time-frequency representation based cepstral processing for speech recognition," *Proc. of IEEE ICASSP*, vol. 1, pp. 25-28, 1996.

[Fischer et al., 1986] W. Fischer, G. Doddington, and K. Goudie-Marshall, "The DARPA speech recognition research database: specifications and status," *DARPA Workshop on Speech Recognition*, pp. 93-99, 1986.

[Fletcher et al., 1933] H. Fletcher and W.A. Munson, "Loudness, its definition, measurement and calculation," *Proc. of JASA*, no. 5, pp. 82-108, 1933.

[Fontaine et al., 1997] V. Fontaine, C. Ris, and J.M. Boite, "Nonlinear Discriminant Analysis for Improved Speech Recognition," *Proc. of Eurospeech*, pp. 2071-2075, 1997.

[Foote et al., 1995] J.T. Foote, G.J.F. Jones, K. Spärck Jones, and S.J. Young, "Talker-independent keyword spotting for information retrieval," *Proc. of Eurospeech*, pp. 2145-2148, 1995.

[Fosler-Lussier et al., 1999] E. Fosler-Lussier and N. Morgan, "Effects of speaking rate and word predictability on conversational pronunciations," *Speech Communication*, vol. 2-4, no. 29, pp. 137-158, 1999.

[Fragoulis et al., 2001] D. Fragoulis, G. Rousopoulos, T. Panagopoulos, C. Alexiou, and C. Papaodysseus, "On the automated recognition of seriously distorted musical recordings," *Proc. of IEEE Transactions on Signal Processing*, vol. 49, no. 4, pp. 898-908, 2001.

[Fukunaga, 1972] K. Fukunaga, *Introduction to statistical pattern recognition*. New York: Academic Press, 1972.

[Gales, 1999] M.J.F. Gales, "Semi-tied covariance matrices for Hidden Markov Models," *IEEE Trans. on SAP*, no. 7, pp. 272-281, 1999.

[Galliano et al., 2005] S. Galliano, E. Geoffrois, D. Mostefa, K. Choukri, J.F. Bonastre, and G. Gravier, "The ESTER phase II evaluation campaign for the rich transcription of French broadcast news," *Proc. of the European Conf. on Speech Communication and Technology*, 2005.

[Galliano et al., 2009] S. Galliano, G. Gravier, and L. Chaubard, "The ESTER 2 evaluation campaign for the rich transcription of French radio broadcasts," *Proc. of Interspeech*, pp. 2583-2586, 2009.

[Ganapathiraju et al., 2004] A. Ganapathiraju, J.E. Hamaker, and J. Picone, "Applications of support vector machines to speech recognition," *IEEE Trans. on Signal Processing*, vol. 52, no. 8, pp. 2348-2355, 2004.

[Garofolo et al., 2000] J.S. Garofolo, C.G.P. Auzanne, and E.M. Voorhees, "The TREC Spoken Document Retrieval track: A success story," *Proc. of RIAO*, pp. 1-20, 2000.

[Garvin et al., 1963] P.L. Garvin and P. Ladefoged, "Speaker identification and message identification in speech recognition," *Phonetica*, no. 9, pp. 193-199, 1963.

[Gas et al., 2000] B. Gas, J.L. Zarader, and C. Chavy, "A new approach to speech coding: the Neural Predictive Coding ," *Journal of Advanced Computational Intelligence and Intelligent Informatics*, vol. 4, no. 1, 120-127 2000.

[Gauvain et al., 1990] J.L. Gauvain, L.F. Lamel, and M. Eskenazi, "Design considerations and text selection for BREF, a large French read-speech corpus," *Proc. of ICSLP*, pp. 1097-1100, 1990.

[Gelin, 1997] P. Gelin, "Détection de mots-clés dans un flux de parole : application à l'indexation de documents multimédia," EPFL, Lausanne, Suisses, Thèse de doctorat, 1997.

[Gemello et al., 2006] R. Gemello, F. Mana, D. Albesano, and R. De Mori, "Multiple resolution analysis for automatic robust speech recognition," *Computer, Speech and Language*, no. 20, pp. 2-21, 2006.

[Gish et al., 2009] H. Gish, M.H. Siu, A. Chan, and B. Belfield, "Unsupervised training of an HMM-based speech recognizer for topic classification ," *Proc. of Interspeech*, pp. 1935-1938, 2009.

[Gomes et al., 2003] L. Gomes, P. Cano, E. Goacutemez, M. Bonnet, and E. Batlle, "Audio watermarking and fingerprinting: for which applications? ," *Journal of New Music Research*, vol. 32, no. 1, pp. 65-81, 2003.

[Gopinath, 1998] R.A. Gopinath, "Maximum likelihood modeling with Gaussian distributions for classification," *Proc. of IEEE ICASSP*, pp. 661-664, 1998.

[Gorin et al., 1997] A.L. Gorin and G. Wright, J.H. Riccardi, "How may I help you?," *Speech Communications*, no. 23, pp. 113-127, 1997.

[Gracianera et al., 2004] M. Gracianera, H. France, J. Zheng, D. Vergyri, and A. Stolcke, "Voicing feature integration in SRI's DECIPHER LVCSR system," *Proc. of IEEE ICASSP*, pp. 921-924, 2004.

[Grangier et al., 2009] D. Grangier, J. Keshet, and S. Bengio, "Discriminative keyword spotting," in *Automatic speech and speaker recognition: large margin and kernel methods*, J. Keshet and S. Bengio, Eds. Chichester, UK: John Wiley & Sons, Ltd, 2009, ch. 11.

[Gravier et al., 2004] G. Gravier, J.F. Bonastre, S. Galliano, E. Geoffrois, K. Mc Tait, and K. Choukri, "The ESTER evaluation campaign of rich transcription of French broadcast news," *Proc. of Language Evaluation and Resources Conference*, 2004.

[Grenier et al., 1981] Y. Grenier, K. Bry, J. Le Roux, and M. Sulpis, "Autoregressive models for noisy speech signals," *Proc. of IEEE ICASSP*, pp. 1093-1096, 1981.

[Grézl et al., 2007] F. Grézl and J. Cernocky, "TRAP-based techniques for recognition of noisy speech," *Proc. of TSD*, pp. 270-277, 2007.

[Haeb-Umbach et al., 1992] R. Haeb-Umbach and H. Ney, "Linear discriminant analysis for improved large vocabulary continuous speech recognition," *Proc. of IEEE ICASSP*, pp. 13-16, 1992.

[Haeb-Umbach et al., 1993] R. Haeb-Umbach, D. Geller, and H. Ney, "Improvements in connected digit recognition using linear discriminant analysis and mixture densities," *Proc. of IEEE ICASSP*, vol. 2, pp. 239-242, 1993.

[Haitsma et al., 2001] J. Haitsma, T. Kalker, and J. Oostveen, "Robust audio hashing for content identification," *Workshop on CBMI*, 2001.

[Haitsma et al., 2002] J. Haitsma and T. Kalker, "A highly robust audio fingerprinting system," *Proc. of ISMIR*, pp. 107-115, 2002.

[Haitsma et al., 2003] J. Haitsma and T. Kalker, "Speed-change resistant audio fingerprinting using auto-correlation," *Proc. of IEEE ICASSP*, vol. 4, pp. 728-731, 2003.

[Hansen, 1996] J.H.L. Hansen, "Analysis and compensation of speech under stress and noise for environmental robustness in speech recognition," *Speech Communications*, vol. 2, no. 20, pp. 151-170, 1996.

[Hansen et al., 2001] J.H.L. Hansen, R. Sarikaya, U. Yapanel, and B.L. Pellom, "Robust speech recognition in noise: an evaluation using SPINE corpus," *Proc. of Eurospeech*, pp. 905-911, 2001.

[Hanson et al., 1990] B.A. Hanson and T. Applebaum, "Robust speaker-independent word recognition using instantaneous, dynamic and acceleration features: experiments with Lombard and noisy speech," *Proc. of IEEE ICASSP*, pp. 857-860, 1990.

[Hardcastle et al., 2010] W.J. Hardcastle, J. Laver, and Gibbon F.E., *The handbook of phonetic sciences*, 2nd ed.: Blackwell Publishing, 2010.

[Hariharan et al., 2001] R. Hariharan, I. Kiss, and O. Viikki, "Noise robust speech parameterization using multiresolution feature extraction," *IEEE Trans. on SAP*, vol. 8, no. 9, pp. 856-865, 2001.

[Harmegnies et al., 1988] B. Harmegnies and A. Landercy, "Intra-speaker variability of the long term speech spectrum," *Speech Communication*, vol. 1, no. 7, pp. 81-86, 1988.

[Harris et al., 1990] R.W. Harris and D.W. Swenson, "Effects of reverberation and noise on speech recognition by adults with various amounts of sensorineural hearing impairment," *International Journal of Audiology*, vol. 6, no. 29, pp. 314-321, 1990.

[Haton et al., 1976] J.P. Haton and J.M. Pierrel, "Organization and operation of a connected speech understanding system at lexical, syntactic and semantic levels," *Proc. of IEEE ICASSP*, vol. 1, pp. 430-433, 1976.

[Haton et al., 1991] J.P. Haton, J. Caelen, J.L. Gauvain, G. Perennou, and J.M. Pierrel, *Reconnaissance automatique de la parole*.: Dunod, 1991.

[Haykin, 1993] S. Haykin, *Adaptive filter theory*, NJ, Ed. USA: Prentice-Hall Publishers, 1993.

[Haykin, 1994] S. Haykin, *Communication systems*, 3rd ed., John Wiley and Sons, Ed. New York, USA, 1994.

[Hazen et al., 2008] T.J. Hazen and A. Margolis, "Discriminative feature weighting using MCE training for topic identification of spoken audio recordings," *Proc. of IEEE ICASSP*, pp. 4965-4968, 2008.

[He et al., 2008] X. He and L. Deng, *Discriminative learning for speech recognition: theory and practice*.: Morgan and Claypool Publishers, 2008.

[Hedge et al., 2004] R.M. Hedge, H.A. Murthy, and V.R.R. Gadde, "Continuous speech recognition using joint features derived from the modified group delay function and MFCC," *Proc. of ICSLP*, pp. 905-908, 2004.

[Hedge et al., 2005] R.M. Hedge, H.A. Murthy, and G.V.R. Rao, "Speech processing using joint features derived from the modified group delay function," *Proc. of IEEE ICASSP*, vol. 1, pp. 541-544, 2005.

[Hellwarth et al., 1968] G. Hellwarth and G. Jones, "Automatic conditioning of speech signals," *IEEE Trans. on Audio and Electroacoustics*, vol. 2, no. 16, pp. 169-179, 1968.

[Hermansky, 1990] H. Hermansky, "Perceptual Linear Predictive (PLP) analysis of speech," *Journal of Acoustic Society of America*, vol. 87, no. 4, pp. 1738-1752, 1990.

[Hermansky et al., 1991] H. Hermansky, N. Morgan, A. Bayya, and P. Kohn, "Compensation for the effect of the communication channel in auditory-like analysis of speech (RASTA-PLP)," *Proc. of ECSCT*, pp. 1367-1370, 1991.

[Hermansky et al., 1993] H. Hermansky, N. Morgan, and H.G. Hirsch, "Recognition of speech in additive and convolutional noise based on RASTA spectral processing," *Proc. of IEEE ICASSP*, vol. 2, pp. 83-86, 1993.

[Hermansky et al., 1994] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Trans. on SAP*, pp. 578-589, 1994.

[Hermansky et al., 1998] H. Hermansky and S. Sharma, "TRAPS - Classifiers of Temporal Patterns," *Proc. of ICSLP*, pp. 1003-1006, 1998.

[Hermansky et al., 1999] H. Hermansky and S. Sharma, "Temporal Patterns (TRAPS) in ASR of noisy speech," *Proc. of IEEE ICASSP*, pp. 289-292, 1999.

[Herre, 2004] J. Herre, "Method and device for producing a fingerprint and method and device for identifying an audio signal," Brevet, USA 0172411, 2004.

[Herre et al., 2002] J. Herre, O. Hellmuth, and M. Cremer, "Scalable robust audio fingerprinting using MPEG-7 content description," *IEEE Workshop on MSP*, pp. 165-168, 2002.

[Higgins et al., 1985] A. Higgins and R. Wohlford, "Keyword recognition using template concatenation," *Proc. of IEEE ICASSP*, pp. 1233-1236, 1985.

[Hirschberg et al., 2004] J. Hirschberg, D. Litman, and M. Swerts, "Prosodic and other cues to speech recognition failures," *Speech Communication*, vol. 1-2, no. 43, pp. 155-175, 2004.

[Hofstetter et al., 1992] E.M. Hofstetter and R.C. Rose, "Techniques for task independent word spotting in continuous speech messages ," *Proc. of IEEE ICASSP*, vol. 2, pp. 101-104, 1992.

[Hofstetter et al., 1992] E.M. Hofstetter and R.C. Rose, "Techniques for task independent word spotting in continuous speech messages," *Proc. of IEEE ICASSP*, vol. 2, pp. 101-104, 1992.

[Holden et al., 1976] A. Holden and E. Strasbourger, "A computer programming system using continuous speech input," *IEEE Trans. on ASSP*, vol. 6, no. 24, pp. 579-582, 1976.

[Holmes et al., 1986] J. Holmes and N. Sedgwick, "Noise compensation for speech recognition using probabilistic models," *Proc. of IEEE ICASSP*, vol. 11, pp. 741-744, 1986.

[Hon et al., 1999] H.W. Hon and K. Wang, "Combining frame and segment based models for large vocabulary continuous speech recognition," *Proc. of IEEE Workshop on ASRU*, 1999.

[Hong et al., 2000] H. Hong, S. Choi, H. Glotin, and F. Berthommier, "Blind acoustic source separation for cocktail party speech recognition," *Proc. of IEEE ICONIP*, 2000.

[Howard et al., 1988] I.S. Howard and M.A Huckvale, "Acoustic-phonetic attribute determination using multi-layer perceptrons," *IEE Colloquium on Speech Processing*, vol. 4, pp. 1-4, 1988.

[Huang et al., 1991] X. Huang and K. Lee, "On speaker-independent, speaker-dependent, and speaker adaptative speech recognition," *Proc. of IEEE ICASSP*, pp. 877-880, 1991.

[Huang et al., 2001] C. Huang, T. Chen, S. Li, E. Chang, and J. Zhou, "Analysis of speaker variability," *Proc. of Eurospeech*, pp. 1377-1380, 2001.

[Hugonnet et al., 1998] C. Hugonnet and P. Walder, *Théorie et pratique de la prise de son stéréophonique*. France: Eyrolles, 1998.

[Hunt et al., 1989] M.J. Hunt and C. Lefebvre, "A comparison of several acoustic representations for speech recognition with degraded and undegraded speech," *Proc. of IEEE ICASSP*, pp. 262-265, 1989.

[Hwang et al., 1989] M.Y. Hwang, H.W. Hon, and K.F. Lee, "Interword coarticulation modeling for continuous speech recognition," *Proc. of JASA*, vol. 1, p. 124, 1989.

[IPA, 1999] International Phonetic Association IPA, *Handbook of the International Phonetic Association: A Guide to the Use of the International Phonetic Alphabet*. Cambridge, United Kingdom: Cambridge University Press, 1999.

[Jankowski et al., 1990] C. Jankowski, A. Kalyanswamy, S. Basson, and J. Spitz, "NTIMIT: a phonetically balanced, continuous speech, telephone bandwidth speech database," *Proc. of IEEE ICASSP*, pp. 109-112, 1990.

[Janse, 2004] E. Janse, "Word perception in fast speech: artificially time-compressed vs naturally produced fast speech," *Speech Communication*, vol. 2, no. 42, pp. 155-173, 2004.

[Jarre et al., 1987] A. Jarre and R. Pieraccini, "Some experiments on HMM speaker adaptation," *Proc. of IEEE ICASSP*, no. 12, pp. 1273-1276, 1987.

[Jelinek, 1976] F. Jelinek, "Continuous speech recognition by statistical methods," *Proc. of the IEEE*, vol. 64, no. 4, pp. 532-556, 1976.

[Jelinek, 1976] F. Jelinek, "Continuous speech recognition by statistical methods," *Proc. of the IEEE*, vol. 4, no. 64, pp. 532-556, 1976.

[Jiang et al., 1999] K. Jiang and X. Huang, "Acoustic feature selection using speech recognizers," *Proc. of IEEE Workshop on ASRU*, 1999.

[Jie et al., 2009] Y. Jie and W. Zhenli, "On the application of variable-step adaptive noise cancelling for improving the robustness of speech recognition," *ISECS International Colloquium on Computing, Communication, Control, and Management*, vol. 2, pp. 419-422, 2009.

[Jitsuhiro et al., 1998] T. Jitsuhiro, S. Takahashi, and K. Aikawa, "Rejection of out-of-vocabulary words using phoneme confidence likelihood," *IEEE Trans. on ASSP*, vol. 1, pp. 217-220, 1998.

[Jiucang et al., 2009] H. Jiucang, H. Attias, S. Nagarajan, T.W. Lee, and T.J. Sejnowski, "Speech enhancement, gain, and noise spectrum adaptation using approximate bayesian estimation," *IEEE Trans. on ASLP*, vol. 1, no. 17, pp. 24-37, 2009.

[Johnson et al., 1999] S.E. Johnson, P. Jourlin, G.L. Moore, K.S. Jones, and P.C. Woodland, "The Cambridge University spoken document retrieval system," *Proc. of IEEE ICASSP*, vol. 1, pp. 49-52, 1999.

[Juang et al., 1987] B.H. Juang, L.R. Rabiner, and J.G. Wilpon, "On the use of bandpassliftering in speech recognition," *IEEE Trans. on ASSP*, no. 35, pp. 947-953, 1987.

[Juang et al., 1992] B.H. Juang and S. Katagiri, "Discriminative learning for minimum error classification," *IEEE Trans. on Signal Processing*, vol. 40, no. 12, pp. 3043-3054, 1992.

[Juang et al., 2005] B.H. Juang and L.R. Rabiner, "Automatic Speech Recognition - A brief history of the technology development," in *Elsevier Encyclopedia of Language and Linguistics*., 2005.

[Junkawitsch et al., 1997] J. Junkawitsch, G. Ruske, and H. Hoege, "Efficient methods for detecting keywords in continuous speech," *Proc. of Eurospeech*, pp. 259-262, 1997.

[Junqua, 1995] J. Junqua, "The influence of acoustics on speech production: a noise-induces stress phenomenon known as the Lombard reflex," *ESCQ-NQTO Workshop on Speech under Stress*, pp. 83-90, 1995.

[Junqua, 1997] J.C. Junqua, "Impact of the unknown communication channel on Automatic Speech Recognition: A review," *Proc. of Eurospeech*, vol. KN, pp. 29-32, 1997.

[Kajarekar et al., 1999] S. Kajarekar, N. Malayath, and H. Hermansky, "Analysis of speaker and channel variability in speech," *Proc. of IEEE Workshop on ASRU*, 1999.

[Kamper et al., 2009] H. Kamper and T.R. Niesler, "Characterisation and simulation of telephone channels using the TIMIT and NTIMIT databases," *Proc. of PRASA*, pp. 47-52, 2009.

[Kanazawa et al., 1995] H. Kanazawa, M. Tachimori, and Y. Takebayashi, "A hybrid wordspotting method for spontaneous speech understanding using word-based pattern matching and phoneme-based HMM," *Proc. of IEEE ICASSP*, vol. 1, pp. 289-292, 1995.

[Kawabata et al., 1988] T. Kawabata, T. Hanazawa, and K. Shikano, "Word spotting method based on HMM phoneme recognition," *Proc. of JASA*, no. 84, p. S62, 1988.

[Ke et al., 2005] Y. Ke, D. Hoiem, and R. Sukthankar, "Computer vision for music identification," *Proc. of IEEE Computer Vision and Pattern Recognition*, vol. 1, pp. 597-604, 2005.

[Kersta, 1962] L.G. Kersta, "Voiceprint identification," *Nature*, vol. 196, pp. 1253-1257, 1962.

[Keshet et al., 2009] J. Keshet, D. Grangier, and S. Bengio, "Discriminative keyword spotting," *Speech Communication*, vol. 51, no. 4, pp. 317-329, 2009.

[Kingsbury, 1998] B. Kingsbury, "Perceptually inspired signal processing strategies for robust speech recognition in reverberant environments," University of California, Berkeley, USA, PhD thesis, 1998.

[Kingsbury et al., 1998] B. Kingsbury, N. Morgan, and S. Greenberg, "Robust speech recognition using the modulation spectrogram," *Speech Communication*, vol. 1-3, no. 25, pp. 117-132, 1998.

[Kingsbury et al., 2002] B. Kingsbury, G. Saon, L. Mangua, M. Padmanabhan, and R. Sarikaya, "Robust speech recognition in noisy environments: the 2001 IBM SPINE evaluation system," *Proc. of IEEE ICASSP*, vol. 1, pp. 53-56, 2002.

[Kirchhoff, 1998] K. Kirchhoff, "Combining articulatory and acoustic information for speech recognition in noise and reverberant environments," *Proc. of ICSLP*, pp. 891-894, 1998.

[Kleinschmidt et al., 2002] M. Kleinschmidt and D. Gelbart, "Improving word accuracy with Gabor feature extraction," *Proc. of ICSLP*, pp. 25-28, 2002.

[Klemm et al., 1995] H. Klemm, F. Class, and U. Kilian, "Word- and phrase spotting with syllable-based garbage modelling," *Proc. of Eurospeech*, pp. 2157-2160, 1995.

[Knill et al., 1995] K. Knill and S. Young, "Techniques for automatically transcribing unknown keywords for open keyword set HMM-based word-spotting," Cambridge University Engineering Dept., Cambridge, UK, Technical report, CUED/F-INFENG/TR 230, 1995.

[Knill et al., 1996] K.M. Knill and S.J. Young, "Fast implementation methods for Viterbi-based word-spotting," *Proc. of IEEE ICASSP*, vol. 1, pp. 522-525, 1996.

[Koreman et al., 1999] J. Koreman, B. Andreeva, and H. Strik, "Acoustic parameters versus phonetic features in ASR," *International Congress of Phonetic Sciences*, pp. 549-553, 1999.

[Korkmazskiy et al., 1997] F. Korkmazskiy, Biing-Hwang Juang, and F. Soong, "Generalized mixture of HMMs for continuous speech recognition," *Proc. of IEEE ICASSP*, vol. 2, pp. 1443-1446, 1997.

[Kristjansson et al., 2001] T. Kristjansson, B. Frey, L. Deng, and A. Acero, "Towards non-stationary model-based noise adaptation for large vocabulary speech recognition," *Proc. of IEEE ICASSP*, vol. 1, pp. 337-340, 2001.

[Kroon et al., 1992] P. Kroon and K. Swaminathan, "A high-quality multirate real-time CELP coder," *IEEE Journal on Selected Areas in Communications*, vol. 10, no. 5, pp. 850-857, 1992.

[Kubala et al., 1994] F. Kubala, A. Anastasakos, J. Makhoul, L. Ngyuen, R. Schwartz, and E. Zavaliagkos, "Comparative experiments on large vocabulary speech recognition," *Proc. of IEEE ICASSP*, pp. 561-564, 1994.

[Kullback et al., 1951] S. Kullback and R. Leibler, "On information and sufficiency," *Annals of Mathematical Statistics*, vol. 22, pp. 79-86, 1951.

[Kumar et al., 1998] N. Kumar and A.G. Andreou, "Heteroscedastic discriminant analysis and reduced rank HMMs for improved speech recognition," *Speech Communication*, vol. 4, no. 26, pp. 283-297, 1998.

[Kumaresan, 1998] R. Kumaresan, "An inverse signal approach to computing the envelope of a real valued signal," *IEEE Signal Processing Letters*, vol. 10, no. 5, pp. 256-259, 1998.

[Kumaresan et al., 1999] R. Kumaresan and A. Rao, "Model-based approach to envelope and positive instantaneous frequency estimation of signals with speech applications," *Proc. of JASA*, vol. 3, no. 105, pp. 1912-1924, 1999.

[Kurth, 2002] F. Kurth, "A ranking technique for fast audio identification," *Proc. of IEEE Multimedia Signal Processing*, pp. 186-189, 2002.

[Kuwabara, 1997] H. Kuwabara, "Acoustic and perceptual properties of phonemes in continuous speech as a function of speaking rate," *Proc. of Eurospeech*, pp. 1003-1006, 1997.

[Kwong et al., 1996] S. Kwong, C.W. Chau, and W.A. Halang, "Genetic algorithm for optimizing the nonlinear time alignment of automatic speech recognition systems," *IEEE Trans. on Industrial Electronics*, vol. 5, no. 43, pp. 559-566, 1996.

[Ladefoged et al., 1996] P. Ladefoged and I. Maddieson, *The sounds of the world's languages*, Wiley-Blackwell, Ed., 1996.

[Lakoff, 1987] G. Lakoff, *Women, fire and dangerous things: what categories reveal about the mind*.: University of Chicago Press, 1987.

[Lamel et al., 1991] L.F. Lamel, J.L. Gauvain, and M. Eskenazi, "BREF, a large vocabulary spoken corpus for French," *Proc. of Eurospeech*, pp. 505-508, 1991.

[Lancini et al., 2004] R. Lancini, F. Mapelli, and R. Pezzano, "Audio content identification by using perceptual hashing," *IEEE International Conference on Multimedia and Expo*, vol. 1, pp. 739-742, 2004.

[Landercy et al., 1982] A. Landercy and R. Renard, *Eléments de phonétique*.: Didier-CIPA, 1982.

[Laroche, 2002] J. Laroche, "Process for identifying audio content," Creative Technology Ltd., United States, Brevet, USA 6453252, 2002.

[Lawson et al., 2003] A.D. Lawson, D.M. Harris, and J.J. Grieco, "Effect of foreign accent on speech recognition in the NATO N-4 corpus," *Proc. of Eurospeech*, pp. 1505-1508, 2003.

[Le Guyader et al., 2000] A. Le Guyader, P. Philippe, and J.B. Rault, "Synthèse des normes de codage de la parole et du son (UIT-T, ETSI ET ISO/MPEG)," *Annals of Telecommunications*, vol. 55, no. 9-10, pp. 425-441, 2000.

[Leblouch, 2009] O. Leblouch, "Décodage acoustico-phonétique et applications à l'indexation audio automatique," Université de Toulouse III, Toulouse, France, Thèse de doctorat, 2009.

[Lebossé, 2008] J. Lebossé, "Méthodes d'identification pour le contrôle de l'utilisation de documents audio," Université de Caen, Caen, France, Thèse de doctorat, 2008.

[Lee, 1997] C.H. Lee, "Adaptive compensation for robust speech recognition," *IEEE Workshop on ASRU*, pp. 357-364, 1997.

[Lee et al., 1988] C.H. Lee, F.K. Soong, and B.H. Juang, "A segment model based approach to speech recognition," *Proc. of IEEE ICASSP*, vol. 1, pp. 501-541, 1988.

[Lee et al., 1988b] K.F. Lee and H.W. Hon, "Large-vocabulary speaker-independent continuous speech recognition using HMM," *Proc. of IEEE ICASSP*, vol. 1, pp. 123-126, 1988b.

[Lee et al., 1989] C.H. Lee, B.H. Juang, F.K. Soong, and L.R. Rabiner, "Word recognition using whole word and subword models," *Proc. of IEEE ICASSP*, vol. 1, pp. 683-686, 1989.

[Lee et al., 1989b] K.F. Lee and H.W. Hon, "Speaker independent phone recognition using hidden markov models," *IEEE Trans. on ASSP*, pp. 1641-1648, 1989b.

[Lee et al., 1996] L. Lee and R.C. Rose, "Speaker normalization using efficient frequency warping procedures," *Proc. of IEEE ICASSP*, vol. 1, pp. 353-356, 1996.

[Lefèvre, 2000] F. Lefèvre, "Estimation de probabilité non-paramétrique pour la reconnaissance markovienne de la parole," Université Pierre et Marie Curie, Paris, France, Thèse de doctorat, 2000.

[Lemaire, 2007] P. Lemaire, *Psychologie cognitive*, De Boeck, Ed. France: Ouvertures Psychologiques, 2007.

[Leonard, 1984] R.G. Leonard, "A database for speaker independent digit recognition," *Proc. of IEEE ICASSP*, pp. 328-331, 1984.

[Levinson et al., 1983] S. Levinson, L. Rabiner, and M. Sondhi, "Speaker independent isolated digit recognition using hidden Markov models ," *Proc. of IEEE ICASSP*, pp. 1049-1052, 1983.

[Levy, 2006] C. Levy, "Modèles acoustiques compacts pour les systèmes embarqués," Université d'Avignon et des Pays de Vaucluse, Avignon, France, Thèse de doctorat, 2006.

[Li et al., 2004] Y. Li and Y. Hou, "Search audio data with the wavelet pyramidal algorithm," *Inf. Proc. Letters*, vol. 1, no. 91, pp. 49-55, 2004.

[Li et al., 2011] Y. Li, J. Le, Y. Yang, and J. Wang, "Improvement algorithm of DTW on isolated-word recognition," *IEEE Int. Conf. on Computer Science and Automation Engineering*, vol. 3, pp. 319-322, 2011.

[Liberman et al., 1957] A.M. Liberman, K.S. Harris, H.S. Hoffman, and B.C. Griffith, "The discrimination of speech sounds winthin and across phoneme boundaries," *Journal of Experimental Psychology*, vol. 54, pp. 358-368, 1957.

[Lin et al., 1977] W. Lin and C. Chan, "An isolated word recognition system based on acoustic-phonetic analysis and statistical pattern recognition," *Proc. of IEEE ICASSP*, vol. 2, pp. 679-682, 1977.

[Lin et al., 2009] H. Lin, A. Stupakov, and J. Bilmes, "Improving multi-lattice alignment based spoken keyword spotting," *Proc. of IEEE ICASSP*, pp. 4877-4880, 2009.

[Lindblom, 1990] B. Lindblom, "Explaining phonetic variation: a sketch of the hyper- and hypospeech theory," in *Speech production and speech modelling*, A. Hardcastle and W.J. Marchal, Ed.: Kluwer Academic Publishers, 1990, pp. 403-439.

[Lindblom et al., 1973] B. Lindblom and S. Svensson, "Interaction between segmental and nonsegmental factors in speech recognition," *IEEE Trans. on Audio and Electroacoustics*, vol. 6, no. 21, pp. 536-545, 1973.

[Lippmann, 1997] R. Lippmann, "Speech recognition by machines and humans," *Speech Communication*, no. 22, pp. 1-15, 1997.

[Liu et al., 1997] L. Liu, J. He, and G. Palm, "Effects of the phase on the perception of intervocalic stop consonants," *Speech Communication*, vol. 4, no. 22, pp. 403-417, 1997.

[Ljolje et al., 1991] A. Ljolje and M.D. Riley, "Automatic segmentation and labeling of speech," *Proc. of IEEE ICASSP*, vol. 1, pp. 473-476, 1991.

[Lleida et al., 1993] E. Lleida, J.B. Mariño, J. Salavedra, A. Bonafonte, E. Monte, and A. Martinez, "Out-of-vocabulary word modelling and rejection for keyword spotting," *Proc. of Eurospeech*, pp. 1265-1268, 1993.

[Lockwood et al., 1992] P. Lockwood, J. Boudy, and M. Blanchet, "Non-linear spectral subtraction (NSS) and hidden Markov models for robust speech recognition in car noise environments," *Proc. of IEEE ICASSP*, vol. 1, pp. 265-268, 1992.

[Lombard, 1911] E. Lombard, "Le signe de l'élévation de la voix," *Annales des Maladies de L'Oreille et du Larynx*, no. 37, pp. 101-119, 1911.

[Loog et al., 2004] M. Loog and R.P.W. Duin, "Linear dimensionality reduction via a heteroscedastic extension of LDA: the Chernoff criterion," *IEEE Trans. on PAMI*, vol. 6, no. 26, pp. 732-739, 2004.

[Mackenzie, 1997] B.J. Mackenzie, "Organic variation of the vocal apparentus," in *The Handbook of Phonetic Sciences*, Blackwell, Ed., 1997, pp. 256-297.

[MacNeilage, 1973] P.F. MacNeilage, "Linguistic units and speech production theory," *Proc. of JASA*, vol. 1, no. 54, pp. 329-330, 1973.

[Mahalanobis, 1936] P.C. Mahalanobis, "On the generalised distance in statistics," *Proc. of the National Institute of Sciences of India*, vol. 2, no. 1, pp. 49-55, 1936.

[Mak et al., 1996] B. Mak and E. Barnard, "Phone clustering using the Bhattacharyya distance," *Proc. of ICSLP*, vol. 4, pp. 2005-2008, 1996.

[Makhoul, 1975] J. Makhoul, "Linear prediction: a tutorial review," *Proc. of the IEEE*, vol. 63, no. 4, pp. 561-580, 1975.

[Mansoo et al., 2006] P. Mansoo, K. Hoi-Rin, R. Yong Man, and K. Munchurl, "Frequency filtering for a highly robust audio fingerprinting scheme in a real-noise environment," *IEICE Trans. on Information and Systems*, vol. 7, no. 89, pp. 2324-2327, 2006.

[Marcus, 1992] J.N. Marcus, "A novel algorithm for HMM words spotting performance evaluation and error analysis," *Proc. of IEEE ICASSP*, vol. 2, pp. 89-92, 1992.

[Mari et al., 1994] J.F. Mari and J.P. Haton, "Automatic word recognition based on second-order Hidden Markov models," *Proc. of ICSLP*, pp. 247-250, 1994.

[Mariani, 2002] J. Mariani, *Analyse, Synthèse et Codage de la Parole*.: Hermès - Lavoisier, 2002.

[Martinez et al., 1997] F. Martinez, D. Tapias, J. Alvarez, and P. Leon, "Characteristics of slow, average and fast speech and their effects in large vocabulary continuous speech recognition," *Proc. of Eurospeech*, pp. 469-472, 1997.

[Mathan et al., 1991] L. Mathan and L. Miclet, "Rejection of extraneous input in speech recognition applications, using multi-layer perceptrons and the trace of HMMs," *Proc. of IEEE ICASSP*, vol. 1, pp. 93-96, 1991.

[Matsu'ura et al., 1994] H. Matsu'ura, Y. Masai, J. Iwasaki, S. Tanaka, H. Kamio, and T. Nitta, "A multimodal, keyword-based spoken dialogue system-MultiksDial," *Proc. of IEEE ICASSP*, vol. 2, pp. 33-36, 1994.

[Mauuari, 1998] L. Mauuari, "Blind equalization in the cepstral domain for robust telephone based speech recognition," *Proc. of EUSIPCO*, no. 9, pp. 359-362, 1998.

[Medress et al., 1978] M. Medress, T. Diller, D. Kloker, L. Lutton, H. Oredson, and T. Skinner, "An automatic word spotting system for conversational speech," *Proc. of IEEE ICASSP*, vol. 3, pp. 712-717, 1978.

[Medress et al., 1979] M. Medress, M. Derr, T. Diller, D. Kloker, L. Lutton, H. Oredson, J. Siebenand, and T. Skinner, "Word spotting in conversational speech," *Proc. of IEEE ICASSP*, vol. 4, pp. 599-602, 1979.

[Medress et al., 1979] M. Medress, M. Derr, T. Diller, D. Kloker, L. Lutton, H. Oredson, J. Siebenand, and T. Skinner, "Word spotting in conversational speech," *Proc. of IEEE ICASSP*, vol. 4, pp. 599-602, 1979.

[Menéndez-Pidal et al., 2001] X. Menéndez-Pidal, R. Chena, D. Wua, and M. Tanaka, "Compensation of channel and noise distortions combining normalization and speech enhancement techniques," *Speech Communication*, no. 34, pp. 115-126, 2001.

[Mercier et al., 1982] G. Mercier, A. Callec, J. Monne, M. Querre, and O. Trevarain, "Automatic segmentation, recognition of phonetic units and training in the KEAL speech recognition system," *Proc. of IEEE ICASSP*, vol. 7, pp. 2000-2003, 1982.

[Mergel et al., 1985] D. Mergel and H. Ney, "Phonetically guided clustering for isolated word recognition," *Proc. of IEEE ICASSP*, pp. 854-857, 1985.

[Mertins et al., 2005] A. Mertins and J. Redemacher, "Vocal tract length invariant features for automatic speech recognition," *Proc. of IEEE Workshop on ASRU*, pp. 308-312, 2005.

[Meunier, 2005] C. Meunier, "Invariants et variabilité en phonétique," in *Phonologie et phonétique*.: Hermès, 2005, pp. 349-374.

[Meyer et al., 2010] B.T. Meyer, T. Jürgens, T. Wesker, T. Brand, and B. Kollmeier, "Human phoneme recognition depending on speech-intrinsic variability," *Proc. of JASA*, vol. 5, no. 128, pp. 3126-3141, 2010.

[Meyer et al., 2011] B.T. Meyer and B. Kollmeier, "Robustness of spectro-temporal features against intrinsic and extrinsic variations in automatic speech recognition," *Speech Communication*, vol. 5, no. 53, pp. 753-767, 2011.

[Miet, 2001] G. Miet, "Towards wideband speech by narrowband speech bandwidth extension: magic effect or wideband recovery?," Université du Maine, Thèse de doctorat, 2001.

[Millar et al., 1988] P.C. Millar, I.R. Cameron, A.J. Greaves, and C.M. McPeake, "A very large telephone-speech database collected using an automated voice-interactive dialogue," *Proc. of IEEE ICASSP*, vol. 1, pp. 647-650, 1988.

[Miller, 1994] J.L. Miller, "On the internal structure of phonetic categories: a progress report," *Cognition*, vol. 50, pp. 271-285, 1994.

[Milner, 1996] B.P. Milner, "Inclusion of temporal information into features for speech recognition," *Proc. of ICSLP*, pp. 256-259, 1996.

[Milner et al., 2011] B. Milner and J. Darch, "Robust Acoustic Speech Feature Prediction From Noisy Mel-Frequency Cepstral CoefficientsMilner, B. Darch, J.," *IEEE Trans. on ASLP*, vol. 2, no. 19, pp. 338-347, 2011.

[Mirghafori et al., 1995] N. Mirghafori, E. Fosler, and N. Morgan, "Fast speakers in large vocabulary continuous speech recognition: analysis and antidotes," *Proc. of Eurospeech*, pp. 491-494, 1995.

[Mokbel, 1992] C. Mokbel, "Reconnaissance de la parole dans le bruit : bruitage/débruitage," ENST, Paris, France, Thèse de doctorat, 1992.

[Mokbel et al., 1995] C.E. Mokbel and G.F.A. Chollet, "Automatic word recognition in cars," *IEEE Trans. on SAP*, vol. 3, no. 5, pp. 346-356, 1995.

[Mokhtari, 1998] P. Mokhtari, "An acoustic-phonetic and articulatory study of speech-speaker dichotomy," University of New South Wales, Australia, PhD thesis, 1998.

[Montacié et al., 1992] C. Montacié, P. Deleglise, F. Bimbot, and M.J. Caraty, "Cinematic techniques for speech processing: temporal decomposition and multivariate linear prediction," *Proc. of IEEE ICASSP*, vol. 1, pp. 153-156, 1992.

[Montacié et al., 1996] C. Montacié, M.J. Caraty, and C. Barras, "Mixture splitting technic and temporal control in a HMM-based recognition system ," *Proc. of ICSLP*, pp. 977-980, 1996.

[Montacié et al., 2011] C. Montacié and M.J. Caraty, "Combining Multiple Phoneme-Based Classifiers with Audio Feature-Based Classifier for the Detection of Alcohol Intoxication," *Proc. of Interspeech*, pp. 3205-3208, 2011.

[Morales et al., 2009] N. Morales, D.T. Toledano, J.H.L. Hansen, and J. Garrido, "Feature compensation techniques for ASR on band-limited speech ," *IEEE Trans. on ASLP*, vol. 4, no. 17, pp. 758-774, 2009.

[Moreno et al., 1994] P.J. Moreno and R.M. Stern, "Sources of degradation of speech recognition in the telephone network," *Proc. of IEEE ICASSP*, vol. 1, pp. 109-112, 1994.

[Morgan et al., 1990] D.P. Morgan, C.L. Scofield, T.M. Lorenzo, E.C. Real, and D.P. Loconto, "A keyword spotter which incorporates neural networks for secondary processing," *Proc. of IEEE ICASSP*, vol. 1, pp. 113-116, 1990.

[Morgan et al., 2004] N. Morgan, B. Chen, Q. Zhu, and A. Stolcke, "TRAPping conversational speech: extending TRAP/tandem approaches to conversational telephone speech recognition," *Proc. of IEEE ICASSP*, vol. 1, pp. 536-539, 2004.

[Moulin, 2010] P. Moulin, "Statistical modeling and analysis of content identification," *Information Theory and Applications Workshop*, pp. 1-5, 2010.

[Murty et al., 2006] K.S.R. Murty and B. Yegnanarayana, "Combining evidence from residual phase and MFCC features for speaker recognition," *IEEE Signal Processing Letters*, vol. 1, no. 13, pp. 52-55 , 2006.

[Myers et al., 1980] C.S. Myers, L.R. Rabiner, and A.E. Rosenberg, "An investigation of the use of Dynamic Time Warping for word spotting and connected speech recognition," *Proc. of IEEE ICASSP*, pp. 173-177, 1980.

[Myers et al., 1981] C.S. Myers and L.R. Rabiner, "A level building dynamic time warping algorithm for connected word recognition," *IEEE Trans. on ASSP*, vol. 2, no. 29, pp. 284-297, 1981.

[Myers et al., 1981b] C.S. Myers and L.R. Rabiner, "Connected digit recognition using a level-building DTW algorithm," *IEEE Trans. on ASSP*, vol. 3, no. 29, pp. 351-363, 1981b.

[Myers et al., 1981c] C.S. Myers and L.R. Rabiner, "A comparative study of several dynamic time-warping algorithms for connected word recognition," *The Bell System Technical Journal*, vol. 7, no. 60, pp. 1389-1409, 1981c.

[Nakamura et al., 1993] S. Nakamura, T. Akabane, and S. Hamaguchi, "Robust word spotting in adverse car environments," *Proc. of Eurospeech*, pp. 1045-1048, 1993.

[Nakazawa et al., 2001] M. Nakazawa and R. Oka, "The slot expression for topic spotting and topic summary in spontaneous speech," *International Conf. on Multisensor Fusion and Integration for Intelligent Systems*, pp. 221-226, 2001.

[Nedel, 2004] J.P. Nedel, "Duration normalization for robust recognition of spontaneous speech via missing feature methods," Carnegie Mellon University, Pittsburgh, USA, PhD thesis, 2004.

[Neumeyer et al., 1995] L. Neumeyer and M. Weintraub, "Robust speech recognition in noise using adaptation and mapping techniques," *Proc. of IEEE ICASSP*, vol. 1, pp. 141-144, 1995.

[Ng et al., 1998] K. Ng and V.W. Zue, "Phonetic recognition for spoken document retrieval," *Proc. of IEEE ICASSP*, vol. 1, pp. 325-328, 1998.

[NIST SDR, 1997] NIST SDR. (1997) National Institute of Standards and Technology, Spoken Document Retrieval. [Online]. http://www.itl.nist.gov/iad/mig/tests/sdr

[NIST STD, 2006] NIST STD. (2006) National Institute of Standards and Technology, Spoken Term Detection. [Online]. http://www.itl.nist.gov/iad/mig/tests/std

[Nolan, 1983] F. Nolan, *The phonetic bases of speaker recognition.*: Cambridge University Press, 1983.

[Ogle et al., 2007] J.P. Ogle and D.P.W. Ellis, "Fingerprinting to identify repeated sound events in long-duration personal audio recordings," *Proc. of IEEE ICASSP*, pp. 233-236, 2007.

[Omar et al., 2002] M.K. Omar, K. Chen, M. Hasegawa-Johnson, and Y. Bradman, "An evaluation of using mutual information for selection of acoustic features representation of phonemes for speech recognition," *Proc. of ICSLP*, pp. 2129-2132, 2002.

[Omar et al., 2002b] M.K. Omar and M. Hasegawa-Johnson, "Maximum mutual information based acoustic features representation of phonological features for speech recognition," *Proc. of IEEE ICASSP*, vol. 1, pp. 81-84, 2002b.

[Ono et al., 1993] Y. Ono, H. Wakita, and Y. Zhao, "Speaker normalization using constrained spectra shifts in auditory filter domain," *Proc. of Eurospeech*, pp. 355-358, 1993.

[O'Shaughnessy, 1974] D. O'Shaughnessy, "Consonant durations in clusters," *IEEE Trans. on ASSP*, vol. 4, no. 22, pp. 282-295, 1974.

[O'Shaughnessy, 1987] D. O'Shaughnessy, *Speech Communications: Human and Machine*, Series in Electrical Engineering ed. USA: Addison-Wesley Publishing Co., 1987.

[Pachet, 2005] F. Pachet, "Knowledge management and musical metadata," in *Encyclopedia of Knowledge Management*, Idea Group, Ed.: Schwartz, 2005.

[Padmanabhan et al., 2005] M. Padmanabhan and S. Dharanipragada, "Maximizing information content in feature extraction," *IEEE Trans. on SAP*, vol. 4, no. 13, pp. 512-519, 2005.

[Paliwal et al., 1991] K.K. Paliwal and B.S. Atal, "Efficient vector quantization of LPC parameters at 24 bits/frame," *Proc. of IEEE ICASSP*, pp. 661-664, 1991.

[Paliwal et al., 2003] K.K. Paliwal and B.S. Atal, "Frequency-related representation of speech," *Proc. of Eurospeech*, vol. 65-68, 2003.

[Pan, 1995] D. Pan, "A tutorial on MPEG/audio compression," *IEEE Multimedia*, vol. 2, no. 2, pp. 60-74, 1995.

[Panagiotakis et al., 2005] C. Panagiotakis and G. Tziritas, "A speech/music discriminator based on RMS and zero-crossings," *IEEE Trans. on Multimedia*, vol. 7, pp. 155-166, 2005.

[Papaodysseus et al., 2001] C. Papaodysseus, G. Roussopoulos, D. Fragoulis, T.H. Panagopoulos, and C. Alexiou, "A new approach to the automatic recognition of musical recordings," *Journal of the Audio Engineering Society*, vol. 49, no. 1, pp. 23-35, 2001.

[Park et al., 2005] A. Park and J.R. Glass, "Towards unsupervised pattern discovery in speech," *IEEE Workshop on Automatic Speech Recognition and Understanding*, pp. 53-58, 2005.

[Paul, 1985] D. Paul, "Training of HMM recognizers by simulated annealing," *Proc. of IEEE ICASSP*, no. 10, pp. 13-16, 1985.

[Peeters et al., 2009] H. Peeters, F. Kuk, C.C. Lau, and D. Keenan, "Subjective and objective evaluation of noise management algorithms," *Proc. of JAAA*, vol. 2, no. 20, pp. 89-98, 2009.

[Peng et al., 2008] Y. Peng, S. Yu, and F. Seide, "Approximate word-lattice indexing with text indexers: Time-Anchored Lattice Expansion," *Proc. of IEEE ICASSP*, pp. 5248-5251, 2008.

[Peters et al., 1999] S.D. Peters, P. Stubley, and J.M. Valin, "On the limits of speech recognition in noise," *Proc. of IEEE ICASSP*, pp. 365-368, 1999.

[Picone, 1993] J. Picone, "Signal modeling techniques in speech recognition," *Proc. of IEEE ICASSP*, vol. 9, no. 81, pp. 1215-1247, 1993.

[Pinquier, 2004] J. Pinquier, "Indexation sonore : recherche de composantes primaires pour une structuration audiovisuelle," Université de Toulouse III, Toulouse, France, Thèse de doctorat, 2004.

[Pinquier et al., 2002] J. Pinquier, C. Sénac, and R. André-Obrecht, "Speech and music classification in audio documents," *Proc. of IEEE ICASSP*, vol. 4, pp. 164-168, 2002.

[Pinquier et al., 2004] J. Pinquier and R. André-Obrecht, "Jingle detection and identification in audio documents," *Proc. of IEEE ICASSP*, vol. 4, pp. 329-332, 2004.

[Pujol et al., 2005] P. Pujol, S. Pol, C. Nadeu, A. Hagen, and H. Bourlard, "Comparison and combination of features in a hybrid HMM/MLP and a HMM/GMM speech recognition system," *IEEE Trans. on SAP*, vol. 1, no. 13, pp. 14-22, 2005.

[Rabiner, 1989] L.R. Rabiner, "A tutorial on Hidden Markov Models and selected applications in speech recognition," *Proc. of IEEE*, vol. 77, pp. 257-286, 1989.

[Rabiner et al., 1977] L.R. Rabiner and M. Sambur, "Voiced-unvoiced-silence detection using the Itakura LPC distance measure," *Proc. of IEEE ICASSP*, vol. 2, pp. 323-326, 1977.

[Rabiner et al., 1980] L.R. Rabiner and C. Schmidt, "Application of dynamic time warping to connected digit recognition," *IEEE Trans. on ASSP*, no. 28, pp. 377-388, 1980.

[Rabiner et al., 1993] L.R. Rabiner and B.H. Juang, *Fundamentals of Speech Recognition*.: Oxford University Press, 1993.

[Ramona et al., 2011] M. Ramona and G. Peeters, "Audio identification based on spectral modeling of bark-bands energy and synchronization through onset detection," *Proc. of IEEE ICASSP*, pp. 477-480, 2011.

[Rault et al., 1995] J.B. Rault, Y.F. Dehery, and M. Lever, "The ISO/MPEG Audio MISICAM family," *IEE Colloquium on MPEG2*, pp. 310-314, 1995.

[Raut et al., 2009] C.K. Raut and M.J.F. Gales, "Bayesian discriminative adaptation for speech recognition," *Proc. of IEEE ICASSP*, pp. 4361-4364, 2009.

[Reeves et al., 1998] L.M. Reeves, K. Hirsh-Pasek, and R. Golinkoff, "Words and Meaning: from primitives to complex organization," in *Psycholinguistics*.: Gleason & Ratner, 1998, ch. 4, pp. 157-226.

[Rennie et al., 2011] S. Rennie, P. Dognin, and P. Fousek, "Robust speech recognition using dynamic noise adaptation," *Proc. of IEEE ICASSP*, pp. 4592-4595, 2011.

[RIAA, 2001] IFPI RIAA, "Request for information on audio fingerprinting technologies," 2001.

[Riley, 1991] M.D. Riley, "A statistical model for generating pronunciation networks," *Proc. of IEEE ICASSP*, vol. 2, pp. 737-740, 1991.

[Rohlicek et al., 1989] J.R. Rohlicek, W. Russell, S. Roukos, and H. Gish, "Continuous Hidden Markov Modeling for speaker-independent wordspotting," *Proc. of IEEE ICASSP*, vol. 1, pp. 627-630, 1989.

[Rohlicek et al., 1993] J.R. Rohlicek, P. Jeanrenaud, K. Ng, H. Gish, B. Musicus, and M. Siu, "Phonetic training and language modeling for word spotting," *Proc. of IEEE ICASSP*, vol. 2, pp. 459-462, 1993.

[Rose, 1992] R.C. Rose, "Discriminant wordspotting techniques for rejection of non-vocabulary utterances in unconstrained speech," *Proc. of IEEE ICASSP*, pp. 105-108, 1992.

[Rose, 1995] R.C. Rose, "Keyword detection in conversational speech utterances using hidden Markov model based continuous speech recognition," *Computer, Speech and Language*, vol. 9, no. 4, pp. 309-333, 1995.

[Rose et al., 1990] R.C. Rose and D.B. Paul, "A hidden Markov model based keyword recognition system ," *Prof. of IEEE ICASSP*, pp. 129-132, 1990.

[Rose et al., 1991] R.C. Rose, E.I. Chang, and R.P. Lippmann, "Techniques for information retrieval from voice messages," *Proc. of IEEE ICASSP*, pp. 317-320, 1991.

[Rose et al., 1993] R.C. Rose and E.M. Hofstetter, "Task independent wordspotting using decision tree based allophone clustering," *Proc. of IEEE ICASSP*, vol. 2, pp. 467-470, 1993.

[Rouvier et al., 2008] M. Rouvier, G. Linares, and B. Lecouteux, "On-the-fly term spotting by phonetic filtering and request-driven decoding," *IEEE Workshop on SLT*, pp. 305-308, 2008.

[Russell et al., 1983] M. Russell, R. Moore, and M. Tomlinson, "Some techniques for incorporating local timescale variability information into a dynamic time-warping algorithm for automatic speech recognition," *Proc. of IEEE ICASSP*, vol. 8, pp. 1037-1040, 1983.

[Sakoe et al., 1978] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *Proc. of IEEE ICASSP*, pp. 43-49, 1978.

[Salton, 1962] G. Salton, "Manipulation of trees in information retrieval," *Commun. ACM*, vol. 5, no. 2, pp. 103-114, 1962.

[Sandness et al., 2000] E.D. Sandness and I.L. Hetherington, "Keyword-based discriminative training of acoustic models," *Proc. of ICSLP*, vol. 3, pp. 135-138, 2000.

[Saon et al., 2000] G. Saon, M. Padmanabhan, R. Gopinath, and S. Chen, "Maximum likelihood discriminant feature spaces," *Proc. of IEEE ICASSP*, pp. 1129-1132, 2000.

[Saunders, 1996] J. Saunders, "Real-time discrimination of broadcast speech/music," *Proc. of IEEE ICASSP*, pp. 993-996, 1996.

[Scheirer et al., 1997] E. Scheirer and M. Slaney, "Construction and evaluation of a robust mainframe speech/music discriminator," *Proc. of IEEE ICASSP*, pp. 1331-1334, 1997.

[Scheirer et al., 1997] E. Scheirer and M. Slaney, "Construction and evaluation of a robust multifeature speech/music discriminator," *Proc. of IEEE ICASSP*, vol. 2, pp. 1331-1334, 1997.

[Schimmel et al., 2005] S. Schimmel and L. Atlas, "Coherent envelope detection for modulation filtering of speech," *Proc. of IEEE ICASSP*, vol. 1, pp. 221-224, 2005.

[Schurer, 1994] T. Schurer, "An experimental comparison of different feature extraction and classification methods for telephone speech," *IEEE Workshop on IVTTA*, pp. 93-96, 1994.

[Schwartz et al., 1984] R. Schwartz, Y. Chow, S. Roucos, M. Krasner, and J. Makhoul, "Improved hidden Markov modeling of phonemes for continuous speech recognition," *Proc. of IEEE ICASSP*, no. 9, pp. 21-24, 1984.

[Scully, 1987] C. Scully, "Linguistic units and units of speech production," *Speech Communication*, vol. 2, no. 6, pp. 77-142, 1987.

[Seidel, 2009] C. Seidel, "Content fingerprinting from an industry perspective," *IEEE Int. Conf. on Multimedia and Expo*, pp. 1524-1527, 2009.

[Shafran et al., 2000] I. Shafran and M. Ostendorf, "Use of higher level linguistic structure in acoustic modeling for speech recognition," *Proc. of IEEE ICASSP*, vol. 2, pp. 1021-1024, 2000.

[Shannon, 1949] C.E. Shannon, "Communication in the presence of noise," *Proc. of Institute of Radio Engineers*, vol. 37, no. 1, pp. 10-21, 1949.

[Shannon et al., 1949] C.E. Shannon and W. Weaver, *The mathematical theory of communication*.: University of Illinois Press, 1949.

[Shlien, 1994] S. Shlien, "Guide to MPEG-1 audio standard," *IEEE Trans. on Broadcasting*, pp. 206-218, 1994.

[Siegler et al., 1995] M.A. Siegler and R.M. Stern, "On the effect of speech rate in large vocabulary speech recognition system," *Proc. of IEEE ICASSP*, pp. 612-615, 1995.

[Siegler et al., 1997] M.A. Siegler, U. Jain, B. Raj, and R.M. Stern, "Automatic segmentation, classification and clustering of broadcast news audio," *Proc. of DARPA Speech Recognition Workshop*, pp. 97-99, 1997.

[Sigasaka, 1988] Y. Sigasaka, "Speech synthesis by rule using an optimal selection of non-uniform synthesis," *Proc. of IEEE ICASSP*, pp. 679-682, 1988.

[Siohan, 1995] O. Siohan, "On the robustness of linear discriminant analysis as a preprocessing step for noisy speech recognition," *Proc. of IEEE ICASSP*, vol. 1, pp. 125-128, 1995.

[Sirigos et al., 1997] J. Sirigos, N. Fakotakis, and G. Kokkinakis, "Improving environmental robustness of speech recognition using neural networks," *Proc. on International Conference on Digital Signal Processing*, vol. 2, pp. 575-578, 1997.

[Slifka et al., 1995] J. Slifka and T.R. Anderson, "Speaker modification with LPC pole analysis," *Proc. of IEEE ICASSP*, pp. 644-647, 1995.

[Smídl et al., 2006] L. Smídl and J. Psutka, "Comparison of keyword spotting methods for searching in speech," *Proc. of Interspeech*, pp. 1894-1897, 2006.

[Sotillo et al., 1998] C. Sotillo and E.G. Bard, "Is hypo-articulation lexically constrained?," *Proc. of SPoSS*, pp. 109-112, 1998.

[Stephenson et al., 2004] T.A. Stephenson, M.M. Doss, and H. Bourlard, "Speech recognition with auxiliary information," *IEEE Trans. on SAP*, vol. 3, no. 12, pp. 189-203, 2004.

[Stevens et al., 1937] S.S. Stevens, J. Volkman, and E. Newman, "A scale for the measurement of the psychological magnitude pitch," *Proc. of JASA*, vol. 3, no. 8, pp. 185–190, 1937.

[Stevens et al., 1978] K.N. Stevens and S.E. Blumstein, "Invariant cues for place of articulation in stop consonants," *Proc. of JASA*, vol. 64, pp. 1358-1368, 1978.

[Suaudeau et al., 1994] N. Suaudeau and R. André-Obrecht, "An efficient combination of acoustic and supra-segmental informations in a speech recognition system," *Proc. of IEEE ICASSP*, vol. 1, pp. 65-68, 1994.

[Sukkar, 1994] R.A. Sukkar, "Rejection for connected digit recognition based on GPD segmental discrimination," *Proc. of IEEE ICASSP*, pp. 373-376, 1994.

[Sukkar et al., 1996] R.A. Sukkar, A.R. Seltur, M.G. Rahim, and C.H. Lee, "Utterance verification of keyword strings using word-based minimum verification error (WB-MVE) training," *Proc. of IEEE ICASSP*, pp. 518-521, 1996.

[Sukkar et al., 1996] R.A. Sukkar and Chin-Hui Lee, "Vocabulary independent discriminative utterance verification for nonkeyword rejection in subword based speech recognition," *IEEE Trans. on SAP*, vol. 4, pp. 320-329, 1996.

[Svendsen et al., 1987] T. Svendsen and F. Soong, "On the automatic segmentation of speech signals," *Proc. of IEEE ICASSP*, pp. 77-80, 1987.

[Svendsen et al., 1989] T. Svendsen, K.K. Paliwal, E. Harborg, and P.O. Husoy, "An improved sub-word based speech recognizer," *Proc. of IEEE ICASSP*, pp. 108-111, 1989.

[Tabibian et al., 2011] S. Tabibian, A. Akbari, and B. Nasersharif, "An evolutionary based discriminative system for keyword spotting," *International Symposium on AISP*, pp. 83-88, 2011.

[Tavenard et al., 2007] R. Tavenard, L. Amsaleg, and G. Gravier, "Machines à vecteurs supports pour la comparaison de séquences de descripteurs. ," *Proc. of CORESA*, pp. 247-251, 2007.

[Thomson et al., 1998] D.L. Thomson and R. Chengalvaryan, "Use of periodicity and jitter as speech recognition feature," *Proc. of IEEE ICASSP*, vol. 1, pp. 21-24, 1998.

[Thosar et al., 1976] R. Thosar and P. Rao, "An approach towards a synthesis-based speech recognition system ," *IEEE Trans. on ASSP*, vol. 2, no. 24, pp. 194-196, 1976.

[Tibrewala et al., 1997] S. Tibrewala and H. Hermansky, "Sub-band based recognition of noisy speech," *Proc. of IEEE ICASSP*, pp. 1255-1258, 1997.

[Tohkura et al., 1992] Y. Tohkura, E. Vatikiotis-Bateson, and Y. Sagisaka, *Speech perception, production and linguistic structure*. Tokyo, Ohmsha and Amsterdam, Japan: IOS Press, 1992.

[Tokuhira et al., 1999] M. Tokuhira and Y. Ariki, "Effectiveness of KL-transformation in spectral delta expansion," *Proc. of Eurospeech*, pp. 359-362, 1999.

[Tolba et al., 2002] H. Tolba, S.A. Selouani, and D. O'Shaughnessy, "Auditory-based acoustic distinctive features and spectral cues for automatic speech recognition using a multi-stream paradigm," *Proc. of IEEE ICASSP*, pp. 837-840, 2002.

[Tomlinson et al., 1997] M.J. Tomlinson, M.J. Russel, R.K. Moore, A.P. Buckland, and M.A. Fawley, "Modelling asynchrony in speech using elementary single-signal decomposition," *Proc. of IEEE ICASSP*, pp. 1247-1250, 1997.

[Tremain, 1982] T.E. Tremain, "The Government standard Linear Predictive Coding algorithm: LPC-10," *Speech Technology*, vol. 1, pp. 40-49, 1982.

[Treurniet et al., 1994] W.C. Treurniet and Y. Gong, "Noise independent speech recognition for a variety of noise types," *Proc. of IEEE ICASSP*, vol. 1, pp. 437-440, 1994.

[Tuerk et al., 1993] C. Tuerk and T. Robinson, "A new frequency shift function for reducing inter-speaker variance," *Proc. of Eurospeech*, vol. 1, pp. 351-354, 1993.

[Tuffelli et al., 1977] D. Tuffelli and B. Groc, "Word recognition and encoding using poles," *Proc. of IEEE ICASSP*, vol. 2, pp. 448-451, 1977.

[Turchi et al., 2010] M. Turchi, J. Steinberger, M. Kabadjov, and R. Steinberger, "Using parallel corpora for multilingual (multi-document) summarisation evaluation ," *Lecture Notes in Computer Science*, no. 6360, pp. 52-63, 2010.

[Tyagi et al., 2003] V. Tyagi, I. McCowan, H. Bourlard, and H. Misra, "Mel-cepstrum modulation spectrum (MCMS) features for robust ASR," *Proc. of IEEE Workshop on ASRU*, pp. 381-386, 2003.

[Tyagi et al., 2005] V. Tyagi and C. Wellekens, "Cepstrum representation of speech," *Proc. of IEEE Workshop on ASRU*, 2005.

[Tyagi et al., 2005b] V. Tyagi, C. Wellekens, and H. Bourlard, "On variable-scale piecewise stationary spectral analysis of speech signals for ASR," *Proc. of Interspeech*, pp. 209-212, 2005b.

[Umesh et al., 1999] S. Umesh, L. Cohen, N. Marinovic, and D. Nelson, "Scale transform in speech analysis," *IEEE Trans. on SAP*, vol. 1, no. 7, pp. 40-45, 1999.

[Vaseghi et al., 1995] S.V. Vaseghi and B.P. Milner, "Speech recognition in impulsive noise," *Proc. of IEEE ICASSP*, vol. 1, pp. 437-440, 1995.

[Vaseghi et al., 1997] S.V. Vaseghi, N. Harte, and B. Miller, "Multi resolution phonetic/segmental features and models for HMM-based speech recognition," *Proc. of IEEE ICASSP*, pp. 1263-1266, 1997.

[Vasiloglou et al., 2004] N. Vasiloglou, R. Schafer, and M. Hans, "Isolated word, speaker dependent recognition under the presence of noise, based on an audio retrieval algorithm," *Asilomar Conf. on Signals, Systems and Computers*, vol. 2, pp. 1809-1812, 2004.

[Vergyri et al., 2007] D. Vergyri, I. Shafran, A. Stolcke, R.R. Gadde, M. Akbacak, B. Roark, and W. Wang, "The SRI/OGI 2006 Spoken Term Detection system," *Proc. of the Interspeech*, pp. 2393-2396, 2007.

[Ververidis et al., 2006] D. Ververidis and C. Kotropoulos, "Emotional speech recognition: Resources, features, and methods," *Speech Communication*, vol. 9, no. 48, pp. 1162-1181, 2006.

[Vintsyuk, 1968] T.K. Vintsyuk, "Speech discrimination by dynamic programming," *Cybernetics and Systems Analysis*, pp. 52-57, 1968.

[Viterbi, 1967] A. Viterbi, "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm," *IEEE Trans. on Information Theory*, vol. 13, no. 2, pp. 260-269, 1967.

[Vogel et al., 2007] D. Vogel, P. McCarthy, G. Bratt, and C. Brewer, "The clinical audiogram: Its history and current use," *Communicative Disorders Review*, vol. 1, pp. 81-94, 2007.

[Wagner, 1981] M. Wagner, "Automatic labelling of continuous speech with a given phonetic transcription using dynamic programming algorithms," *Proc. of IEEE ICASSP*, vol. 6, pp. 1156-1159, 1981.

[Waibel et al., 1990] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, and K.J. Lang, "Phoneme recognition using time-delay neural networks," *Readings in speech recognition*, pp. 393-404, 1990.

[Wakita, 1977] H. Wakita, "Normalization of vowels by vocal-tract length and its application to vowel identification," *IEEE Trans. on ASSP*, no. 25, pp. 183-192, 1977.

[Wang, 2003] A. Wang, "An industrial-strength audio search algorithm," *Proc. of ISMIR*, pp. 582-588, 2003.

[Wang, 2006] A. Wang, "The Shazam music recognition service," *Communications of the ACM*, vol. 49, no. 8, pp. 44-48, 2006.

[Wang et al., 1996] X. Wang, L.C.W. Pols, and L.F.M. ten Bosch, "Analysis of context-dependent segmental duration for automatic speech recognition," *Proc. of ICSLP*, vol. 2, pp. 1181-1184, 1996.

[Wang et al., 2003] A. Wang and D. Culbert, "Robust and invariant audio pattern matching," Shazam Entertainment Ltd, Technical Patent, USA 20090265174, 2003.

[Warren, 1970] R.M. Warren, "Perceptual restoration of missing speech sounds," *Science*, vol. 167, pp. 392-395, 1970.

[Weintraub, 1993] M. Weintraub, "Keyword-spotting using SRI's DECIPHER large-vocabulary speech-recognition system," *Proc. of IEEE ICASSP*, vol. 2, pp. 463-466, 1993.

[Weintraub, 1995] M. Weintraub, "LVCSR log-likelihood ratio scoring for keyword spotting," *Proc. of IEEE ICASSP*, vol. 1, pp. 297-300, 1995.

[Weintraub et al., 1997] M. Weintraub, F. Beaufays, Z. Rivlin, Y. Konig, and A. Stolcke, "Neural-network based measures of confidence for word recognition," *Proc. of IEEE ICASSP*, vol. 2, pp. 887-890, 1997.

[Welling et al., 2002] L. Welling, H. Ney, and S. Kanthak, "Speaker adaptive modeling by vocal tract normalization," *IEEE Trans. on SAP*, vol. 6, no. 10, pp. 415-426, 2002.

[Westphal, 1997] M. Westphal, "The use of cepstral means in conversational speech recognition," *Proc. of Eurospeech*, vol. 3, pp. 1143-1146, 1997.

[Wilcox et al., 1991] L.D. Wilcox and M.A. Bush, "HMM-based wordspotting for voice editing and indexing," *Proc. of Eurospeech*, pp. 25-28, 1991.

[Wilpon et al., 1989] J.G. Wilpon, C.H. Lee, and L.R. Rabiner, "Application of hidden Markov models for recognition of a limited set of words in unconstrained speech ," *Proc. of IEEE ICASSP*, vol. 1, pp. 254-257, 1989.

[Wilpon et al., 1990] J.G. Wilpon, L.R. Rabiner, C.H. Lee, and E.R. Goldman, "Automatic recognition of keywords in unconstrained speech using hidden Markov models," *Proc. of IEEE TASSP*, pp. 1870-1878, 1990.

[Wöllmer et al., 2009] M. Wöllmer, F. Eyben, J. Keshet, A. Graves, B. Schuller, and G. Rigoll, "Robust discriminative keyword spotting for emotionally colored spontaneous speech using bidirectional LSTM networks," *Proc. of IEEE ICASSP*, pp. 3949-3952, 2009.

[Wong et al., 1998] P.H.W. Wong, O.C. Au, J.W.C. Wong, and W.H.B. Lau, "Reducing computational complexity of dynamic time warping-based isolated word recognition with time scale modification," *Int. Conf. on Signal Processing Proceedings*, vol. 1, pp. 722-725, 1998.

[Wood et al., 1991] L.C. Wood, D.J.B. Pearce, and F. Novello, "Improved vocabulary-independent sub-word HMM modelling," *Proc. of IEEE ICASSP*, vol. 1, pp. 181-184, 1991.

[Woodland et al., 1993] P.C. Woodland and S.J. Young, "The HTK tied-state continuous speech recogniser," *Proc. of Eurospeech*, pp. 2207-2210, 1993.

[Yang et al., 1996] X. Yang, J.B. Millar, and I. Macleod, "On the sources of inter- and intra-speaker variability in the acoustic dynamics of speech," *Proc. of ICSLP*, pp. 1792-1795, 1996.

[Yang et al., 2007] C. Yang, F.K. Soong, and T. Lee, "Static and dynamic spectral features: their noise robustness and optimal weights for ASR," *IEEE Trans. on ASSP*, vol. 3, no. 15, pp. 1087-1097, 2007.

[Young et al., 2006] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The HTK Book (for HTK version 3.4)*, Cambridge University Engineering Department, Ed., 2006.

[Zeppenfeld et al., 1992] T. Zeppenfeld and A.H. Waibel, "A hybrid-neural network, dynamic programming word spotter," *Proc. of IEEE ICASSP*, vol. 2, pp. 77-80, 1992.

[Zhan et al., 1997] P. Zhan and M. Westphal, "Speaker normalization based on frequency warping," *Proc. of IEEE ICASSP*, vol. 2, pp. 1039-1042, 1997.

[Zhang et al., 2001] Y. Zhang, R. Lee, and A. Madievski, "Confidence measure (CM) estimation for large vocabulary speaker-independent continuous speech recognition system," *Proc. of Eurospeech*, pp. 2545-2548, 2001.

[Zhang et al., 2005] B. Zhang and S. Matsoukas, "Minimum phoneme error based heteroscedastic linear discriminant analysis for speech recognition," *Proc. of IEEE ICASSP*, vol. 1, pp. 925-928, 2005.

[Zhu et al., 2000] Q. Zhu and A. Alwan, "AM-demodulation of speech spectra and its application to noise robust speech recognition," *Proc. of ICSLP*, vol. 1, pp. 341-344, 2000.

[Zhu et al., 2004] Q. Zhu, B. Chen, N. Morgan, and A. Stolcke, "On using MLP features in LVCSR," *Proc. of ICSLP*, pp. 921-924, 2004.

[Zhu et al., 2004b] D. Zhu and K.K. Paliwal, "Product of power spectrum and group delay function for speech recognition," *Proc. of IEEE ICASSP*, pp. 125-128, 2004b.

[Zolnay et al., 2002] A. Zolnay, R. Schlüter, and H. Ney, "Robust speech recognition using a voiced-unvoiced feature," *Proc. of ICSLP*, vol. 2, pp. 1065-1068, 2002.

[Zolnay et al., 2005] A. Zolnay, R. Schlüter, and H. Ney, "Acoustic feature combination for robust speech recognition," *Proc. of IEEE ICASSP*, vol. 1, pp. 457-460, 2005.

[Zue et al., 1997] V. Zue, S. Seneff, J. Glass, L. Hetherington, E. Hurley, H. Meng, C. Pao, J. Polifroni, R. Schloming, and P. Schmid, "From interface to content: translingual access and delivery of on-line information," *Proc. of Eurospeech*, pp. 2227-2230, 1997.

[Zwicker, 1961] E. Zwicker, "Subdivision of the audible frequency range into critical bands," *Proc. of JASA*, vol. 2, no. 33, p. 248, 1961.

[Zwicker et al., 1981] E. Zwicker and R. Feldtkeller, *Psychoacoustique - L'oreille récepteur d'information*, CNET - ENST, Ed.: Collection technique et scientifique des télécommunications, 1981.