Segmentation parole/musique pour la transcription automatique de parole continue

Emmanuel Didiot

Nancy-Université
Université Henri Poincaré

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

Segmentation parole/musique pour la transcription automatique de parole continue

THÈSE

présentée et soutenue publiquement le 13 Novembre 2007

pour l'obtention du

Doctorat de l'université Henri Poincaré – Nancy 1

(spécialité informatique)

par

Emmanuel Didiot

Composition du jury

Rapporteurs : Paul Deléglise , *Professeur, LIUM-CNRS Le Mans*
Christian J. Wellekens , *Professeur, Institut EURECOM Sophia Antipolis*

Examinateurs : Jean-Paul Haton , *Professeur, UHP-LORIA Nancy (Directeur)*
Jean-François Bonastre , *Maître de conférence, LIA-CERI Avignon*
Irina Illina , *Maître de conférence, UHP-LORIA Nancy (Co-directrice)*
Dominique Fohr , *Chargé de recherche CNRS, UHP-LORIA Nancy*
Laurent Besacier , *Maître de conférence, CLIPS-IMAG Grenoble*
Jean-Pierre Thomesse , *Professeur, INPL - ENSEM Nancy*

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

Loria

Mis en page avec la classe thloria.

Remerciements

La partie “remerciements” est, je pense, la partie la plus difficile à écrire dans un mémoire. On a toujours peur d'oublier quelqu'un. Je vais donc commencer par remercier le lecteur de ce mémoire, qui prend même le temps de lire les remerciements (enfin j'espère qu'il ne s'arrêtera pas là !).

Je ne sais pas par où commencer, je vais donc débuter en remerciant mes rapporteurs Paul Deléglise et Christian Wellekens ainsi que les membres du jury : Jean-François Bonastre, Laurent Besacier, et Jean-Pierre Thomesse pour le temps qu'ils ont consacré à la lecture de mon manuscrit et pour les corrections qu'ils ont pu y apporter.

Je tiens à remercier Irina Illina et Jean-Paul Haton pour avoir encadrer ma thèse. Ils m'ont laissé une totale liberté dans mes recherches tout en sachant me recadrer lorsque je m'éparpillais, je leur en suis très reconnaissant. Je tiens également à remercier Dominique Fohr et Odile Mella, qui m'ont apporté énormément. Dominique, par ses connaissances et son expérience en reconnaissance de la parole, a toujours su m'expliquer très clairement et simplement des concepts qui ne le sont pas toujours. Odile, par sa rigueur et sa méticulosité, a toujours été présente lorsqu'il s'agissait d'écrire et de peaufiner (c'est un faible mot) des articles ou des présentations. J'en arrive à l'équipe Parole dont j'ai fait parti à plusieurs reprises durant mon parcours. Je remercie Yves Laprie, chef de l'équipe, pour son accueil lors de mon stage de maîtrise pour lors de ma thèse. Je tiens également à remercier dans l'équipe mes collègues et amis : Alexandre (“l'ingénieur”), Wassim, Guillaume (qui nous a quitté trop tôt... je vous rassure il va très bien le petit !), Murat (“l'enfumeur”), Blaise, Farid, Ghazi, Marina (la petite dernière). Il y a maintenant les inclassables, que je souhaite remercier plus particulièrement :

- Joseph, je ne trouve aucun qualificatif pour toi, juste un fan d'Iron Maiden et de Valentino Rossi... oui des gens comme ça existent...
- Slim, “le râleur”, il trouve toujours quelque chose ou quelqu'un après qui râler.
- Seb, “le stressé tranquille”, perdu en Belgique, j'espère que tu retrouveras rapidement le chemin de la France.
- Jean-Pierre (JP pour les intimes), “le grand dadet qui parle fort” comme dirait Joseph, mon compatriote de la Meuse. A quand la prochaine sortie escalade ?
- Sabine, “bibine”, avec qui j'ai passé de bonnes soirées salsa mais plus important grâce à qui j'ai connu ma douce Caro, je t'en remercie.
- Pavel, “l'ami Tchèque”, avec qui j'ai passé d'inoubliables soirées (en intérieur ou au grand air).

Je n'oublie pas ma famille : mes parents (Christine et Régis), mes frères (Baptiste, Clément, Benjamin), mes grand-parents (qui sont partis trop vite, heureusement il reste “la vovone”) qui sont un moteur pour moi et qui m'apportent énormément. Ils ont toujours été là pour moi et je sais qu'ils le seront toujours, je les en remercie. Et enfin la plus belle : Caro, avec qui je partage tout (enfin presque... non je ne suis pas radin !!). Je lui dédie d'ailleurs cette thèse ainsi qu'à mes parents.

J'allais oublier... Je voudrais aussi remercier les deux entreprises qui ont financées cette thèse : Presse+ puis TNS Media Intelligence et plus particulièrement Stéphane Gérard et Jean-Michel Vieillard avec qui j'ai vraiment apprécié travailler.

Enfin, un dernier merci à tout ceux que j'aurais maladroitement oublier...

« C'est l'inconnu qui m'attire. Quand je vois un écheveau bien enchevêtré, je me dis qu'il serait bien de trouver un fil conducteur. » Pierre Gilles de Gennes

Table des matières

[1] *Zero Crossing Rate* en anglais

[2] *Multi-Layer Perceptron* (MLP) en anglais

Table des figures

Liste des tableaux

Introduction générale

"Encore un mot juste une parole
Parole, parole, parole
Ecoute-moi.
Parole, parole, parole
Je t'en prie.
Parole, parole, parole
Je te jure.
Parole, parole, parole, parole, parole
encore des paroles que tu sèmes au vent..."

Dalida & Alain Delon (Paroles, paroles)

Communiquer, que se soit en parlant, en criant, en chantant, pour donner des informations pertinentes ou inutiles, est une action naturelle et essentielle chez l'homme. L'homme ne peut pas s'empêcher de communiquer. On le constate aujourd'hui avec l'explosion des marchés des téléphones portables et d'Internet. Nous sommes arrivés à l'ère de la communication, à une époque où une personne privée de moyen de communication se sent perdue. Et c'est naturellement que, dès la naissance des premiers systèmes mécaniques, l'homme a rêvé de pouvoir un jour communiquer avec eux. Aujourd'hui, l'homme a réussi à atteindre ce rêve. Il peut enfin communiquer avec les machines qu'il construit. Au départ mécanique en utilisant des cartes perforées puis électronique par l'intermédiaire d'un clavier, la communication avec la machine se fait désormais par la parole. L'homme peut ainsi manipuler son environnement, c'est-à-dire les machines qui l'entourent, uniquement avec sa voix. Ce qui n'était encore qu'un fantasme il y a un peu plus d'un demi-siècle est devenu une réalité grâce aux systèmes de reconnaissance automatique de la parole. Ces systèmes, souvent couplés à des systèmes de synthèse vocale pour que la communication soit totale, sont de plus en plus présents dans notre vie. On les retrouve dans nos téléphones, nos voitures, à des guichets, etc. Ces machines communicantes sont devenues indispensables. Si pour nous elles permettent d'avoir un interlocuteur toujours disponible à n'importe quel moment, elles permettent aussi de faciliter la vie aux utilisateurs handicapés, comme exécuter des commandes par la voix, mais également aux personnes inexpérimentées face aux nouvelles technologies. La parole reste le moyen le plus facile pour communiquer et

interagir avec son environnement.
Les avancées technologiques nous donnant la possibilité d'interagir avec les machines par la voix, ont été permises grâce aux nombreux travaux de recherche sur la reconnaissance vocale réalisés dès le milieu du XXème siècle. Ainsi, le premier système pouvant être considéré comme faisant de la reconnaissance vocale date de 1952. Ce système électronique développé par Davis, Biddulph, et Balashek aux laboratoires Bell Labs [Davis 52] était essentiellement composé de relais et ses performances se limitaient à reconnaître des chiffres isolés. La recherche s'est ensuite considérablement accrue durant les années 70 avec l'utilisation croissante des ordinateurs. C'est d'ailleurs à la fin des années 70 qu'apparaît la première génération de systèmes commercialisés. Aujourd'hui, avec les possibilités sans cesse croissantes de l'informatique et de l'électronique et les nombreux travaux de recherches dans le domaine, les systèmes de reconnaissance automatique de la parole sont devenus de plus en plus performants. Cependant, nous sommes encore loin des performances humaines.

Depuis quelques années, les systèmes de reconnaissance automatique de la parole intéressent les sociétés de veille plurimédia. Le travail de ces sociétés consiste à récolter pour leurs clients toutes sources d'informations les concernant et à retranscrire tout ou parties d'émissions télévisées ou radiophoniques. Pour savoir ce que recherchent leurs clients, une liste de mots clés est définie. Cette liste évolue quotidiennement au rythme des événements nationaux ou internationaux et des actualités des clients. Ce travail colossal nécessite un grand nombre de personnes chargées de lire et écouter les différents médias, tâche qui s'avère éprouvante physiquement et psychologiquement pour ces pigistes. Aussi, face à l'explosion multimédia actuelle, que ce soit à la télévision, à la radio ou sur Internet (blog audio, radio internet, podcast, etc.), les sociétés de veilles sont désormais obligées de trouver des solutions leur permettant de continuer à couvrir un maximum de médias sans augmenter de manière drastique leur masse salariale, ce qui ne serait pas viable économiquement. La reconnaissance automatique de la parole apparaît être une bonne solution. En effet, en couplant un système de transcription automatique de la parole à un système de détection de mots clés, la veille continue de toute les informations audiovisuelles serait effectuée par des machines. Une personne serait tout de même chargé de contrôler les alarmes de détection de ces machines et de réécrire les transcriptions en corrigeant les erreurs commises.En effet, la transcription ne sera jamais parfaite. Améliorer les performances des systèmes de reconnaissance afin d'obtenir des transcriptions de meilleure qualité est l'un des objectifs des recherches actuelles en reconnaissance automatique de la parole. Une difficulté à surmonter lorsque l'on travaille dans le domaine de la veille audiovisuelles, est le problème du flux audio continu. En effet, la transcription d'un flux audio ininterrompu comporte son lot de difficultés. Outre le problème du temps réel qui peut être réglé en diminuant légèrement les performances du système de reconnaissance, il y a le problème de la nature du contenu du flux audio. En effet, celui-ci n'est pas uniquement composé de parole que l'on peut directement fournir au moteur de reconnaissance. Le flux

audio peut ainsi comporter :

- de la parole qui peut se trouver dans un environnement plus ou moins bruité ou en présence de musique,
- de la musique seule,
- du bruit ou du silence.

Le silence est généralement très rare en radio. A contrario, la musique y est omniprésente. La musique en fond sonore est ainsi un bon moyen de combler les moments de silence et permet d'éviter le malaise ou l'ennui que pourrait engendrer de trop longs silences. La parole se retrouve aujourd'hui noyée dans un magma de musique et de bruit. La segmentation du flux audio est donc nécessaire et doit être l'une des premières étapes dans un système de reconnaissance automatique de la parole. Cette tâche doit permettre à la machine de faire la différence entre la parole, la musique et le bruit. Mais elle ne s'arrête pas seulement à cela. En effet, l'étape de segmentation peut également découper le flux suivant :

- le canal de transmission : le locuteur peut être au téléphone ou dans un studio d'enregistrement.
- le genre du locuteur : homme/femme.
- le locuteur : il est possible de découper le signal en tour de parole, c'est à dire séparer les différents locuteurs.
- les respirations et les silences : ces indicateurs permettent de découper les segments de parole en plus petits segments (en phrases) afin d'être exploitables par le moteur de reconnaissance.

En ce qui nous concerne, nous nous intéressons seulement dans nos recherches à la séparation parole/musique. Cette étape est très importante au sein d'un système de transcription automatique d'émissions radiophoniques. En effet, le passage au moteur de reconnaissance de tout événement sonore n'étant pas de la parole peut provoquer des erreurs de transcription, faisant ainsi baisser les performances du système.

Dans cette thèse nous nous sommes donc intéressés à la segmentation parole/musique. Ce travail sur la segmentation parole/musique a été réalisé dans le cadre d'une application de détection de mots clés. La segmentation parole/musique fait ici partie d'un ensemble de pré-traitements du flux audio en amont d'un système de transcription automatique de la parole. Mais elle peut être utilisée dans de nombreuses autres applications. Parmi ces applications, nous pouvons citer :

- l'indexation de documents multimédia : l'indexation permet de retrouver rapidement et facilement des zones musicales ou de la parole dans de grandes bases de données

multimédias.
- l'aide au sous-titrage : séparer la parole de la musique peut permettre à la personne sous-titrant des documents audiovisuels d'aller plus vite. En effet, elle pourra facilement passer d'une zone de parole à une autre et éviter ainsi toutes les zones sans discours qu'elle aurait dû écouter.

Cette thèse comporte 7 chapitres.

Le chapitre 1 présente le problème de la segmentation parole/musique sur lequel notre travail s'est porté durant cette thèse. Un état de l'art des différentes paramétrisations et méthodes de classification utilisées dans le domaine est alors présenté.

Le chapitre 2 brosse un aperçu des méthodes de classification et des paramètres pour la segmentation parole/musique mis en place dans des systèmes complets de transcription automatique. Cet aperçu nous permet de nous positionner par rapport aux systèmes existants et d'apporter une approche originale dépassant les méthodes classiquement utilisées.

Le chapitre 3 concerne nos travaux dans le domaine de la segmentation parole/musique. Nous commençons par présenter ce sur quoi tous nos travaux sont basés : les ondelettes. Après avoir introduit les ondelettes, nous décrivons la méthode pour obtenir notre nouvelle paramétrisation. Celle-ci est basée sur une décomposition multi-échelles du signal à l'aide d'ondelettes et sur le calcul de différentes énergies.

Le chapitre 4, quant à lui, présente les différents corpus utilisés lors des expérimentations présentées au chapitre suivant. Nous y présentons les corpus d'apprentissage, de développement et de validation.

Vient alors le chapitre 5 au début duquel, les détails sur notre système de segmentation sont donnés. Ensuite, les différentes expériences sur les corpus de développement et de validations sont présentées. Ce chapitre se termine par des essais sur la combinaison de paramètres et sur une fusion de classifieurs par vote majoritaire.

Dans le chapitre 6, nous présentons notre système de transcription et de détection de mots-clés réalisé dans le cadre d'une convention CIFRE en partenariat avec l'entreprise TNS Média Intelligence.

Nous terminons ce manuscrit avec le chapitre 7. Ce chapitre est la conclusion générale de notre travail sur la discrimination parole/musique. Nous résumons les résultats importants obtenus au cours des diverses expérimentations. Enfin, nous présentons nos perspectives concernant la segmentation parole/musique et concernant les améliorations à apporter à notre système de transcription et de détection de mots clés dans les émissions radiophoniques.

1

La segmentation parole/musique

Sommaire

[3] *Zero Crossing Rate* en anglais

1.1 Position du problème

1.1.1 Introduction

Le traitement rapide, voir en temps réel, de l'information est devenu capital dans le monde d'aujourd'hui. La multiplication des sources d'information : radio, télévision, Internet (podcast, webradio) et la volonté d'analyser le contenu de tous ces médias nécessitent une automatisation maximum de la chaîne de traitement de l'information. L'analyse et le traitement des flux audio, que ce soit dans un but de transcription, d'indexation ou encore de détection de mots clés, comportent nécessairement une étape de segmentation du flux audio. En effet, un flux audio est constitué de nombreuses composantes acoustiques. Ces composantes peuvent être regroupées en trois grandes catégories : la parole, la musique et le bruit. Le bruit désigne ici tout ce qui ne correspond pas à de la parole, de la musique ou de la parole sur de la musique. La segmentation d'un signal consiste pour nous à découper le signal suivant les composantes acoustiques que nous avons décrites : parole, musique, bruit. Afin de pouvoir séparer ces composantes, il nous faut tout d'abord définir ce qui caractérise la parole, la musique et le bruit.

[4] *Multi-Layer Perceptron* (MLP) en anglais

1.1.2 La parole

Un signal de parole, au delà du déplacement d'air et de l'onde acoustique, peut être vu comme une suite de "sons élémentaires" aussi appelés phonèmes. Cette suite de phonèmes est marquée par des transitions plus ou moins franches comme nous pouvons l'observer sur le spectrogramme de la figure 1.1. Pour rappel, un spectrogramme est une représentation temps-fréquences-énergie d'un signal : l'axe des abscisses représente le temps, l'axe des ordonnées les fréquences et l'énergie du signal est représentée par les différents niveaux de couleur. Parmi les phonèmes, les voyelles jouent un rôle très important. En effet, la fréquence d'apparition des voyelles donne une sorte de rythme à la parole, c'est ce qu'on appelle la "fréquence syllabique". Une autre observation du signal de parole montre que son énergie n'est pas dispersée sur toutes les bandes de fréquence mais se concentre dans certaines bandes de fréquence. L'énergie se retrouve donc généralement dans les basses fréquences (correspondant aux formants) et dans les hautes fréquences lorsque le son émis correspond à une "fricative" comme les sons "s" ou "f". Le spectrogramme de la figure 1.1 illustre ces caractéristiques de la parole, les transitions entre les phonèmes, les concentrations d'énergie dans certaines bandes de fréquence.

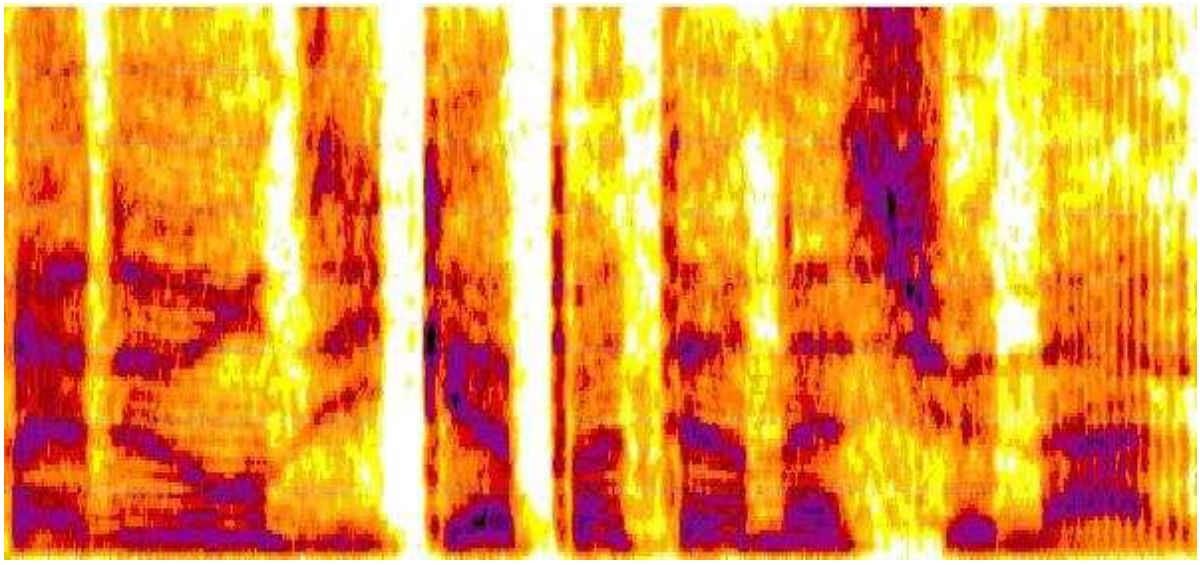

FIG. 1.1 – Spectrogramme d'un signal de parole

Enfin, la parole n'est pas seulement une suite de phonèmes, mais aussi une suite de mots séparés par des silences, plus ou moins longs suivant le débit de parole du locuteur. Ces silences sont aussi caractéristiques d'un signal de parole car ils ne se retrouvent pas dans un signal de musique.

1.1.3 La musique

La musique est un art qui semble avoir toujours existé. Elle peut être très évoluée comme la musique contemporaine, la musique classique, le jazz, le rock, ou très basique comme des chants, des battements de mains, des chocs de pierres ou de morceaux de bois, etc. La musique est donc beaucoup plus difficile à caractériser que la parole. Certains ce sont essayés à définir la musique à différents niveaux : anthropocentrique, philosophique, esthétique, etc. Mais nous nous intéressons ici à un plus bas niveau, au signal lui-même, et ce que nous pouvons dire avec certitude, c'est qu'au niveau spectral, la musique possède des harmoniques bien visibles (cf. Figure 1.2). Les harmoniques [Moore 92, Huang 01] sont des sons élémentaires dont les fréquences sont des multiples entiers de la fréquence fondamentale f_0. Par exemple, le son produit par un instrument à vent contient de nombreuses harmoniques naturelles, alors que certains instruments comme les percussions émettent des fréquences inharmoniques, elles sont des multiples non entiers de la fréquence fondamentale. Le timbre d'un instrument est ainsi révélé par l'ensemble des fréquences de son spectre harmonique. Un autre élément distinctif de la musique, c'est le rythme (le tempo) qui est beaucoup plus rapide que celui de la parole, c'est-à-dire bien supérieur à la fréquence syllabique. Il est rare d'observer des silences et des transitions marquées dans un signal de musique. De plus, au niveau énergétique, nous pouvons constater que, dans un signal de musique, l'énergie est plus uniformément distribuée dans les bandes de fréquences au fil du temps (cf. Figure 1.2).

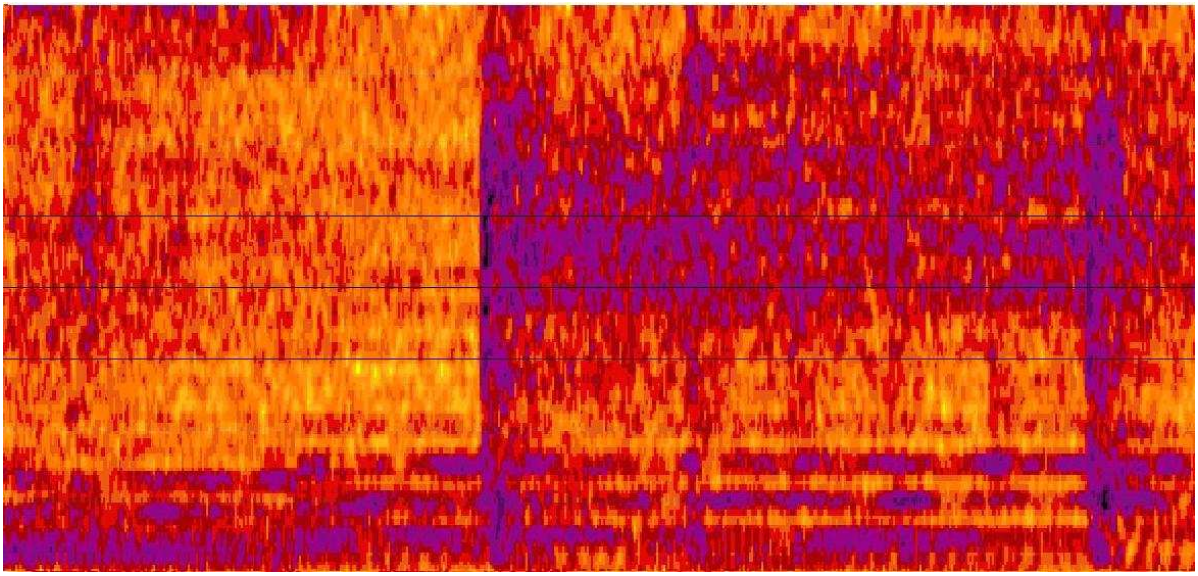

FIG. 1.2 – Spectrogramme d'un signal de musique

Finalement, au niveau spectral et temporel (figures 1.1 et 1.2) des différences apparaissent déjà entre la parole et la musique.

1.1.4 Le bruit

En acoustique et plus particulièrement dans le domaine de la parole, le bruit est caractérisé par tout ce qui peut gêner lors de la tâche de reconnaissance ou de discrimination par exemple. On ne parlera pas des différents bruits existants, comme le bruit blanc, rose, etc., mais plutôt du bruit en général qui pour nous représente tout ce qui n'est ni de la parole, ni de la musique. Le bruit est très gênant car il peut masquer les caractéristiques spécifiques de la parole ou de la musique (cf. Figure 1.3). En effet, l'addition de bruits extérieurs ou encore l'ajout de bruit dû au canal de transmission (téléphone, par exemple), amenant une détérioration du signal, rendent plus difficile la séparation entre parole et musique.

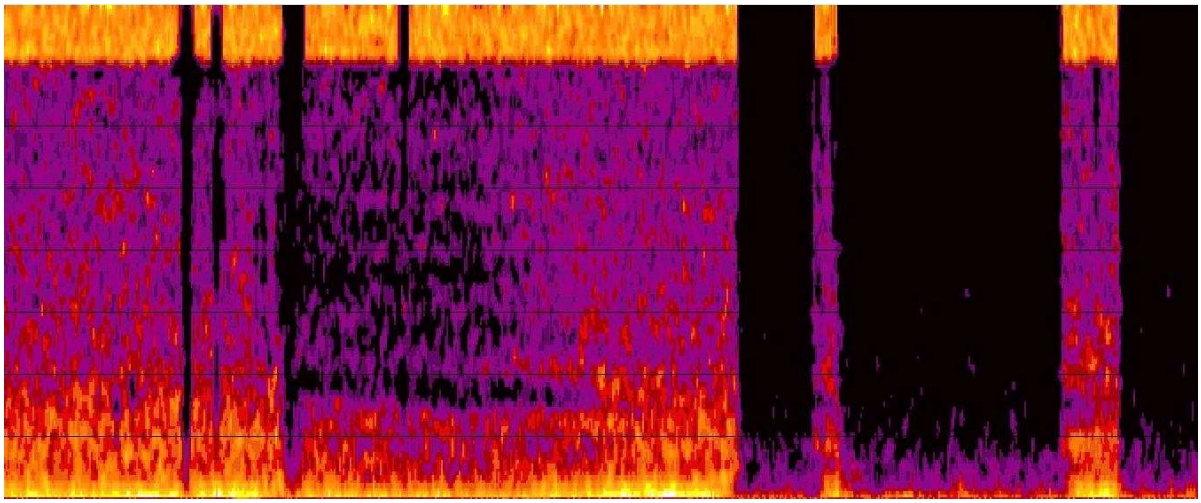

FIG. 1.3 – Spectrogramme d'un signal de parole très bruité. La parole est inaudible.

Pour finir sur le bruit, la parole aussi peut être considérée comme du bruit. Ainsi, lorsqu'une personne ou une foule (effet "cocktail party") couvre par leur voix le locuteur que l'on écoute, il devient difficile de reconnaître ce qui a été prononcé.

Nous avons défini trois grandes catégories de son : parole, musique et bruit. Malheureusement ces catégories ne sont pas totalement disjointes. Nous pouvons avoir de la parole ou de la musique bruitée ou encore de la parole sur fond musical. Tout ceci implique des difficultés supplémentaires pour différencier la parole de la musique. En effet, dans des enregistrements radiophoniques, il est très fréquent d'avoir des chansons, de la parole sur un fond musical, de la parole bruitée, etc. Il est donc intéressant de définir une catégorie supplémentaire pour ces segments qui sont un mélange des trois catégories que nous avons défini. Cette classe qui correspond aux segments pouvant être à la fois de la parole et de la musique, nous l'appellerons dans la suite du manuscrit "parole sur fond musical", "parole sur musique" ou encore "parole en présence de musique".

1.1.5 La parole sur fond musical

Il est très fréquent, dans des enregistrements radio, d'avoir de la parole sur un fond musical. Ce fond musical se retrouve dans des jingles, des émissions musicales et même parfois dans des émissions ou des flashs d'information. Le spectre de la parole ainsi "parasité" (figure 1.4) nous rend plus difficile la décision de dire si le segment correspond à de la parole ou à de la musique. De plus, les effets de fondu (*fade-in fade-out*[5]) apportent une difficulté supplémentaire dans la détection des frontières entre la parole, la musique et la parole sur musique. Bien qu'un être humain fasse souvent totalement abstraction de ce bruit de fond et reconnaisse parfaitement ce qui est prononcé, une machine, quant à elle, a besoin de modèles pour reconnaître les sons prononcés.

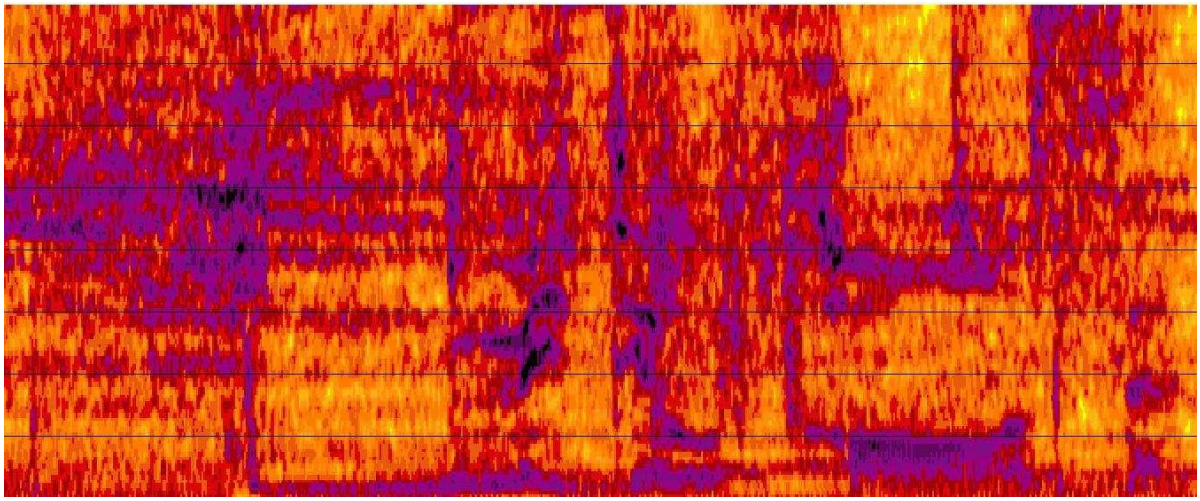

FIG. 1.4 – Spectrogramme d'un signal de parole sur fond musical

L'intérêt de pouvoir distinguer le cas de la parole sur fond musical apparaît tout de suite. En détectant les segments de parole sur musique, nous pourrions utiliser des modèles de parole adaptés, c'est-à-dire appris dans un contexte musical. Les performances de la transcription, succédant en général à l'étape de segmentation, seraient ainsi améliorées.

1.1.6 Quelques applications

Il existe un grand nombre d'applications nécessitant une phase préliminaire de segmentation du signal en parole et musique. Parmi celles-ci nous pouvons citer :

– La transcription automatique de flux audio :

[5]diminution ou augmentation progressive d'un son

La transcription automatique de flux audio consiste à reconnaître et à transcrire la suite de mots prononcés dans un document sonore. Au final, cette application permet de mettre par écrit tout ce qui est prononcé. C'est l'application la plus classique, et qui a fait l'objet de nombreuses campagnes d'évaluation, comme dernièrement la campagne ESTER[6] [Gravier 04] ou encore les nombreuses campagnes NIST[7]. L'intérêt de segmenter le document en parole/musique est clair. Cela permet de ne fournir en entrée du module de transcription que les segments contenant de la parole et de rejeter les segments inutiles comme la musique.

– L'indexation de documents multimédia :

  L'indexation consiste à découper virtuellement et à étiqueter un document pour en faciliter l'exploitation, et en particulier pour faciliter les recherches ultérieures dans ce document. Un document est donc indexé suivant la tâche que l'on voudra effectuer ultérieurement. Cela peut être la recherche de mots clés, de locuteurs, de morceaux de musique, etc. La tâche d'indexation est en plein essor. Ceci est dû à l'augmentation du nombre de média, notamment avec l'explosion du nombre de chaînes de télévision (TNT, satellite), de radios, particulièrement sur Internet (podcast, webradios, webtv, etc.). Le "WEB" est aujourd'hui devenu la plus grande source d'informations et son contenu multimédia est en pleine expansion. Par conséquent, le nombre de sources audio à traiter est lui-même en constante progression. La segmentation parole/musique va ici permettre de faire une première indexation en indiquant les zones correspondant à de la parole et celles correspondant à de la musique.

– L'aide au sous-titrage de films et d'émissions télévisées :

  Le sous-titrage consiste à afficher, au bas de l'écran, une traduction synchrone avec le dialogue de ce qui est prononcé durant un film, ou tout du moins, une transcription assez proche. Lors de la tâche de sous-titrage, nous pouvons considérer que le texte des dialogues est connu et que le "sous-titreur" n'a plus qu'à placer les différents sous-titres sur les zones de parole de la vidéo. La segmentation préliminaire en parole et musique du document audio-visuel permet un gain de temps au "sous-titreur" en lui permettant de passer d'une zone de parole à une autre plus rapidement. En France, avec la multiplication des chaînes de télévision et l'apparition de supports nouveaux comme le DVD, la demande en sous-titrage s'est beaucoup développée au cours des dernières années. De plus, le sous-titrage pour les sourds et les malentendants est une obligation inscrite aux cahiers des charges des chaînes publiques depuis 1984.

[6]Evaluation des Systèmes de Transcription d'Emissions Radiophoniques
[7]*National Institute of Standards and Technology*

Une loi de 2005 prévoit même que dans un délai maximum de cinq ans, les chaînes dont l'audience moyenne annuelle dépasse 2.5% de l'audience totale des services de télévision devront rendre la totalité de leurs programmes accessibles aux personnes sourdes et malentendantes. Cependant, nous sommes encore loin de cet objectif et la proportion actuelle de programmes sous-titrés, très variable d'une chaîne à l'autre, reste globalement en dessous de 30% d'émissions sous-titrées. L'aide au sous-titrage va donc devenir une application majeure dans les années à venir.

1.1.7 Premières conclusions

Pour différencier la parole de la musique, nous devons trouver les caractéristiques les plus discriminantes de la parole et de la musique. Cela peut se faire à différents niveaux :

- Au niveau de la paramétrisation du signal :

  La paramétrisation du signal consiste à extraire l'information pertinente du signal acoustique dans le but d'en fournir une description aussi compacte et représentative que possible. De nombreux travaux ont été réalisés sur la paramétrisation du signal dans le but de mieux distinguer la parole de la musique et même de distinguer la parole sur fond musical. Les paramètres étudiés peuvent être classés en trois grandes catégories : paramètres temporels, fréquentiels et enfin mixtes. Nous reviendrons plus en détails sur ces paramètres dans la section suivante. Les paramètres, extraits à intervalles réguliers, forment ce que nous appelons des trames acoustiques ou vecteurs d'observations.

- Au niveau de la classification :

  Les trames acoustiques extraites dans une étape préliminaire de paramétrisation du signal doivent être classées en différentes catégories. Cela nécessite de définir une méthode de classification (HMM, kNN, réseau de neurones...). Le choix de la méthode de classification en fonction des paramètres considérés peut aussi avoir une influence sur les performances du système de segmentation Parole/Musique.

Nous allons voir plus en détails les deux grandes étapes habituelles en segmentation parole/musique : la paramétrisation du signal et la classification.

1.2 La paramétrisation

La segmentation en parole et musique nécessite une représentation discriminante et compacte du signal. De nombreux descripteurs ont été étudiés , la plupart sont issus de l'analyse du signal et des études dans le domaine de la reconnaissance de la parole.
On peut classer les différents paramètres suivant leur mode de calcul, c'est-à-dire suivant la caractéristique du signal utilisée : caractéristique temporelle, fréquentielle, mixte ou encore psycho-acoustique.

1.2.1 Les paramètres temporels

Ils sont calculés en prenant en compte les variations temporelles des caractéristiques extraites du signal acoustique.

1.2.1.1 Le taux de passage par zéro (ZCR[8])

Le taux de passage par zéro [Saunders 96] représente le nombre de fois que le signal, dans sa représentation amplitude/temps, passe par la valeur centrale de l'amplitude (généralement zéro) sur une fenêtre temporelle. Il peut être défini comme suit :

$$ZCR(i) = \frac{1}{2N}\left(\sum_{n=1}^{N} sign(x_n(i)) - sign(x_{n-1}(i))\right) \tag{1.1}$$

avec $x_n(i)$ le n^{ieme} échantillon de la trame i et N le nombre total d'échantillons de la trame i.
La valeur du ZCR donne principalement une idée sur le voisement du signal de parole. Le voisement est une propriété de certains sons de la parole. Un son est dit voisé si sa production s'accompagne d'une vibration des cordes vocales, et non voisé dans le cas contraire. Le ZCR aura une valeur élevée dans les zones non-voisées et faible dans les zones voisées. Un signal de parole étant constitué d'une alternance de sons voisés et non-voisés, le ZCR a donc une variation importante. Inversement la musique n'ayant pas cette alternance, le ZCR n'aura pas une grande variation au cours du temps. C'est donc la variation du ZCR qui est fréquemment utilisée en classification parole/musique [Zhang 98, Scheirer 97].

1.2.1.2 La mesure de rythmicité (*Pulse Metric*)

Cette mesure, décrite dans [Scheirer 97], est basée sur une méthode d'autocorrélation à long terme et permet d'évaluer la "rythmicité" d'un flux audio dans une fenêtre de 5 secondes. Elle fonctionne très bien pour des musiques ayant un tempo (une pulsation) bien marqué, comme la techno, la salsa, le rock. Au contraire, on ne peut pas déterminer

[8] *Zero Crossing Rate* en anglais

la mesure de rythmicité pour un signal ayant des changements de tempo ou ayant un rubato, c'est-à-dire ayant un ralentissement ou une accélération du rythme.
La mesure de rythmicité est basée sur le fait qu'un tempo marqué induit une variation dans toutes les bandes de fréquences d'un signal. Ainsi, quelque soit la bande de fréquences observée, le même tempo (rythme) apparaît. L'algorithme permettant de calculer cette mesure divise d'abord le signal en six bandes de fréquences puis cherche les pics dans les enveloppes de chacune des bandes. Une enveloppe correspond à l'évolution de l'amplitude du signal au cours du temps. Ensuite, par une méthode d'autocorrélation, on peut extraire pour chaque bande une pulsation potentielle. Plus on retrouvera les mêmes pics dans les différentes bandes de fréquence, plus la mesure de rythmicité aura une valeur élevée. Ainsi, la musique ayant un rythme très présent et régulier aura une valeur élevée pour sa mesure de rythmicité contrairement à un signal de parole où la notion de rythme est beaucoup moins présente.

1.2.1.3 Le pourcentage de trames de faible énergie

Saunders [Saunders 96] fût le premier à montrer que le contour énergétique d'un signal audio était capable de discriminer la parole et la musique. Ainsi, le contour ne varie presque pas pendant plusieurs secondes pour un signal de musique alors que pour un signal de parole, ce contour subit des variations dûes à l'alternance entre voyelles et consonnes caractéristique de la parole.
Scheirer et Slaney [Scheirer 97] utilisent cette propriété pour définir un nouveau paramètre : le pourcentage de trames de faible énergie. Il correspond au nombre de trames dont la puissance RMS est inférieure à 5O% de la puissance RMS[9] moyenne dans une fenêtre d'une seconde. Du fait que la parole a une proportion importante de trames peu intenses comparée à la musique, cette mesure aura une valeur élevée en présence d'un signal de parole. Ce paramètre est fréquemment utilisé en discrimination parole/musique, par exemple [Lu 01] ou [Wang 03] qui proposent d'utiliser une version modifiée du pourcentage de trames de faible énergie.

1.2.1.4 L'entropie moyenne par trame

Ajmeira, McCowan et Bourlard [Ajmera 02, Ajmera 03] utilisent des paramètres basés sur des probabilités *a posteriori*. En effet, ils utilisent un réseau de neurones qui, à partir des vecteurs acoustiques d'entrée, génère des estimateurs de probabilités *a posteriori* de phonèmes. C'est à partir de ces probabilités que l'entropie est estimée. Ils définissent ainsi

[9]La puissance RMS (*Root Mean Square*) ou puissance efficace désigne la moyenne quadratique de la puissance instantanée.

l'entropie moyenne à la trame n :

$$H_n = \frac{1}{N} \sum_{t=n-N/2}^{n+N/2} h_t \tag{1.2}$$

où n correspond à l'indice de la trame acoustique courante, N à la taille de la fenêtre de moyennage et h_t à *l'entropie instantanée* à la trame t.
Soit q_k, $k = 1, \ldots, K$ les K classes correspondant aux différents phonèmes de la parole, l'entropie instantanée à la trame t est définie comme suit :

$$h_t = -\sum_{k=1}^{K} P(q_k|x_t) \log_2 P(q_k|x_t) \tag{1.3}$$

x_t représente le vecteur acoustique au temps t, et $P(q_k|x_t)$ la probabilité *a posteriori* de la classe (phonème) q_k étant donné x_t en entrée.
L'entropie à la sortie du réseau de neurones, i.e. l'entropie des probabilités *a posteriori* de phonèmes, sera en moyenne plus élevée pour des segments de non-parole (musique, bruit, etc.) que pour des segments de parole. En d'autres termes, l'entropie, qui est une mesure de "désordre" d'une distribution, sera faible quand les phonèmes seront bien reconnus (segments de parole). Par contre, dans la situation où plusieurs classes se répartissent presque équitablement les probabilités (segments de non-parole), l'entropie sera plus élevée.

1.2.2 Les paramètres fréquentiels

Les paramètres fréquentiels sont calculés à partir de la DSP (Densité Spectrale de Puissance) du signal. La DSP d'un signal est issue de la transformée de Fourier de la fonction d'autocorrélation. Ces paramètres caractérisent l'enveloppe spectrale du signal. Ils permettent ainsi de capter le contenu fréquentiel du signal à un moment donné, comme par exemple les formants, les harmoniques, etc.

1.2.2.1 Le centre de gravité spectral (*Spectral Centroïd*)

Le centre de gravité spectral ou centroïde spectral (SC) [Peeters 03] est le centre de gravité de la distribution d'énergie du spectre de magnitude de la trame de signal. On peut le considérer comme le point d'équilibre du spectre. La hauteur spectrale donnée par la position du centre de gravité des composantes du spectre définit ce que l'on appelle la *brillance* du son. Le centre de gravité spectral peut être calculé de la manière suivante :

$$SC = \frac{\sum_k f_k \; a_k}{\sum_k a_k} \tag{1.4}$$

où a_k correspond à l'amplitude de la composante spectrale de fréquence f_k.
Le centre de gravité spectral devrait être plus élevé et plus constant pour la musique que pour la parole, à cause de l'alternance de sons voisés et non-voisés dans un signal de parole.

1.2.2.2 La variation de l'amplitude du spectre (*Delta Spectrum Magnitude* ou *Spectral «flux»*)

La variation du spectre aussi appelée "flux spectral" (SF) est une mesure permettant de voir avec quelle rapidité le spectre de puissance d'un signal varie au cours du temps. Cette mesure est calculée à partir de la corrélation croisée normalisée entre deux amplitudes successives du spectre $a_k(t-1)$ et $a_k(t)$.

$$SF = 1 - \frac{\sum_k a_k(t-1).a_k(t)}{\sqrt{\sum_k a_k(t-1)^2}\sqrt{\sum_k a_k(t)^2}} \tag{1.5}$$

En d'autres termes, la valeur SF correspond à la différence d'amplitude du vecteur spectral entre deux trames successives. Elle est proche de 0 si les spectres successifs sont similaires, et de 1 pour des spectres successifs très différents. Cette valeur est élevée pour la musique, car la musique varie fortement d'une trame à l'autre. Pour la parole, avec l'alternance de périodes de stabilité (voyelle) et de transitions (consonne-voyelle), cette mesure prend des valeurs très différentes et varie fortement au cours d'une phrase [Carey 99, Scheirer 97].

1.2.2.3 Le point spectral de coupure (*Spectral Rolloff Point*)

Le point spectral de coupure ou "*Spectral Rolloff Point*" [Scheirer 97] est la position de l'échantillon fréquentiel en dessous duquel est contenu 95% de la puissance du spectre (en général, ce pourcentage est un paramètre à fixer). La figure 1.5 illustre ceci.

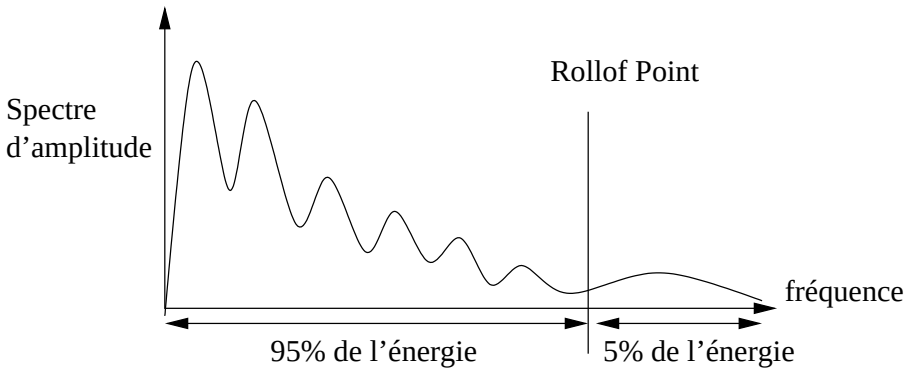

FIG. 1.5 – Définition du "*Spectral Rollof Point*"

Cette mesure permet de caractériser l'aternance parole voisée, parole non voisée. En effet, l'énergie est concentrée dans les basses fréquences pour les voyelles, d'où une petite

valeur du point spectral de coupure, alors qu'elle se situe plus dans les hautes fréquences pour les fricatives et nous avons par conséquent une valeur plus élevée du point spectral de coupure. Pour la musique, dont l'énergie est plus uniformément répartie sur toutes les bandes de fréquences, cette mesure ne varie que très faiblement.

1.2.3 Les paramètres mixtes (temps-fréquence)

Ces paramètres sont issus d'analyses du signal à la fois dans le domaine temporel et dans le domaine fréquentiel. Ils permettent ainsi de lier les deux caractéristiques importantes d'un signal audio : le temps et la fréquence. Ainsi, les variations temporelles et fréquentielles du signal peuvent être analysées en même temps.

1.2.3.1 La modulation de l'énergie à 4Hz

L'origine de ce paramètre vient de l'observation [Houtgast 85] d'un pic caractéristique de modulation d'énergie autour de 4Hz pour un signal de parole. Cette fréquence de modulation correspond au rythme syllabique. En effet, sous l'hypothèse qu'une syllabe soit la combinaison d'une zone de faible énergie (consonne) et d'une zone de forte énergie (voyelle), les changements de syllabes (variations d'énergie) pour la parole vont se trouver aux alentours de cette fréquence de 4Hz. Nous allons présenter ici brièvement la procédure d'extraction de ce paramètre [Pinquier 04a] :

- Découpage du signal d'entrée en trames de 16ms sans recouvrement.
- Pour chaque trame, après avoir appliqué une fenêtre de Hamming, 40 coefficients spectraux sont extraits suivant l'échelle Mel. Ces 40 coefficients correspondent à 40 bandes de fréquences.
- L'énergie de chaque bande est ensuite filtrée à l'aide d'un filtre FIR[10] passe-bande de fréquence centrale 4Hz.
- Les énergies filtrées des 40 bandes sont sommées et normalisées par l'énergie moyenne.
- Finalement, la modulation de l'énergie à 4Hz est obtenue en calculant la variance de l'énergie filtrée, sur une seconde de signal.

D'après sa définition, ce paramètre aura une valeur élevée pour un signal de parole alors qu'il aura une faible valeur pour de la musique qui, *a priori*, n'a pas cette modulation d'énergie spécifique à la fréquence syllabique de la parole (cf. Figure 1.6).

1.2.3.2 Caractéristiques basées sur les modulations basse fréquence : LF-MAD, 4Hz ASD

Ce paramètre se base, comme la modulation de l'énergie à 4Hz, sur le fait que la modulation d'amplitude d'un signal de parole suit le rythme syllabique. Mais contrairement

[10] FIR (RIF en français) signifie "Réponse Impulsionnelle Finie"

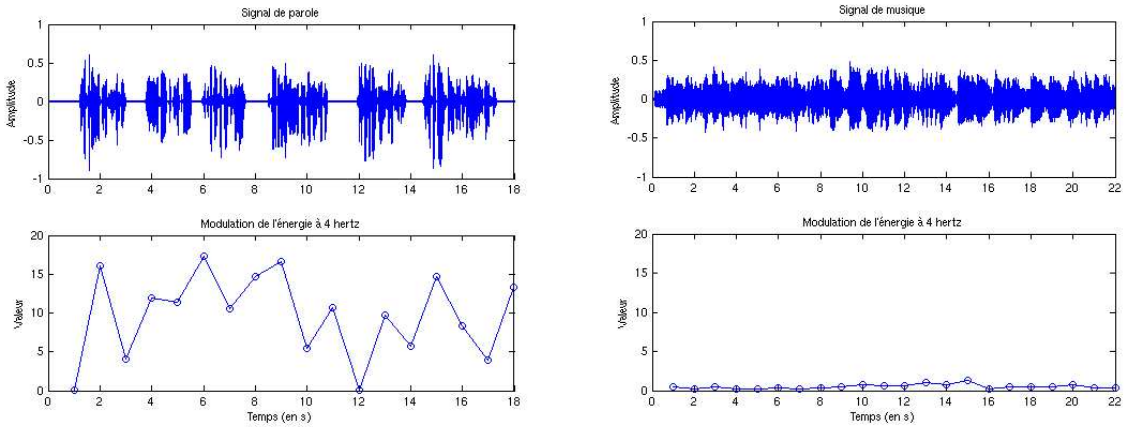

FIG. 1.6 – Modulation de l'énergie à 4Hz pour un signal de parole (à gauche) et pour un signal de musique (à droite) (image extraite de la thèse de Pinquier [Pinquier 04a])

à la modulation de l'énergie à 4Hz qui est uni-dimensionnel, le paramètre "4Hz ASD" (*4Hz Amplitude and Standard Deviation*) de Karnebäck [Karneback 01] est multi-dimensionnel. Ce paramètre, en plus de la modulation à 4Hz, prend en compte les corrélations de modulation à 4Hz entre les différentes bandes et la variance dans chaque bande pour montrer les différents comportements de la musique et de la parole. Il permet ainsi d'extraire d'autres informations concernant la parole et la musique telles que :

- La parole est plus régulière que la musique d'un point de vue des corrélations.
- Les variations d'amplitude sont plus uniformément distribuées à travers les fréquences pour la musique que la parole.
- Pour un signal de parole, des bandes de fréquence adjacentes sont plus corrélées que des bandes distantes.
- Les variations de l'amplitude et de la variance sont plus grandes dans les fréquences moyennes que dans les hautes ou basses fréquences pour la parole.
- Une forte corrélation peut apparaître entre plusieurs bandes en présence d'un instrument seul (par exemple un piano), en supposant que plusieurs tons ou accords sont joués à la fois.
- Au contraire, il y aura une faible corrélation entre les bandes lorsqu'un orchestre accompagnera un instrument seul.

Ce paramètre a été comparé à la modulation d'énergie à 4Hz [Scheirer 97] et aux MFCC. Par rapport à la modulation d'énergie à 4Hz, il donne de meilleurs résultats en discrimination parole/musique, mais ses performances restent en deçà de celles des MFCC (surtout pour discriminer la musique) [Karneback 01, Karneback 02]. Cependant la combinaison de ce paramètre avec les paramètres MFCC permet d'obtenir des performances supérieures à celles obtenues avec l'emploi des paramètres MFCC seuls. Enfin, Karnebäck

[Karneback 02] a voulu aller plus loin dans l'étude des modulations basse fréquence et ne pas se limiter à une fréquence de 4Hz. Il a donc fait varier la fenêtre d'analyse de 37.5 millisecondes à 1 seconde ainsi que les fréquences analysées : de 2Hz à 27Hz. Il a montré que ce paramètre améliorait les performances en classification parole/musique en présence de bruit, comparativement aux MFCC.

1.2.3.3 Le Coefficient Harmonique

Le "Coefficient Harmonique" définit par [Chou 01] est basé sur une méthode utilisée initialement en codage de la parole pour faire une estimation précise du pitch [Cho 98]. Le pitch est la fréquence fondamentale perçue d'un son. Ce coefficient a été développé pour améliorer les performances en détection de voix chantées. Il permet de représenter les caractéristiques des structures harmoniques d'un signal de parole voisé. Il est calculé à partir d'une évaluation spectrale et temporelle de l'autocorrélation. Le détail des calculs est donné ci-dessous.

Soit $s_t(n)$ un signal de parole, l'autocorrélation temporelle pour le pitch candidat τ est définie par :

$$R^T(\tau) = \frac{\sum_{n=0}^{N-\tau-1}[\tilde{s}_t(n).\tilde{s}_t(n+\tau)]}{\sqrt{\sum_{n=0}^{N-\tau-1}\tilde{s}_t^2(n).\sum_{n=0}^{N-\tau-1}\tilde{s}_t^2(n+\tau)}} \tag{1.6}$$

où $\tilde{s}_t(n)$ est la version moyennée à 0 ("zero-mean") de $s_t(n)$ et N est le nombre d'échantillons pour l'analyse.

L'autocorrélation spectrale est donnée par l'équation suivante :

$$R^S(\tau) = \frac{\int_0^{\pi-\omega_\tau}\tilde{S}_f(\omega)\tilde{S}_f(\omega+\omega_\tau)d\omega}{\sqrt{\int_0^{\pi-\omega_\tau}\tilde{S}_f^2(\omega)\int_0^{\pi-\omega_\tau}\tilde{S}_f^2(\omega+\omega_\tau)d\omega}} \tag{1.7}$$

où $\omega_t = 2\pi/\tau$, $S_f(\omega)$ correspond au spectre de magnitude de $s_t(n)$, et $\tilde{S}_f(\omega)$ est la version moyennée à 0 de $S_f(\omega)$.
L'autocorrélation spectro-temporelle est obtenue en faisant une moyenne pondérée de l'autocorrélation temporelle et l'autocorrélation spectrale :

$$R(\tau) = \beta.R^T(\tau) + (1-\beta).R^S(\tau) \tag{1.8}$$

$\beta = 0.5$ est le poids donnant les meilleurs résultats après expérimentation [Cho 98].
Finalement, le coefficient harmonique H_a est défini par :

$$H_a = max_\tau R(\tau) \tag{1.9}$$

H_a aura une plus grande valeur pour des sons voisés que pour des sons non-voisés. Ce coefficient harmonique est associé dans [Chou 01] à la modulation de l'énergie à 4Hz pour

améliorer la discrimination parole/musique.

1.2.4 Les paramètres cepstraux

Les paramètres cepstraux sont largement utilisés dans le domaine de la reconnaissance de la parole depuis de nombreuses années. De récents travaux [Logan 00] ont montré que les paramètres cepstraux pouvaient également être utilisés dans le cadre de la discrimination parole/musique.

1.2.4.1 L'analyse cepstrale

Le signal de parole résulte de l'intermodulation entre un signal émis par une source et le conduit vocal [Haton 06]. La séparation source-conduit permet de voir les contributions du conduit (les fréquences formantiques) et de la source (la fréquence fondamentale f_0). Cependant, dans le domaine spectral, la séparation entre source et conduit est difficile. Le passage dans le domaine log-spectral, avec l'analyse spectrale [Rabiner 78], permet de surmonter cela et de déconvoluer le signal.
Le cepstre réel d'un signal numérique $s(n)$ est obtenu en prenant le logarithme de son spectre $S(f)$ auquel on applique ensuite une transformation de Fourier inverse (Figure 1.7).

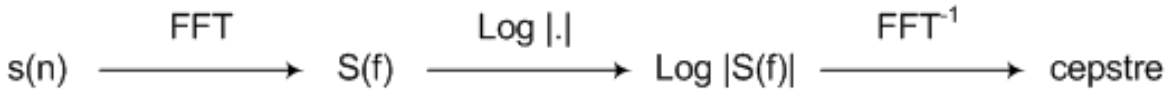

FIG. 1.7 – Principe de l'analyse cepstrale. FFT (*Fast Fourier Transform*) correspond à la transformée de Fourier rapide et FFT^{-1} correspond à la transformée de Fourier inverse.

En reconnaissance de la parole, l'analyse cepstrale n'est pas utilisée directement sous cette forme. Mais c'est une paramétrisation du signal dérivant de cette analyse cepstrale qui est largement répandue. Il s'agit de la paramétrisation MFCC que nous allons maintenant présenter.

1.2.4.2 Les MFCC

Les coefficients MFCC (*Mel Frequency Cepstrum Coefficients*) [Davis 80] ont été très largement utilisés en reconnaissance automatique de la parole et en identification du locuteur [Woodland 98, Gauvain 02], mais aussi en discrimination parole/musique [Scheirer 97, Carey 99, Williams 99, West 04]. Les coefficients MFCC c_i sont calculés en utilisant une échelle de fréquences non linéaire Mel, sous la forme d'un ensemble de filtres passe-bande triangulaires (cf. Figure 1.8). Cette échelle est une approximation de la perception de l'oreille humaine.

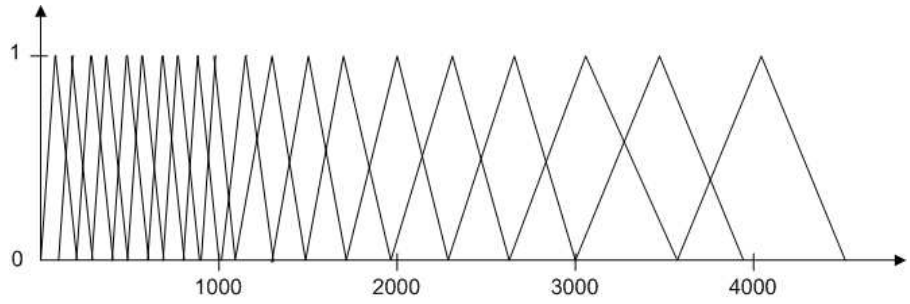

FIG. 1.8 – Banc de filtre à échelle Mel pour le calcul des coefficients MFCC

Les coefficients cepstraux sont calculés sur des trames successives du signal, selon la transformée en cosinus discrète (DCT) :

$$c_i = \sum_{j=0}^{M} S(j) cos(i(j - \frac{1}{2})\frac{\pi}{N_f}) \text{ pour } i = 0, 1, \ldots, M-1 \tag{1.10}$$

M correspond au nombre de coefficients cepstraux à calculer et N_f désigne le nombre de filtres. La figure 1.9 illustre les étapes successives nécessaires au calcul des coefficients MFCC. Il est intéressant de noter que l'on obtient des coefficients MFCC décorrélés grâce à la DCT. De plus, le premier coefficient est fortement corrélé avec l'énergie moyenne du signal de la trame. Il est ainsi souvent utilisé pour la représenter.

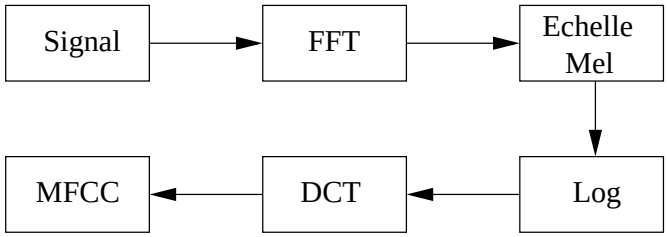

FIG. 1.9 – Schéma représentant les différentes étapes du calcul des coefficients MFCC. *FFT* (*Fast Fourier Transform*) correspond à la transformée de Fourier rapide et *DCT* (*Discrete Cosine Transform*) correspond à la transformée en cosinus discrète.

Pour finir, bien que les MFCC aient l'inconvénient d'être très difficile à interpréter, leur utilisation est quasi incontournable à l'heure actuelle dans le domaine de la parole.

1.2.4.3 les LFCC

Les LFCC (*Linear Frequencies Cepstrum Coefficients*), utilisés par Karnebäck en discrimination parole/musique [Karneback 04], sont une variante des MFCC. L'extraction des coefficients LFCC est similaire à celui des MFCC, seule l'échelle de fréquence diffère. L'échelle Mel est remplacée par une échelle de fréquence linéaire. La haute résolution dans les basses fréquences obtenue par les MFCC grâce à l'échelle Mel, pourra être obtenue en

utilisant un plus grand nombre de coefficients LFCC. Toutefois, à nombre égal de coefficients, les LFCC sont moins performants que les MFCC en discrimination parole/musique [Logan 00].

1.2.4.4 L'amplitude des résidus de resynthèse cepstrale

L'amplitude des résidus de resynthèse cepstrale [Scheirer 97] est définie comme la norme 2 ($||.||_2$) de la différence entre le spectre d'amplitude à une trame donnée et le spectre d'amplitude reconstruit après liftrage à cette même trame. Le liftrage correspond ici à un filtrage passe-bas dans le domaine cepstral. Ce liftrage donne un "spectre lissé". La différence entre le spectre d'origine et ce spectre "lissé" reconstruit après liftrage est à l'origine de ce paramètre.
Le principe de la méthode est le suivant [Rossignol 00] :

Tout d'abord, le cepstre réel $\hat{C}$ est calculé pour une trame de signal. Le cepstre réel a la particularité d'être symétrique par rapport à la quéfrence 0. Pour information, la "quéfrence" est l'équivalent de la fréquence dans le domaine cepstral.

$$\hat{C} = Partie\_reelle(FFT^{-1}(log(|\hat{S}|))) \tag{1.11}$$

Avec $|\hat{S}| = |FFT(x)|$ le spectre d'amplitude à la trame x.

Le cepstre filtré $\hat{C}'$ est ensuite obtenu par « liftrage ». Pour obtenir ce liftrage, nous conservons parmi les coefficients cepstraux obtenus :

- les n coefficients correspondant aux plus petites quéfrences positives,
- le coefficient correspondant à la quéfrence 0,
- et les n coefficients correspondant aux plus petites quéfrences négatives en valeur absolue.

La valeur des autres coefficients est mise à 0.
La transformée inverse est alors calculée pour obtenir le spectre d'amplitude reconstruit après liftrage $|\hat{S}'|$:

$$|\hat{S}'| = |exp(FFT(\hat{C}'))| \tag{1.12}$$

L'amplitude des résidus de resynthèse cepstrale est alors :

$$|||\hat{S}'| - |\hat{S}|||_2 \tag{1.13}$$

Dans le cas de sons voisés, les coefficients d'ordre élevé (au dessus de $|n|$) permettent de remonter, après transformée inverse, au train d'impulsions qui est émis par la source.

La périodicité de ces impulsions correspond à la période fondamentale.
Au contraire, les coefficients compris entre $-n$ et n permettent de revenir, après transformée inverse, à la réponse impulsionnelle du filtre (passage dans la gorge, entre les lèvres...). Ici, seule la partie correspondant à la réponse impulsionnelle du filtre (gorge, lèvres) est conservée. Ainsi, le spectre d'amplitude d'un bruit (son non voisé) est mieux "approximé" que le spectre d'amplitude d'un signal harmonique (son voisé).
De ce fait, ce paramètre variera pour un signal de parole (alternance de sons voisés et non voisés) mais restera relativement stable pour un signal de musique.

1.2.5 Autres paramètres et paramètres psychoacoustiques

D'autres caractéristiques, parmi lesquelles des caractéristiques dites psychoacoustiques ou perceptives, ont été étudiées afin de discriminer parole et musique. Nous pouvons ainsi citer, sans entrer dans les détails :

- Le pitch et sa variation [Carey 99].

- Le vibrato [Gerhard 02] :
  C'est une caractéristique du pitch permettant de distinguer la parole de la chanson.

- *Roughness* (Agitation), *Loudness* (Sonie), *Sharpness* (Netteté)[Breebaart 03] :
  "L'agitation" correspond à la perception des modulations de l'enveloppe temporelle dans l'intervalle de fréquence 20-150Hz et est maximale pour des modulations proche de 70Hz.
  "La sonie" est une mesure de la perception acoustique de l'intensité. Elle permet de mesurer l'intensité des sons telle qu'elle est perçue chez l'homme.
  "La netteté" est une perception relative à la densité spectrale et l'intensité relative de l'énergie dans les hautes fréquences.

- Les paramètres basés sur des modulations spectro-temporelles multi-échelles [Mesgarani 06] :
  Ici, l'ensemble des paramètres utilisés est inspiré des différentes étapes effectuées par le système auditif humain. Ils sont calculés en utilisant un modèle du cortex auditif qui va, étant donné un son, lui faire correspondre, dans un espace de plus grande dimension, une représentation de ces modulations spectro-temporelles. Contrairement aux paramètres conventionnels, ces nouveaux paramètres ont des résolutions spectrales et temporelles multi-échelles.

- Paramètres basés sur les probabilités *a posteriori* [Williams 99].

1.3 La classification

Cette section présente les méthodes statistiques pour la classification utilisées dans le cadre de la discrimination parole/musique. Nous ne parlerons que des méthodes avec apprentissage supervisé. Le but de la classification est d'assigner à une observation o une classe C_i, $i \in \{1, \ldots, N\}$ (N est le nombre de classes).
La décision de classification d'une observation peut être prise soit dans un cadre statistique en fonction des distributions de probabilité des différentes classes, soit dans un cadre géométrique à partir d'une partition de l'espace des observations en sous-régions correspondant aux différentes classes.

Nous pouvons ainsi distinguer deux catégories d'approches différentes pour la tâche de classification : les méthodes dites "génératives" et les méthodes dites "discriminantes".

1.3.1 Méthodes "génératives"

1.3.1.1 Description des méthodes "génératives"

Cette famille de méthodes se fonde sur la connaissance des distributions de probabilités des classes étudiées [Devijver 82].

Le processus de décision peut être formalisé de la manière suivante. Une observation o, représentée par un vecteur de d paramètres o_d, doit être classée dans une des N classes $C_1, C_2, \ldots, C_N$.
Le problème est donc le suivant : connaissant une observation o, quelle est la probabilité (ou densité de probabilité dans le cas continu) $p(C_i, o)$ pour que o appartienne à la classe C_i ? Un classifieur qui assigne o à la classe C_i en suivant la règle :

$$p(C_i|o) > p(C_j|o) \forall j \neq i$$

commet un nombre minimum d'erreurs de classification. Cette règle, aussi appelée la règle du maximum *a posteriori* (MAP), est optimale. En d'autres termes, aucune autre règle ne donnera un risque de mauvaise classification plus faible. On ne peut pas en général calculer directement $p(C_i|o)$. Il est cependant possible de l'estimer à partir des données d'apprentissage. Les algorithmes d'apprentissage vont permettre d'estimer seulement la probabilité $p(o|C_i)$ d'observer le vecteur o connaissant sa classe C_i. Le vecteur d'observation o appartenant à la classe C_i est considéré comme une observation tirée au hasard en fonction de la distribution de probabilités conditionnelles $p(o, C_i)$. C'est la formule de Bayes qui permet de calculer $p(C_i|o)$:

$$p(C_i|o) = \frac{p(o|C_i)p(C_i)}{p(o)} \tag{1.14}$$

$p(C_i|o)$ est appelée probabilité *a posteriori* et $p(o|C_i)$ la vraisemblance (aussi appelée *likelihood* en anglais). Enfin, $p(C_i)$ est la probabilité *a priori* de la classe C_i et $p(o)$ la

probabilité de l'observation o. Le terme "génératif" s'explique ainsi par le fait que la règle de décision est basée sur une modélisation de la probabilité $p(o|C_i)$ qui va "générer" les observations o pour une classe donnée C_i .

Ces méthodes utilisent les données d'apprentissage pour modéliser les densités de probabilité $p(o, C_i)$ de chaque classe par une famille de fonctions paramétriques. Aussi, lorsque cela est possible, ces méthodes peuvent tenir compte des probabilités *a priori* $p(C_i)$ d'apparition de chaque classe. Si nous ne pouvons pas connaître les probabilités *a priori* d'apparition des classes, il est d'usage de considérer les classes comme équiprobables *a priori*.

Nous allons maintenant présenter les méthodes génératives les plus utilisées en classification parole/musique : les mélanges de modèles gaussiens et les modèles de Markov Cachés.

1.3.1.2 Les mélanges de modèles gaussiens (GMMs : *Gaussian Mixture Models*)

Un GMM ou mélange de modèles gaussiens est une densité de probabilité vectorielle qui peut s'écrire sous la forme d'une combinaison linéaire positive de lois gaussiennes.
Les GMMs permettent de modéliser les fonctions de densité de probabilité des variables observées, en utilisant une densité de distribution gaussienne multivariée.
Soit o un vecteur d'observations de $\mathbb{R}^d$. La densité de probabilité associée s'écrit :

$$p(o|\theta) = \sum_{g=1}^{G} \omega_g \mathcal{N}(o|\mu_g, \Sigma_g) \tag{1.15}$$

$$\text{avec } \theta = \{\omega_g, \mu_g, \Sigma_g\}_g \in (\mathbb{R}^+ \times \mathbb{R}^d \times \mathbb{R}^{d^2}) \text{ et la contrainte } \sum_g \omega_g = 1.$$

θ correspond au jeu de paramètres à apprendre : les coefficients de pondérations ω_g, les moyennes μ_g et les matrices de covariances Σ_g des gaussiennes. En général, ces matrices de covariances sont souvent diagonales, ce qui a pour effet de réduire la complexité des modèles et les calculs. Aussi, la densité de probabilité associée à la loi gaussienne $\mathcal{N}(x|\mu, \Sigma)$ peut être formulée ainsi :

$$\mathcal{N}(x|\mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^d det\Sigma}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)} \tag{1.16}$$

L'apprentissage des GMMs se fait en général en réglant le jeu de paramètres θ à l'aide de l'algorithme *EM* (Expectation Maximization) [Dempster 77]. Le principe de l'algorithme *EM* est d'estimer et d'optimiser la vraisemblance des données d'apprentissage aux modèles de manière itérative, jusqu'à obtenir une certaine stationnarité.
L'algorithme *EM* se déroule comme suit. Soit N vecteurs d'apprentissage o_i $(i = 1 \dots N)$ et K gaussiennes.
Après une première étape consistant à initialiser les moyennes μ_g, les matrices de cova-

riance Σ_g et les poids ω_g, des phases d'estimation (*Expectation*) et d'optimisation (*Maximisation*) vont alterner jusqu'à ce que le critère d'arrêt défini soit vérifié.
L'étape d'estimation calcule pour chacune des gaussiennes g ($g = 1, .., K$) les probabilités que chaque vecteur d'observation o_i ait été généré par la gaussienne g :

$$p(g|o_i) = \frac{\omega_g \mathcal{N}(o_i|\mu_g, \Sigma_g)}{\sum_{l=1}^{K} w_l \mathcal{N}(o_i|\mu_l, \Sigma_l)}$$

Une fois l'étape d'estimation terminée, les différents paramètres sont réévalués lors de la phase dite de *Maximisation* :

$$\omega_g = \frac{1}{N} \sum_{i=1}^{N} p(g, o_i)$$

$$\mu_g = \frac{\sum_{i=1}^{N} (p(g,o_i) o_i}{\sum_{i=1}^{N} p(g,o_i)}$$

$$\Sigma_g = \frac{\sum_{i=1}^{N} (p(g,o_i)(o_i - \mu_g)(o_i - \mu_g)^T)}{\sum_{i=1}^{N} p(g,o_i)}$$

Cette alternance de phases d'estimation et de réestimation se termine soit lorsque le nombre maximum d'itérations fixé est atteint, soit lorsque la variation de la vraisemblance normalisée du corpus d'apprentissage, calculée lors de l'étape d'estimation, devient suffisamment petite.

Les GMMs sont considérés comme des "approximateurs universels" et peuvent donc modéliser n'importe quelle distribution, à condition d'avoir un nombre suffisant de gaussiennes. La complexité de la modélisation GMM se régle sur ce point, c'est-à-dire le nombre de gaussiennes nécessaires pour modéliser une distribution donnée d'observations. De par leur grande capacité de modélisation, les GMMs sont très souvent utilisés dans le domaine de la reconnaissance de la parole mais aussi en discrimination parole/musique [Saunders 96, Scheirer 97].

1.3.1.3 Les modèles de Markov Cachés

Les modèles de Markov cachés (HMMs) connaissent un succès considérable en reconnaissance de la parole depuis plusieurs décennies [Jelinek 76, Rabiner 93]. Le modèle de Markov caché peut être vu comme un automate probabiliste contrôlé par deux processus stochastiques (cf. Figure 1.10). Le premier processus (caché) contrôle les transitions entre les états du HMM en respectant les transitions imposées par la topologie de l'automate. Le second (observable) génère la suite des observations. Une observation est générée à chaque passage dans un état du HMM. En d'autres termes, nous avons deux suites. Une première suite observable qui correspond à la suite d'observations générée $o_1, \ldots, o_T$ où les o_i sont des vecteurs d'observations du signal à reconnaître. Une second suite cachée qui correspond à une suite d'états $q_0, q_1, \ldots, q_T$, où les q_i puisent leurs valeurs parmi

l'ensemble des N états du modèle $\{S_1, S_2, \ldots, S_N\}$ (cf. Figure 1.10).

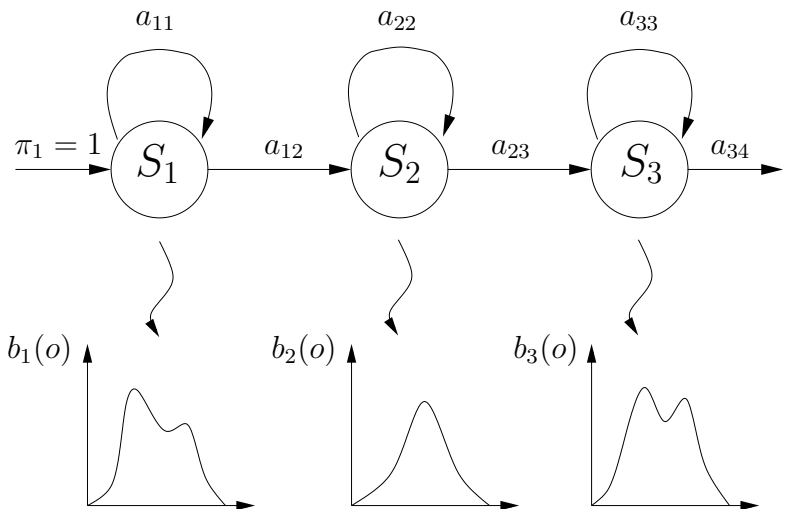

FIG. 1.10 – *HMM gauche-droite à 3 états usuellement utilisé pour la modélisation de phonèmes. Les lois de probabilité $b_i(o)$ fournissant les probabilités qu'une observation o ait été générée par un état S_i sont modélisées par des modèles à mélange de gaussiennes (GMM).*

Un HMM est défini par :

- un ensemble de N états

- Un vecteur Π de probabilités initiales, $\Pi = \pi_i$ $(i = 1..N)$. π_i est la probabilité d'être dans l'état i au temps 0,

$$\pi_i = P(q_0 = i), 1 \leq i \leq N$$

  Avec : $\pi_i \geq 0$ $\forall i$ et $\sum_{i=1}^{N} \pi_i = 1$

- Une matrice A de probabilités de transition entre les états, telle que a_{ij} soit la probabilité de transition pour aller de l'état i à l'état j,

$$a_{ij} = P(q_t = j | q_{t-1} = i) \quad 1 \leq i, j \leq N, \quad 1 \leq t \leq T$$

  Avec : $a_{ij} \geq 0$ $\forall i, j$ et $\sum_{j=1}^{N} a_{ij} = 1$

- Une matrice B de probabilités d'observation des symboles dans chacun des états,

telle que $b_i(o)$ soit la probabilité d'observer o quand le modèle est à l'état i.

$$b_i(o) = P(o_t|q_t = i),\ 1 \leq i \leq N,\ \ 1 \leq t \leq T$$

Avec : $b_i(o) \geq 0\ \forall i, k$ et $\int_o b_i(o)\ \mathrm{d}o = 1$

Au niveau de la topologie, il existe plusieurs types de HMMs. Les plus fréquemment utilisés sont le modèle ergodique et le modèle gauche-droit. Le modèle ergodique est sans contrainte sur les transitions, toutes les transitions entre états sont possibles. Le modèle gauche-droit, illustré par la figure 1.10, est le plus usité en reconnaissance de la parole. Il interdit toutes transitions vers un état de rang inférieur à l'état courant. Les HMMs peuvent être vus comme des processus permettant de générer des séquences d'observations. Il peuvent ainsi être utilisés comme règles de classification de séquences. Cependant, pour ce faire, il faut pouvoir d'une part calculer la probabilité qu'une séquence soit engendrée par un HMM et d'autre part, pouvoir apprendre les paramètres des HMM à partir d'exemples. Trois problèmes principaux se posent :

- "Evaluation" : étant donnée une séquence d'observations O et un HMM $\Lambda = (A, B, \pi)$, comment évaluer la probabilité d'observation $P(O, \Lambda)$?

- "Reconnaissance" : étant donnée une suite d'observations O et un HMM Λ, comment trouver la séquence d'états $q = (q_0, q_1; \ldots, q_T)$ qui maximise la probabilité d'observation de la séquence ?

- "Apprentissage" : étant donné un ensemble de J séquences d'observations O_j représentant chacune la même entité acoustique et donc associées au même HMM, comment choisir les paramètres (A, B, Π) de ce HMM afin de maximiser la probabilité que le HMM engendre la suite d'observations O_j ?

Chacun de ces problèmes a bien heureusement une solution.

Evaluation

L'évaluation de la probabilité de la séquence d'observations se fait grâce à l'approche de Baum [Baum 72], par les fonctions *forward-backward*. L'émission de la séquence d'observations se fait ici en deux temps :

- l'émission du début de la séquence d'observations, de o_1 à o_t en arrivant à l'état q_i au temps t,
- l'émission de la fin de la séquence d'observations, de o_{t+1} à o_T en partant de l'état q_i au temps t.

La probabilité de la séquence d'observations peut s'écrire ($O = (o_1 \dots o_T)$) :

$$P(O|\Lambda) = \sum_{i=1}^{N} \alpha_t(i)\beta_t(i) \tag{1.17}$$

Avec :

$$\alpha_t(i) = P(o_1 o_2 \dots o_t,\ q_t = q_i|\Lambda)$$
$$\beta_t(i) = P(o_t o_{t+1} \dots o_T|\ q_t = q_i,\ \Lambda)$$

$\alpha_t(i)$ est la probabilité d'émettre le début de la séquence d'observations O et d'arriver à l'état i au temps t, et β_i est la probabilité d'émettre la fin de la séquence d'observations en partant de l'état i au temps t.
L'appellation *forward-backward* vient du fait que le calcul de $\alpha_t(i)$ se fait avec t croissant :

$$\alpha_t(i) = \left[\sum_{j=1}^{N} \alpha_{t-1}(j)a_{ij}\right] b_i(o_t)$$

tandis que celui de $\beta_t(i)$ se fait avec t décroissant :

$$\beta_t(i) = \sum_{j=1}^{N} a_{ij}b_j(o_{t+1})\beta_{t+1}(j)$$

Ce calcul *forward-backward* a une complexité de l'ordre de N^2T. Des détails sur les algorithmes de calcul de α et β peuvent être trouvés dans [Cornuéjols 02].

Reconnaissance
Le problème de la reconnaissance consiste à déterminer le meilleur chemin correspondant à une séquence d'observations $O = (o_1, o_2, \dots, o_T)$, c'est-à-dire trouver la meilleure séquence d'états Q du modèle HMM Λ qui maximise la quantité suivante :

$$P(Q, O|\Lambda)$$

Pour trouver cette séquence, on définit la variable intermédiaire $\delta_t(i)$ comme la probabilité, connaissant les t premières observations, du meilleur chemin menant à l'état q_i :

$$\delta_t(i) = Max_{Q_1,\dots,Q_{t-1}} P(Q_1, Q_2, \dots, Q_t = q_i, o_1, o_2, \dots, o_t|\Lambda)$$

L'algorithme de Viterbi [Viterbi 67], une variante de la programmation dynamique, permet d'obtenir la séquence optimale d'états qui donne le meilleur chemin menant à l'état q_i au temps t.

Cet algorithme calcule $\delta_t(i)$ par récurrence :

- Initialisation : $\delta_0(i) = \pi_i$
- Récurrence : $\delta_t(i) = max_j(\delta_{t-1}(j).a_{ij}).b_i(o_t)$, j étant l'un des états du HMM.
- Terminaison : $P = max_i(\delta_T(i))$ est obtenue en cherchant l'état qui maximise la valeur δ à la dernière observation T.

Et il fournit en sortie la probabilité P^* de l'émission de la séquence par la meilleure suite d'états.

Apprentissage

L'apprentissage des paramètres d'un HMM se déroule en suivant une variante de l'algorithme *EM*, l'algorithme de Baum-Welch [Baum 70], sur un ensemble de séquences d'apprentissage $\mathcal{O} = O^1, \ldots O^J$. L'idée est d'utiliser une procédure de réestimation qui affine le modèle à chaque itération afin de maximiser :

$$P(\mathcal{O}|\Lambda) = \prod_{k=1}^{J} P(O^k|\Lambda)$$

Pour chaque étape p d'apprentissage, on dispose de Λ_p et on cherche Λ_{p+1} qui doit améliorer la probabilité d'émission des observations de l'ensemble d'apprentissage :

$$P(\mathcal{O}|\Lambda_{p+1}) \geq P(\mathcal{O}|\Lambda_p)$$

Pour calculer le nouveau Λ_{p+1}, un comptage de l'utilisation des transitions A et des distributions B et π du modèle Λ_p est effectué, lorsqu'il produit l'ensemble O. Les fréquences obtenues fournissent de bonnes approximations *a posteriori* des distributions de probabilités A,B et π si l'ensemble O est assez important. Nous avons ainsi :

$$\overline{b_j}(o_l) = \frac{\text{nombre de fois où le HMM était à l'état } q_j \text{ et générant } o_l}{\text{nombre de fois où le HMM était à l'état } q_j}$$

$$\overline{a}_{ij} = \frac{\text{nombre de transitions de l'état } q_i \text{ à l'état } q_j}{\text{nombre de transitions partant de } q_i}$$

Les formules de réestimation de $\overline{A}$ et $\overline{B}$ établies par Baum sont décrites dans ([Baum 72]). Pour plus de détails sur l'algorithme d'apprentissage des HMMs de Baum-Welsh, on peut se référer au livre de Rabiner et Juang [Rabiner 93].

1.3.2 Méthodes "discriminantes"

1.3.2.1 Description des méthodes "discriminantes"

Pour ces méthodes, on peut considérer que les points dans l'espace de paramètres correspondant à des observations de même classe se trouvent en général regroupés dans une même région de cet espace. Il est ainsi possible de définir pour chaque classe une fonction discriminante. Ces fonctions vont former une partition de l'espace des observations en régions mutuellement exclusives correspondant au domaine de chacune des classes. La surface de décision permettant de séparer les classes C_i et C_j est donnée par l'équation :

$$g_i(x) - g_j(x) = 0 \tag{1.18}$$

où $g_i(x)$ (respectivement $g_j(x)$) est la fonction discriminante de la classe C_i (respectivement C_j).
L'apprentissage pour ces méthodes consiste donc à apprendre ces surfaces de décision. Si les observations sont linéairement séparables alors les surfaces de décision seront des hyperplans. Mais en pratique, ce cas est rare et soit les surfaces de décision seront non-linéaires ou linéaires par morceaux, soit il faudra se ramener à un problème linéairement séparable en modifiant l'espace des observations (comme par exemple, avec les Machines à Vecteurs Support (SVM)).

Nous allons présenter les méthodes discriminantes les plus usitées en classification parole/musique : les k-plus proches voisins, les réseaux de neurones et les Machines à Vecteurs Support.

1.3.2.2 Les k-plus proches voisins (k-ppv)

La méthode des k-plus proches voisins (k-ppv) fait partie des méthodes par fonction noyau. Pour rappel, une fonction noyau K est une fonction bornée sur l'espace de représentation des observations d'intégrale égale à 1. On suppose en général que K est centré en 0 et par conséquent, $K(x_i, x_j)$ détermine une mesure de proximité entre les exemples x_i et x_j. La règle de décision des k-ppv est très simple : une observation nouvelle est classée en prenant la classe majoritaire parmi les k observations d'apprentissage les plus proches (cf. Figure 1.11). Cela revient à choisir une fonction noyau simple, constante dans l'hypersphère contenant les k voisins et nulle ailleurs.
Cette méthode n'a pas réellement de phase d'apprentissage, c'est-à-dire qu'il n'y a pas de construction de modèle. Tout repose sur :

- l'ensemble d'apprentissage stocké en mémoire,

- une mesure de distance, c'est-à-dire la fonction noyau. Parmi les distances exis-

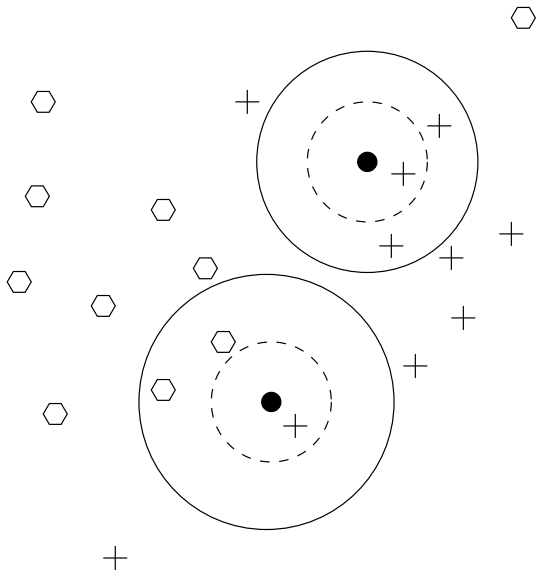

FIG. 1.11 – Décision par 1-ppv (cercle pointillé) et 3-ppv (cercle en trait plein) sur un ensemble d'observations appartenant à 2 classes.

tantes, on peut citer les distances euclidienne et de Mahanalobis.

– une méthode de choix de la classe, en général, la méthode consiste à choisir la classe majoritaire parmi les k observations d'apprentissage les plus proches.

Notons que la capacité de généralisation de cette méthode dépend du paramètre k. Le réglage de k permet de lisser la modélisation. En effet, un k élevé permet d'englober plus de voisins et ainsi d'être moins sensible aux erreurs d'apprentissage. Cette méthode a l'avantage de pouvoir s'appliquer à des cas de discrimination faisant intervenir un nombre élevé de classes. L'algorithme des k-ppv peut être résumé ainsi :

```
Début
Soit  x l'exemple à classer
et C = C_1,... ,C_n l'ensemble des classes
Pour chaque exemple (y,C_i) de l'ensemble d'apprentissage  faire
       calculer la distance D(y, x) entre y et  x
FinPour
Dans les k plus proches voisins de  x
       compter le nombre d'occurrences de chaque classe
Classer  x dans la classe apparaissant le plus souvent.
Fin
```

La règle des k-ppv fait une approximation de la décision bayésienne. En effet, elle fait implicitement une estimation comparative de toutes les densités de probabilité des classes

apparaissant dans le voisinage d'une observation à classer et choisit la plus probable. Les k-ppv ont été utilisés avec succès en classification audio, notamment dans [Scheirer 97, Lu 01]. Pour une étude plus approfondie sur les k-ppv, il est conseillé de se reporter à [Duda 73, Devijver 82] ou encore [Caran 96].

1.3.2.3 Les réseaux de neurones : le perceptron multi-couches (PMC[11])

Les réseaux de neurones (RN) constituent un domaine de recherche très intéressant et sont très couramment utilisés lorsque l'on parle de classification. Ils ont été notamment appliqués à des problèmes tels que : la reconnaissance de visage, le contrôle de robot, la reconnaissance de la parole, l'identification du locuteur, la segmentation parole/musique, etc.
Les RN réalisent un traitement d'informations distribué et sont composés d'unités de calcul primitives (les neurones formels) fonctionnant en parallèle et reliées entre elles par des connexions. Un neurone formel reçoit un nombre variable d'entrées en provenance de neurones en amont. A chacune de ces entrées est associé un poids représentant la force de la connexion. Il est aussi doté d'une sortie unique qui se ramifie ensuite pour alimenter les neurones en aval. Le principe de fonctionnement du neurone est simple, il calcule la somme pondérée de ses entrées et passe cette valeur à une fonction d'activation qui détermine l'excitation de ce neurone. La figure 1.12 illustre l'architecture d'un neurone formel. La sortie du neurone $y = F(\sum_{i=1}^{n} w_i x_i)$ dépend de la fonction d'activation choisie : fonc-

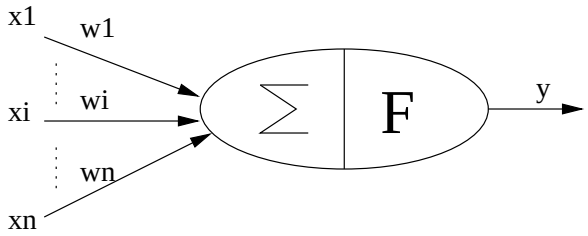

FIG. 1.12 – Architecture d'un neurone formel à n entrées.

tion seuil, linéaire par morceaux, sigmoïde, gaussienne etc.
Dans un réseau, la connaissance se trouve dans la topologie même du réseau et dans les poids des connexions. L'apprentissage d'un RN est réalisé à l'aide de méthodes d'apprentissage automatique utilisant la descente du gradient de l'erreur et se fait par modification des poids des connexions du réseau en fonction des données d'apprentissage. Aucune hypothèse sur la distribution des données n'est nécessaire.
Enfin, les RN ont de nombreuses propriétés très intéressantes telles que leur robustesse au bruit, leur flexibilité et leur capacité importante de généralisation. Nous allons présenter rapidement le réseau de neurones le plus souvent utilisé dans le domaine de la reconnaissance automatique de la parole et de la classification parole/musique : le perceptron

[11] *Multi-Layer Perceptron* (MLP) en anglais

multi-couches (PMC).

Le Perceptron Multi-Couches (PMC)

Le perceptron Multi-Couches est issu des travaux de F. Rosenblatt sur le perceptron monocouche [Rosenblatt 62]. Un PMC est un réseau dont les neurones sont disposés en plusieurs couches successives et où chaque neurone d'une couche est connecté à tous les neurones de la couche suivante et de la couche précédente mais pas aux neurones de la même couche. Le PMC est un réseau passe-avant (*feed-forward*), c'est-à-dire que les informations ou activations ne vont circuler que dans un seul sens, des neurones de la couche d'entrée vers les neurones de la couche de sortie (cf. Figure 1.13).

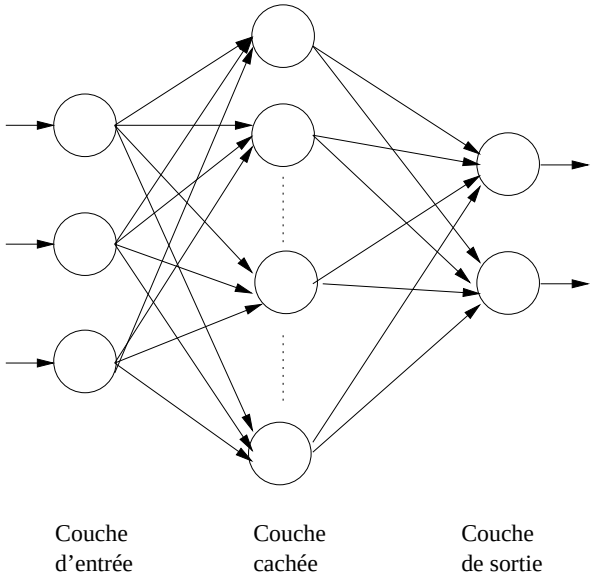

FIG. 1.13 – Architecture d'un Perceptron Multi-Couches à une couche cachée.

Une couche cachée dans un PMC correspond à une couche qui n'est ni la couche d'entrée, ni celle de sortie. De plus, un PMC peut avoir autant de couches cachées que désiré mais il a été montré [Hornik 89] que quelque soit le nombre de couches cachées dans un PMC, il existe un PMC équivalent avec une seule couche cachée. Cette couche cachée permet de modéliser des fonctions de décisions complexes et non linéaires entre n'importe quels espaces d'entrée et de sortie.
L'apprentissage des PMC se fait par rétropropagation du gradient de l'erreur [LeCun 85]. Le principe est d'adapter les différents poids des connexions en propageant l'erreur commise en sortie du réseau.

1.3.2.4 Les Machines à Vecteurs Support (SVM)

Les SVM, introduites par Vapnik et ses collègues [Boser 92] comme une nouvelle classe d'algorithmes d'apprentissage, constituent une application directe du principe inductif de minimisation structurel du risque [Vapnik 82, Vapnik 98]. Elles sont utilisées dans les trois problèmes classiques en apprentissage (régression, estimation de densité et discrimination). Ces différents algorithmes se caractérisent par le choix de maximiser les capacités en généralisation d'une fonction de discrimination f en minimisant une borne supérieure sur le risque. Le risque est l'erreur en généralisation de la fonction de discrimination f et correspondant à la probabilité que le résultat de f soit erroné. La borne supérieure sur le risque est ce que l'on appelle le *risque garanti*.

Dans le cadre de la discrimination, la SVM, à l'instar d'un perceptron, tente de séparer linéairement les données. Cependant, dans l'espace où elles se trouvent, les données ne sont généralement pas linéairement séparables (voir figure 1.14(a)). Dans ce cas, il devient utile d'effectuer un pré-traitement sur les données avant de les séparer avec des hyperplans. Ainsi, dans l'exemple représenté sur la figure 1.14(b), on peut pré-traiter les points de $\mathbb{R}^2$, en les projetant sur la surface d'un paraboloïde bien choisi. D'une manière générale, on projette les données, à l'aide d'une fonction Φ, dans un espace de plus grande dimension, appellé "espace de représentation", où l'on espère qu'elles seront linéairement séparables. On parle alors de SVM linéaire lorsque cette application Φ correspond à la fonction identité, i.e. lorsqu'elle ne renvoie pas les données dans un nouvel espace de représentation, et de SVM non linéaire dans le cas contraire.

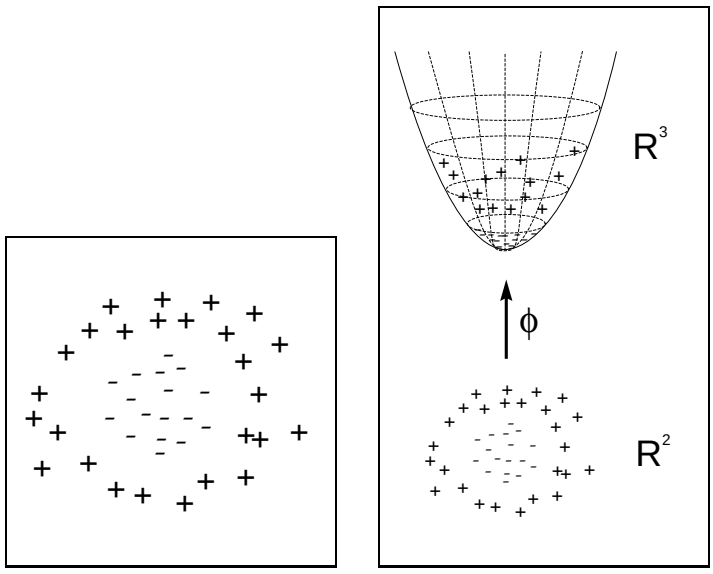

FIG. 1.14 – **(a)** données non linéairement séparables. **(b)** Pré-traitement des données, choix d'une transformation Φ (projection sur un paraboloïde) rendant les données linéairement séparables.

Enfin, les SVM ont été développés initialement dans le cadre d'une classification bi-

classes, mais des extensions multi-classes ont été proposés, comme la M-SVM [Guermeur 05]. Les SVMs ont récemment été introduites en reconnaissance de la parole et ont donné des résultats prometteurs [Wan 05a, Wan 05b, Wan 07], ainsi qu'en discrimination parole/musique [Mesgarani 06].

1.3.3 Méthodes "hybrides"

Les HMMs sont largement utilisés dans le domaine de la parole, que ce soit en reconnaissance ou en discrimination parole/musique. Mais ils présentent aussi quelques limitations comme le besoin de faire des hypothèses simplificatrices pour leur fonctionnement qui entraîne une limitation de leur généralité. De plus, leur apprentissage n'est en général pas discriminant.
La combinaison des HMMs avec des méthodes discriminantes semble intéressante et a été utilisée avec succès en reconnaissance de la parole et en discrimination parole/musique [Ajmera 03, Ganapathiraju 00]. Deux associations sont souvent utilisées : HMM-RN et HMM-SVM.

1.3.3.1 HMM et réseaux de neurones

Dans cette approche hybride, le réseau de neurones (la plupart du temps un PMC) se situe en aval d'un HMM et est utilisé comme estimateur de probabilités *a posteriori* d'appartenance à une classe. En effet, il a été démontré [Hopfield 87, Bourlard 90] qu'un PMC entraîné dans des conditions adéquates est équivalent à un estimateur de probabilités *a posteriori* d'appartenance à une classe.
Un perceptron peut ainsi apprendre les probabilités *a posteriori* des classes de phonèmes. Ces probabilités, grâce à la formule de Bayes, permettent d'obtenir les vraisemblances des observations qui vont être utilisées à la place de celles fournies par un mélange de gaussiennes dans un HMM classique.
Une autre façon d'utiliser la sortie du PMC comme entrée d'un HMM est illustrée par la figure 1.15. Ce système est celui de [Ajmera 03].

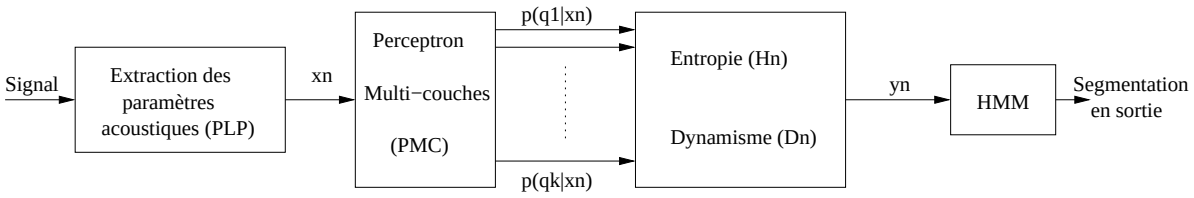

FIG. 1.15 – Système de segmentation parole/musique [Ajmera 03].

Des coefficients cepstraux (PLP) sont extraits tous les 16ms. Un PMC reçoit ces coefficients en entrée et donne en sortie des probabilités *a posteriori* pour les différentes classes de phonèmes. Les probabilités *a posteriori* des classes de phonèmes sont ensuite analysées

selon leur "entropie" et "dynamisme" pour finalement arriver en entrée du classifieur HMM qui effectuera la segmentation (les probabilités d'émission du HMM ont été estimées en utilisant soit un GMM, soit un deuxième PMC). Ce système de classification hybride a été utilisé avec succès par [Ajmera 02] en discrimination parole/musique. Diverses expérimentations ont montré qu'un tel système hybride améliore les performances des HMMs tout en conservant un temps de calcul et un encombrement mémoire raisonnable.

1.3.3.2 HMM et SVM

L'hybridation HMM/PMC donnant de bons résultats, il est donc normal de vouloir coupler les HMMs avec d'autres méthodes discriminantes telles les SVM. Contrairement aux PMC qui estiment des distributions de probabilité, les SVM estiment directement, à partir des données d'apprentissage, des surfaces de décision. Différentes méthodes ont été proposées pour convertir la distance d'une observation inconnue à une surface de décision fournie par une SVM en probabilités *a posteriori* exploitables par un HMM. Une de ces implantations a consisté à entraîner une SVM sur des données segmentales, en transformant les informations de distance fournie par une SVM en estimation de probabilités *a posteriori* pour les HMMs [Ganapathiraju 00]. Utilisée notamment en reconnaissance de la parole bruitée, cette hybridation donne déjà des résultats prometteurs. L'hybridation HMM/SVM n'a pas été mise en oeuvre, à notre connaissance, en discrimination parole/musique mais pourrait l'être à la manière du système hybride de [Ganapathiraju 00].

1.4 Conclusions

Nous venons de voir qu'il existe de nombreuses méthodes de classification utilisées en segmentation parole/musique. Chacune de ces méthodes à ses avantages et ses inconvénients. L'idée de combiner différentes méthodes en conservant leurs avantages et en limitant leurs inconvénients s'est concrétisée par l'intermédiaire des méthodes hybrides. Les résultats prometteurs obtenus permettent de continuer dans ce sens.
Nous avons aussi constaté au cours de ce chapitre que, pour la tâche de discrimination parole/musique, de nombreuses paramétrisations ont été mises en oeuvre. Cependant, ce sont les MFCC qui sont utilisés pour la séparation parole/musique dans la plupart des systèmes de reconnaissance automatique. Nous pourrons l'observer au chapitre suivant lors de l'étude de différents systèmes de reconnaissance de la parole comportant une phase de segmentation parole/musique. Ceci peut s'expliquer pour deux raisons, la première est que les coefficients MFCC, calculés lors de l'étape de segmentation, seront réutilisés dans la phase ultérieure de reconnaissance, permettant ainsi un gain de temps. La deuxième raison pour expliquer l'utilisation des coefficients MFCC vient du fait qu'il a été montré dans de nombreuses études que les coefficients MFCC étaient performants pour la tâche de discrimination parole/musique [Scheirer 97, Logan 00, Karneback 02, Razik 04].
Enfin, il a été montré dans [Scheirer 97] que le choix de la méthode de classification n'avait

pas une grande influence sur les résultats en discrimination parole/musique.
Il est donc important de se pencher sur l'étape de paramétrisation. La recherche de nouveaux paramètres doit donc se focaliser sur des paramètres compacts, rapides à calculer, et donnant des performances significativement supérieures aux MFCC.

2

Exemples de systèmes pour la segmentation Parole/Musique

Sommaire

Dans cette section, nous décrivons brièvement quelques systèmes de segmentation parole/musique intégrés dans des systèmes de transcription automatique de parole continue.

Nous présentons de tels systèmes car notre but est d'intégrer notre propre méthode de segmentation parole/musique à un système global de détection de mots clés reposant sur une transcription automatique du flux audio. Nous avons choisi de présenter des systèmes francophones et anglophones car d'une part nous travaillons sur la reconnaissance automatique de la parole en langue française et d'autre part, les systèmes en langue anglaise que nous présentons (DRAGON, IDIAP, Cambridge) sont connus pour être très performants en reconnaissance automatique de la parole. De plus, les systèmes ont été sélectionnés car ils proposent des approches différentes pour séparer la parole de la musique.

2.1 Le système du LIMSI

Le module de segmentation parole/musique exposé ici fait partie d'un système plus complet de partitionnement des données [Gauvain 02] développé dans le cadre d'un travail de recherche sur la reconnaissance de la parole initié par le DARPA.
Le DARPA (*Defense Advanced Research Projects Agency*) est une agence américaine (du département de la défense) responsable du développement de nouvelles technologies à usage militaire. Les recherches sur la transcription d'émissions radios ont permis d'évaluer et d'améliorer les technologies de reconnaissance de la parole.
Dans le module de segmentation du système de transcription du LIMSI, nous nous limiterons à la partie effectuant la segmentation parole/musique. Ce sous-module de segmentation est illustré par la figure 2.1.

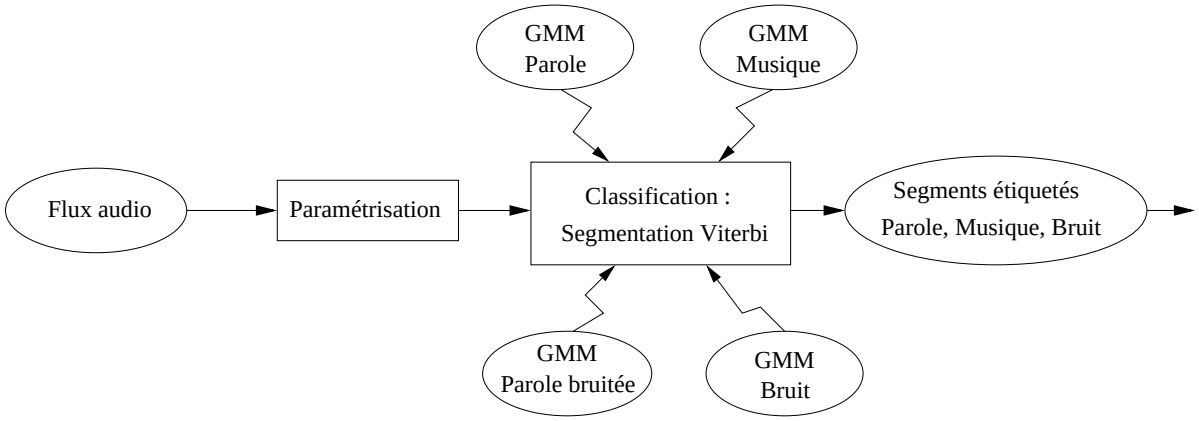

FIG. 2.1 – Architecture du système de segmentation parole/musique faisant partie du système de transcription d'émissions radiophoniques du LIMSI

2.1.1 Paramétrisation

Les paramètres utilisés sont des vecteurs acoustiques formés de :

– 12 coefficients cepstraux normalisés sans la log-énergie ($c0$),

- les dérivées premières (Δ) des 12 coefficients cepstraux et de la log-énergie,
- les dérivées secondes ($\Delta\Delta$) des 12 coefficients cepstraux et de la log-énergie.

Au final le vecteur de paramètres comprend 38 coefficients : 12 coefficients cepstraux, 13 Δ et 13 $\Delta\Delta$. La normalisation des coefficients cepstraux est basée sur la soustraction de la moyenne cepstrale (*CMR*) et sur la normalisation de la variance des coefficients. Ces paramètres sont les mêmes que ceux utilisés par leur module de reconnaissance de la parole mais sans la log-énergie. Ils sont estimés toutes les 10 millisecondes sur une fenêtre de Hamming de 30 millisecondes.

2.1.2 Classification

Le signal audio est alors segmenté à l'aide de l'algorithme de Viterbi en utilisant quatre GMMs pour modéliser les données. Chacun de ces GMMs est composé de 64 gaussiennes. Ces GMMs permettent de distinguer :

1. la parole,
2. la musique pure,
3. le bruit de fond,
4. la parole bruitée.

Ce quatrième GMM modélisait au départ la parole dans un environnement musical. Il devait permettre d'éviter de rejeter des segments contenant de la parole sur un fond musical [Gauvain 98]. Mais n'ayant pas apporté d'amélioration notable, ce GMM a été remplacé dans [Gauvain 99, Gauvain 02] par un GMM modélisant la parole bruitée. Ce nouveau GMM ne se limite donc plus à modéliser la parole avec un fond musical mais modélise la parole dans toutes sortes d'environnements bruités (environnement musical inclus).

2.2 Le système du LIA

Ce système a été réalisé dans le cadre de la phase I de la campagne d'évaluation ESTER (Evaluation des Systèmes de Transcription enrichie d'Emissions Radiophoniques) [Gravier 04]. Il a été développé exclusivement pour le traitement d'émissions radiophoniques françaises.

2.2.1 Paramétrisation

Comme pour le système du LIMSI, le système du LIA utilise une paramétrisation basée sur les MFCC. Cette fois le vecteur acoustique possède 39 coefficients : 12 coefficients MFCC et la log-énergie normalisée auxquels on ajoute les dérivées premières et secondes

de ces 13 premiers coefficients. Ces coefficients sont estimés toutes les 10 millisecondes sur une fenêtre de Hamming de 25 millisecondes. Il est à noter qu'aucune normalisation des paramètres n'est effectuée.

2.2.2 Classification

Dans le système de segmentation du LIA [Fredouille 04], la segmentation parole/musique se fait en deux étapes, comme l'illustre la figure 2.2.

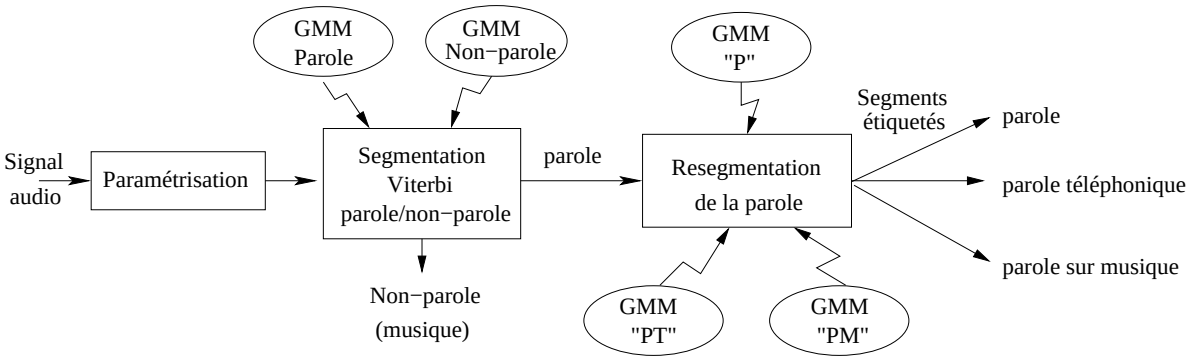

FIG. 2.2 – Architecture du système de segmentation du LIA dans le cadre de la campagne ESTER. "P" représente la parole, "PT" la parole téléphonique et "PM" la parole sur fond musical.

La première étape a pour but de séparer les segments de parole et de non parole, cette dernière catégorie incluant la musique. Deux modèles (GMMs) représentant la parole et la non-parole sont utilisés. Le processus de segmentation repose sur une recherche du meilleur modèle au niveau de la trame suivi par l'application de règles de regroupement des trames étiquetées.
La seconde étape resegmente les portions de parole en trois nouvelles classes : parole ("P"), parole téléphonique ("PT"), parole sur fond musical ("PM"). La segmentation repose ici sur un décodage Viterbi appliqué sur un HMM ergodique à trois états, chaque état représentant un des modèles GMM "P", "PT" ou "PM".

2.3 Le système de Cambridge

Nous décrivons ici la partie segmentation du système de transcription de journaux de 1997 [T.Hain 98], en nous limitant à la segmentation parole/musique. Ce module de segmentation automatique est repris au fil des années dans leur système de transcription sans grandes modifications au niveau de la segmentation parole/musique. Nous nous limitons à la première étape du système de segmentation car c'est dans cette étape que la parole est séparée de la musique. La première étape de segmentation dans le système

HTK consiste donc à classer les données audio en trois grandes catégories : parole studio (S), parole téléphonique (T) et la dernière catégorie (M) englobant la musique et tout autre bruit de fond ne correspondant pas à de la parole. Cette étape permet de séparer et rejeter tous les segments qui ne correspondent pas à de la parole.

2.3.1 Paramétrisation

Les paramètres sont des vecteurs acoustiques composés de 12 MFCC, du logarithme de l'énergie normalisée, ainsi que des dérivées premières Δ et secondes $\Delta\Delta$ des 13 coefficients précédents. Ces paramètres sont classiquement utilisés en reconnaissance automatique de la parole.

2.3.2 Classification

La segmentation du flux audio en parole et musique (et non-parole) s'effectue ici en deux grandes étapes se répétant. La première étape consiste en une première classification des trames du signal. Pour cela, quatre GMMs à 1024 gaussiennes sont utilisés pour modéliser les données. Ces quatre modèles correspondent à :

- la parole studio "pure",
- la parole téléphonique "pure",
- la parole sur fond musical,
- la musique.

Un décodeur basé sur l'algorithme de Viterbi est mis en oeuvre avec les 4 modèles en parallèle pour étiqueter les trames. Une pénalité de transition inter-classe est mise en place dans le décodeur pour qu'il produise des segments plus longs. Une autre pénalité est ajoutée lorsque l'on quitte le modèle de musique dans le but de réduire le nombre de mauvaises classifications.
Cette première classification est suivie d'une étape d'adaptation des modèles GMMs (cf Figure 2.3).

Les modèles sont adaptés en utilisant des transformations MLLR (*Maximum Likelihood Linear Regression*) calculées pour chaque classe. Quinze itérations de MLLR sont effectuées en utilisant les résultats de la première étape de classification. Ce grand nombre d'itérations est dû au grand nombre de gaussiennes de ces modèles. Aussi, cette étape d'adaptation n'est effectuée que pour les classes contenant au moins 15s de données. Dans [Woodland 98], ce type de schéma permet ainsi une augmentation du nombre de trames bien étiquetées et une diminution du nombre de trames perdues (étiquetées comme de la musique ou supprimées par erreur). Le décodage de Viterbi de la première étape est alors réeffectué en utilisant les nouveaux modèles. Cette opération est réitérée plusieurs fois.

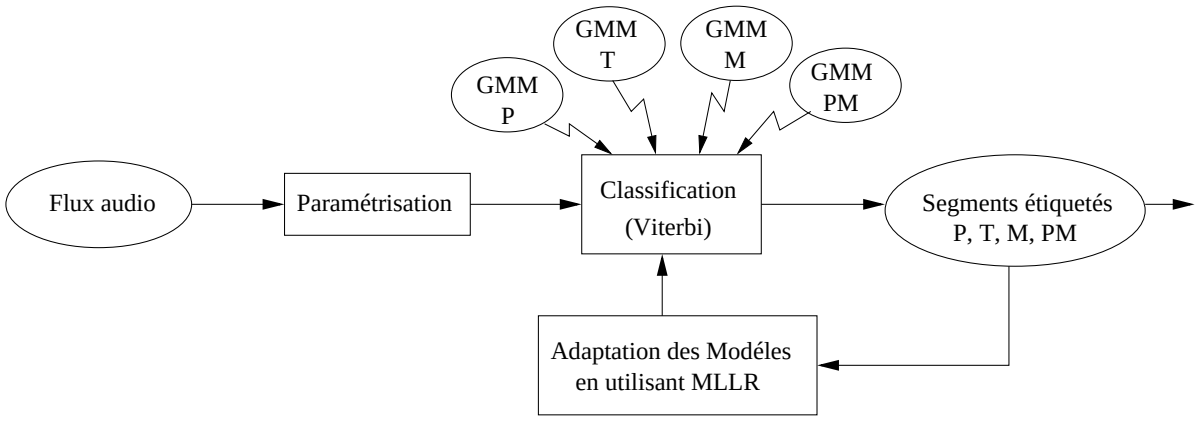

FIG. 2.3 – Partie de l'architecture du module de segmentation du système de transcription de journaux de Cambridge (HTK). Cette partie correspond à la segmentation parole/musique. Les symboles P, T, M, PM signifient respectivement la parole, la parole téléphonique, la musique et la parole sur de la musique.

2.4 Le système hybride de l'IDIAP

Le système de segmentation parole/musique proposé par Ajmera et Bourlard [Ajmera 02, Ajmera 03] suit une approche très différente des autres systèmes. Ils ne se basent pas sur la nature acoustique du signal pour différencier la parole de la musique, mais plutôt sur le comportement de la sortie du reconnaisseur lorsque différents types de signaux (parole, musique) sont fournis en entrée.

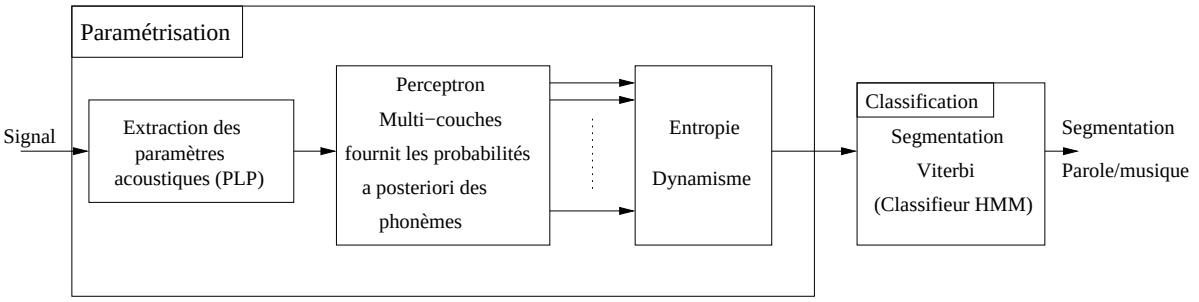

FIG. 2.4 – Système de segmentation parole/musique de l'IDIAP. Les paramètres extraits sont basés sur le calcul de l'Entropie et du Dynamisme des probabilités *a posteriori* d'émission des phonèmes.

2.4.1 Paramétrisation

Des coefficients cepstraux (PLP) sont extraits tous les 16ms et définissent la paramétrisation du signal. Ces coefficients ne sont pas utilisés directement pour discriminer la parole de la musique. Les paramètres utilisés pour cela sont basés sur les probabilités *a posterori*

d'émission des phonèmes utilisés en reconnaissance. Ces probabilités d'émission sont estimées par un perceptron multi-couches (PMC) entraîné sur de la parole pure. L'examen du comportement des ces probabilités permet de discriminer la parole de la non-parole (musique). Deux paramètres sont finalement extraits, l'entropie et le dynamisme, pour analyser le comportement des probabilités *a posterori*. Ainsi, dans le cas d'un segment de parole, la valeur de la probabilité *a posterori* d'un phonème (le phonème reconnu) sera plus grande que pour les autres phonèmes et l'entropie sera proche de zéro. Le dynamisme mesure la variation des valeurs des probabilités, c'est-à-dire leur dynamique. Comme la parole est un signal variant beaucoup, les valeurs des probabilités en sortie du PMC vont elles aussi varier rapidement, entraînant ainsi un dynamisme élevé. Au contraire, la musique est un signal harmonique et donc la dynamique des probabilités sera faible. Ces deux paramètres forment un vecteur à deux dimensions qui servira d'entrée au classifieur de l'étape suivante.

2.4.2 Classification

L'étape de classification est basée sur un classifieur HMM. La topologie de ce HMM est illustrée par la figure 2.5. Le HMM est un modèle entièrement connecté à deux états, où une durée minimum est imposée pour chaque état. Les deux états sont en fait représentés par des "hyper-états" composés de plusieurs sous-états concaténés représentant la même fonction de densité de probabilité et imposant ainsi une contrainte de durée.

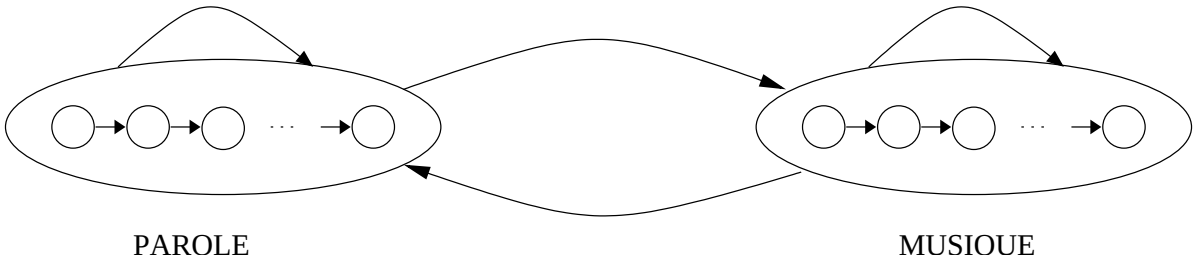

FIG. 2.5 – Topologie du HMM utilisé pour la classification parole/musique de Ajmera/Bourlard.

N'ayant aucune information *a priori* sur les probabilités de transitions entre les segments de parole et de musique, les probabilités de transitions entre états (parole et musique) sont réglées manuellement, en favorisant la stationnarité dans l'état courant (parole ou musique). Concernant le début du signal audio, les probabilités initiales sont elles aussi réglées manuellement de manière équiprobable entre parole et musique.
Enfin, les probabilités d'émissions des états du HMM sont estimées soit par un GMM, soit par un PMC.
L'algorithme de Viterbi est utilisé pour trouver la meilleure séquence d'états (parole/musique) possible qui pourrait être à l'origine de la séquence d'observations.

2.5 Le système de l'IRIT

Le système proposé par le laboratoire de Toulouse (IRIT) [Pinquier 04b, Mauclair 04] est différent des autres systèmes vus jusqu'ici puisqu'il traite séparément la détection de la parole et la détection de la musique. Il est basé sur un schéma "classe/non-classe" que nous avons repris dans notre propre système. Le système se décompose donc en deux sous-systèmes traitant chacun soit la parole, soit la musique (voir figure 2.6).

2.5.1 Paramétrisation

Le système séparant la détection de la parole et la détection de la musique, il est logique d'essayer d'utiliser des paramétrisations spécifiques dans chaque cas. C'est ce qui est fait dans le système de l'IRIT. Ces paramètres sont décrits en détails dans [Pinquier 02a].

Le sous-système de détection de la parole utilise la modulation de l'entropie et la modulation de l'énergie à 4Hz. La modulation à 4 Hz est un indicateur de la fréquence syllabique, ce paramètre aura une valeur plus élevée pour les segments de parole que pour les segments de musique. La modulation de l'entropie, quant à elle, nous renseigne sur la perturbation du signal : un signal de parole est très perturbé comparé à un signal de musique du fait des silences suivis d'explosions provenant des occlusives.

Le sous-système de détection de musique utilise des paramètres différents et originaux. Une segmentation issue de l'algorithme de "Divergence Forward-Backward" (DFB) [André-Obrecht 88] est réalisée sur une seconde de signal. Cet algorithme est basé sur une étude statistique du signal dans le domaine temporel. L'hypothèse de départ est que le signal de parole est décrit par une suite de zones quasi-stationnaires. Deux paramètres sont ensuite extraits de cette segmentation :

- Le nombre de segments :
  Ce paramètre correspond au nombre de segments présents durant chaque seconde de signal. Les signaux de parole présentent une alternance de périodes de transition (voisées/non-voisées) et de périodes relativement stables (les voyelles en général), alors qu'en général, la musique étant plus tonale (ou harmonique), ne possède pas de telles variations. Le nombre de segments par seconde est donc plus important pour la parole que pour la musique.
- La durée des segments :
  Les segments sont généralement plus longs pour la musique que pour la parole. La durée des segments est modélisée avec une loi gaussienne inverse. La fonction de

densité de probabilité est donnée par :

$$p(g) = \sqrt{\frac{\lambda}{2\pi g^3}} \times e^{\frac{-\lambda(g-\mu)^2}{2\mu^2 g}}, \ g \geq 0$$

avec μ la valeur moyenne de g et $\frac{\mu^3}{\lambda}$ la variance de g (g est la variable aléatoire représentant la durée des segments).

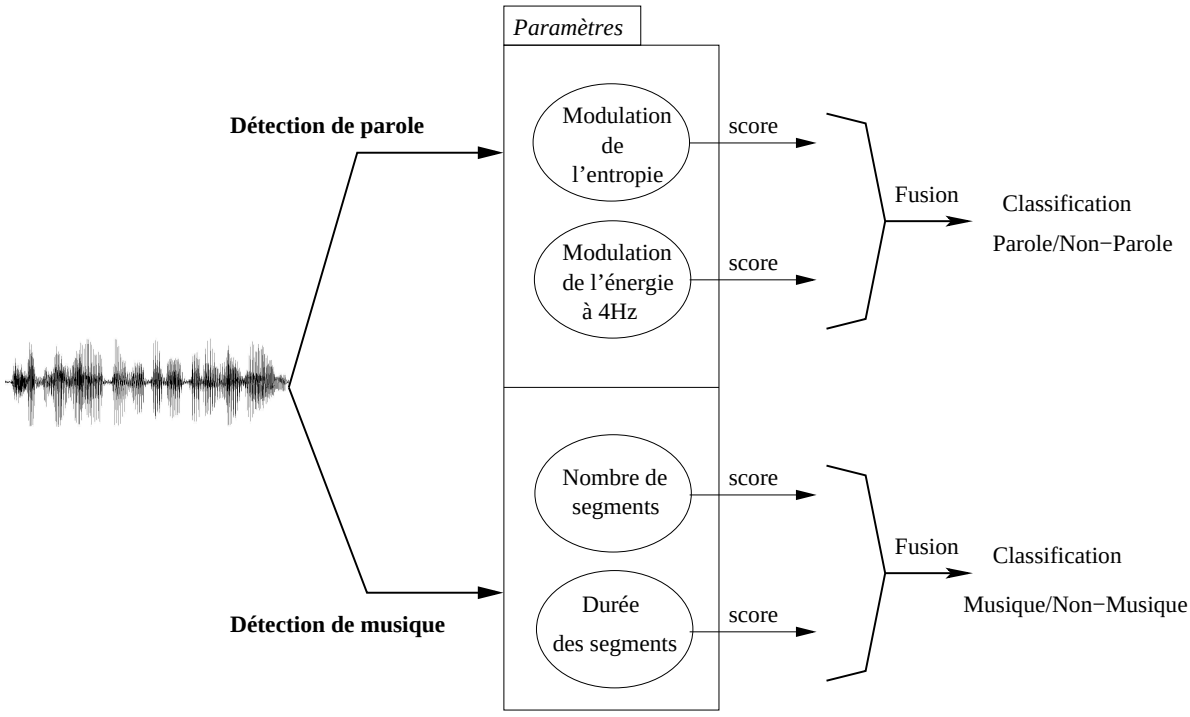

FIG. 2.6 – Le système de classification Parole/Musique de l'IRIT. Les scores de vraisemblance donnés pour chaque paramètre sont fusionnés pour donner une décision : parole/non-parole dans un cas et musique/non-musique dans l'autre.

2.5.2 Classification

Les deux systèmes que nous avons décrits précédemment vont prendre une décision sur une fenêtre d'une seconde du signal. Chaque sous-système (parole/non-parole et musique/non-musique) correspond en fait à deux "classifieurs" statistiques liés à chacun des deux paramètres utilisés. Les distributions probabilistes, associées à chaque paramètre et classe, sont des lois gaussiennes sauf celle associée à la durée des segments qui est une loi inverse gaussienne.
Chaque "classifieur" délivre alors un score de vraisemblance. Il faut ensuite fusionner les scores obtenus pour les deux paramètres pour obtenir la décision du sous-système. La fusion est basée sur la maximisation des scores, c'est-à-dire que le score de vraisemblance

le plus important détermine le choix (parole ou non-parole, musique ou non-musique). Au final les deux sous-systèmes nous indiquent si nous sommes en présence de parole ou non et en présence de musique ou non, comme l'illustre la figure 2.6.

2.6 Le système de "DRAGON Systems"

L'étape de segmentation parole/musique présentée fait partie d'un système de transcriptions de journaux radiophoniques en langue Mandarin [Zhan 99]. La segmentation s'effectue en deux phases comme l'illustre la figure 2.7. La première phase consiste à découper automatiquement le signal afin d'obtenir des segments d'une durée comprise entre 2 et 30 secondes. La seconde phase consiste à détecter les segments de musique dans les segments précédemment obtenus afin de les éliminer.

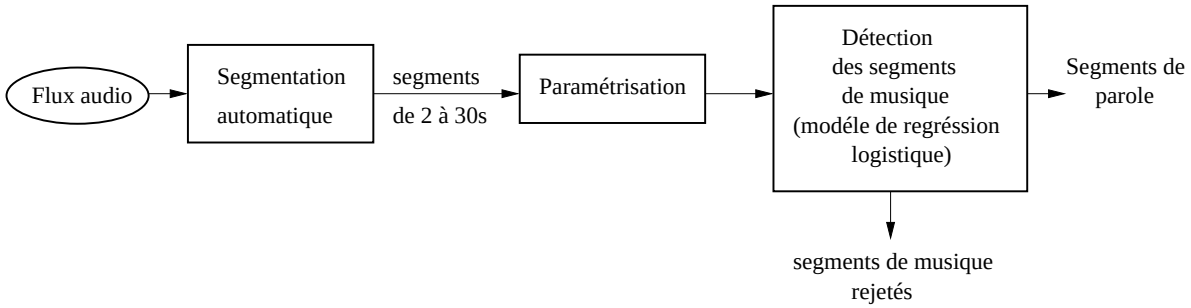

FIG. 2.7 – Le système de segmentation Parole/Musique de DRAGON Systems. Le flux audio est d'abord découpé en segments d'une durée de 2 à 30 secondes avant d'être paramétré. La deuxième étape permet alors de séparer les segments de parole et de musique afin de ne conserver que les segments de parole. Cette deuxième étape utilise un modèle de régression logistique.

2.6.1 Segmentation automatique

Avant de séparer les segments de musique des segments de parole, le flux audio subit une première segmentation automatique. Cette procédure de segmentation effectue les étapes suivantes :

- Un détecteur basé sur l'amplitude du signal est utilisé pour découper le flux audio en morceaux d'une durée de 20 à 30 secondes.
- Ces morceaux sont ensuite redécoupés en segments d'une durée comprise entre 2 et 30 secondes en se basant sur les zones de silence produites par une rapide reconnaissance en mots. Cette reconnaissance rapide est approximative, mais ce n'est pas important car la seule information intéressante ici est l'information de silence.

– Enfin, les frontières de ces segments sont encore affinées en utilisant un système de détection de changements de locuteur, en utilisant le test T^2 d'Hotelling ou le critère BIC (*Bayesian Information Criterion*).

2.6.2 Détection des segments de musique

2.6.2.1 Paramétrisation

Pour détecter les segments de musique afin de les éliminer, plusieurs paramètres ont été sélectionnés :

– Variance du point spectral de coupure (*spectral rolloff point*),
– Moyenne et variance du centre de gravité spectral (*spectral centroïd*)
– Moyenne du flux spectral (*delta spectrum magnitude* ou *spectral flux*),
– 5 moyennes et 8 variances des coefficients cepstraux, parmi les moyennes et les variances de 12 coefficients cepstraux et de leurs premières et secondes dérivées.

Les coefficients cepstraux sont calculés toutes les 10ms en utilisant une fenêtre d'analyse de 20ms. Leurs moyennes et leurs variances sont calculées sur une fenêtre d'une seconde, c'est-à-dire sur 100 trames.

2.6.2.2 Classification

Un modèle de régression logistique est utilisé comme méthode de classification. Ce modèle est entraîné en utilisant les paramètres précédents. Le modèle créé pourra ainsi prédire la catégorie (parole ou musique) d'une nouvelle observation. De plus, lors de l'entraînement du modèle, des segments de musique ont été étiquetés en parole. Cet ajout de musique dans les données de parole a été réalisé afin d'inciter le modèle à classer en parole les segments contenant de la parole sur fond musical.

2.7 Conclusions

Une première conclusion s'impose : la moitié des systèmes présentés utilise le même type de paramètres, les coefficients cepstraux (MFCC) pour discriminer la parole de la musique. C'est le cas pour la plus grande majorité des systèmes existants, lorsque la segmentation parole/musique intervient dans un système complet de reconnaissance automatique de parole continue. L'utilisation des MFCC au lieu de paramètres spécialement étudiés pour la discrimination parole/musique peut s'expliquer de deux manières. D'une part, les coefficients MFCC sont généralement utilisés dans différents modules des systèmes de transcription audio, par exemple au niveau du moteur de reconnaissance. Ainsi, utiliser les MFCC pour la segmentation parole/musique permet d'éviter de perdre du

temps à calculer d'autres paramètres. D'autre part, bien que les MFCC soient utilisés à l'origine pour modéliser l'information spectrale à court terme de la parole dans le cadre de la reconnaissance de la parole, des recherches ont montré leur capacité à modéliser également la musique [Logan 00]. En effet, les deux principales caractéristiques des MFCC, à savoir l'échelle Mel et la transformée en cosinus discrète (DCT), sont bien adaptées à la modélisation de la musique.

L'utilisation des MFCC en paramétrisation étant devenue naturelle, les recherches se sont alors principalement concentrées sur la méthode de classification. Bien que cette seconde étape soit importante pour la discrimination parole/musique, elle n'en est pas moins dépendante de la paramétrisation du signal. En effet, si les paramètres extraits du signal ne sont pas discriminants, il sera très difficile, voire impossible de séparer parole et musique. Enfin, Scheirer et Slaney [Scheirer 97] ont eux aussi montré l'importance des paramètres par rapport au classifieur dans les performances d'un système de segmentation parole/musique.

Cela permet de justifier notre choix de nous concentrer principalement sur la paramétrisation du signal pour améliorer la segmentation parole/musique et remplacer les MFCC dans le module de segmentation. Il nous faut donc trouver une paramétrisation dont les performances sont significativement supérieures aux MFCC.

3

Une nouvelle approche pour la segmentation Parole/Musique

Sommaire

Dans ce chapitre nous proposons une nouvelle approche de discrimination parole/musique fondée sur l'utilisation de la décomposition en ondelettes du signal. Nous avons étudié différentes décompositions en ondelettes du signal, en extrayant des paramètres fondés sur l'énergie. Pour valider l'approche proposée, nous avons effectué des expériences sur différents corpus. Les corpus seront présentés au chapitre 4 et les résultats de nos expériences au chapitre 5.

3.1 Présentation des ondelettes

3.1.1 Un peu d'histoire

Cette section présente rapidement la base de notre approche en paramétrisation du signal, à savoir la décomposition du signal en ondelettes. Cette présentation est un rapide aperçu des fondements théoriques des ondelettes. Pour aller plus loin sur cette théorie du traitement du signal à l'aide d'ondelettes, le lecteur pourra se reporter au livre de Mallat [Mallat 98, Mallat 00].

Au quotidien, notre attention (visuelle ou auditive) est attirée par le mouvement et les phénomènes transitoires, au contraire des stimuli stationnaires qui sont vite ignorés. Cette stratégie qui donne la priorité aux phénomènes transitoires permet de sélectionner les informations importantes de notre environnement, informations qui, en des temps anciens, nous ont permis de survivre. Pourtant le traitement du signal classique s'est surtout concentré sur l'étude d'opérateurs invariants dans le temps et dans l'espace, qui modifient les propriétés stationnaires des signaux. Cela a conduit à l'hégémonie indiscutable de la transformée de Fourier. La transformée de Fourier est un outil fondamental pour une grande variété d'applications, telles que les transmissions ou les traitements des signaux stationnaires. Néanmoins, si nous nous intéressons à des phénomènes transitoires, la transformée de Fourier s'avère inadéquate. En effet, nous pouvons définir un morceau musical comme un ensemble de "fréquences" sonores qui varient dans le temps. De telles évolutions temps-fréquence peuvent être mises en évidence en décomposant le signal en fonctions élémentaires bien concentrées en temps et en fréquence. La transformée de Fourier à fenêtre et la transformée en ondelettes sont deux exemples importants de décomposition temps-fréquence. C'est en 1946 que le physicien Gabor [Gabor 46] propose

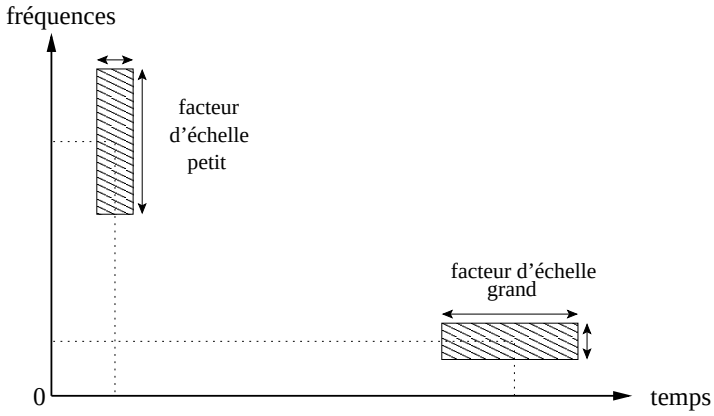

FIG. 3.1 – Boîtes de Heisenberg correspondant au pavage du plan temps-fréquence de la transformée en ondelettes à des échelles différentes. Une échelle plus petite réduit l'étalement en temps mais augmente la taille du support fréquentiel.

d'analyser les signaux sonores avec des atomes élémentaires qui sont des fonctions bien

concentrées en temps et en fréquence. En montrant que de telles décompositions sont étroitement liées à notre perception des sons, et qu'elles isolent les structures importantes des signaux de parole et de musique, les travaux de Gabor furent à la base de l'analyse temps-fréquence. Gabor introduit ainsi en 1946 les atomes de Fourier à fenêtre afin de mesurer les "variations fréquentielles" des sons.

La résolution temps-fréquence de la transformée de Fourier à fenêtre dépend de l'étalement de la fenêtre en temps et en fréquence. Cet étalement correspond à la surface de la boîte de Heisenberg. En effet, les concentrations en temps et en fréquence sont limitées par le principe d'incertitude d'Heisenberg. Ce principe, qui dit que l'énergie d'une fonction et de sa transformée de Fourier ne peuvent être simultanément concentrées sur des intervalles arbitrairement petits, a une interprétation importante en mécanique quantique, en tant qu'incertitude sur la position et la quantité de mouvement d'une particule libre. En d'autres termes, le principe d'incertitude d'Heisenberg indique qu'un signal ne peut pas être simultanément connu avec des précisions en temps Δt et en fréquence Δf quelconques, le produit de ces deux quantités étant borné inférieurement [Mallat 00] :

$$\Delta t \Delta f \geq \frac{1}{2}$$

Un inconvénient de la transformée de Fourier à fenêtre est le réglage de la taille de la fenêtre d'analyse. Ce réglage est un compromis entre résolution temporelle et résolution fréquentielle. On perd en localisation fréquentielle ce qu'on a gagné en localisation temporelle, ceci à cause du principe d'incertitude de Heisenberg (cf. Figure 3.1). Ainsi, une repré-

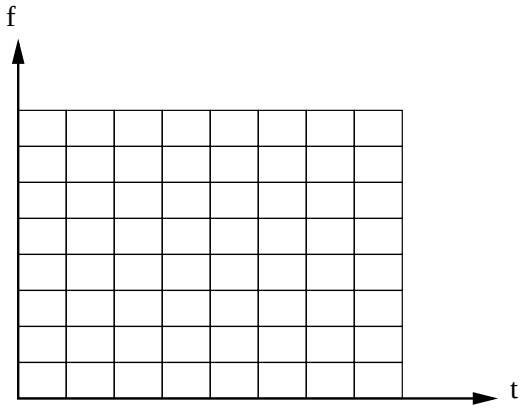

FIG. 3.2 – Exemple de couverture temps-fréquence avec la transformée de Fourier à fenêtre. Les résolutions temporelle et fréquentielle restent inchangées quelque soit le temps et la fréquence.

sentation satisfaisante de la structure temporelle fine du signal permettant par exemple de voir les transitions entre phonèmes se fera au détriment de la résolution fréquentielle (analyse large bande). Inversement, une analyse permettant de bien faire apparaître les

composantes harmoniques du signal se fera au détriment de la résolution temporelle et ne rendra pas compte des événements temporels brefs (analyse bande étroite). Une fois ce réglage effectué, la taille de la fenêtre sera fixée et la résolution de la transformée de Fourier à fenêtre restera la même sur tout le plan temps-fréquence (cf. Figure 3.2). Mais pour analyser des composantes transitoires de durées différentes comme c'est souvent le cas en parole et en musique, il est nécessaire d'utiliser des atomes dont les supports temporels ont des tailles variables. La transformée en ondelettes en est la solution (cf. Figure 3.3).

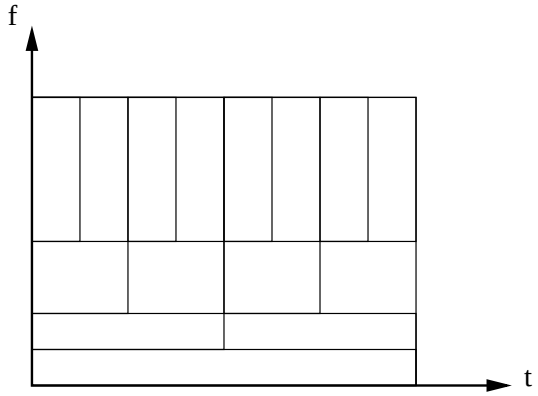

FIG. 3.3 – Un exemple de couverture temps-fréquence avec la transformée en ondelettes

3.1.2 Définitions

Nous avons vu qu'une alternative, pour dépasser les limitations de la transformée de Fourier à fenêtre, se trouve être l'utilisation de la transformée en ondelettes. Nous pouvons à présent définir ce qu'est une ondelette [Mallat 00] et comment réaliser une transformée en ondelettes du signal.

3.1.2.1 Les ondelettes

Une ondelette [Mallat 00] est une fonction $\Psi \in L^2(\mathbb{R})$ de moyenne nulle :

$$\int_{-\infty}^{+\infty} \Psi(t)dt = 0$$

et à énergie finie :

$$\int_{-\infty}^{+\infty} |\Psi(t)|^2 < +\infty$$

Elle est normalisée à $||\Psi|| = 1$, et centrée au voisinage de $t = 0$. Une famille d'atomes temps-fréquence s'obtient en dilatant l'ondelette par un facteur s, et en la translatant par

u :

$$\Psi_{u,s}(t) = \frac{1}{\sqrt{s}}\Psi\left(\frac{t-u}{s}\right) \quad \text{avec } s \in \mathbb{R}_+^*$$

Si on considère $||\Psi_{u,s}|| = 1$, alors les ondelettes dilatées et translatées restent de norme unitaire. L'ondelette peut être réelle ou analytique complexe. Selon les applications, on peut choisir l'une ou l'autre. Pour notre part, nous avons opté pour une ondelette réelle. Nous allons maintenant définir la transformée en ondelettes.

3.1.2.2 La transformée en ondelettes

La transformée en ondelettes d'un signal $f(t)$ à l'échelle s et au temps u se calcule en corrélant $f(t)$ avec l'ondelette $\Psi_{u,s}$ correspondante. Ceci nous donne la définition suivante de la transformée en ondelettes :

$$Wf(u,s) = < f, \Psi_{u,s} > = \int_{-\infty}^{+\infty} f(t)\frac{1}{\sqrt{s}}\Psi^*\left(\frac{t-u}{s}\right)dt,$$

où

- W est l'initiale de *Wavelet* qui signifie ondelette en anglais,
- Ψ^* est le complexe conjugué de Ψ.

Nous utiliserons par la suite uniquement des transformées en ondelettes réelles car elles permettent de mesurer la variation de $f(t)$ dans un certain voisinage de u (dépendant de Ψ) de taille proportionnelle à s. Il a été démontré que lorsque s tend vers 0, la décroissance des coefficients d'ondelettes caractérisent la régularité de $f(t)$ au voisinage de u. Cette propriété est très importante pour nous car elle permet de détecter des transitoires.
Enfin, une transformée en ondelettes réelles est complète et préserve l'énergie tant que l'ondelette Ψ satisfait une condition d'admissibilité donnée par le théorème suivant :

Théorème 1 (Calderón, Grossmann, Morlet) *Soit $\Psi \in L^2(\mathbb{R})$ une fonction réelle (ou un signal réel) vérifiant :*

$$C_\Psi = \int_0^{+\infty} \frac{|\hat{\Psi}(\omega)|^2}{\omega}d\omega \ < +\infty$$

où $\hat{\Psi}$ est la transformée de Fourier de Ψ.
Toute fonction $x(t) \in L^2(\mathbb{R})$ vérifie :

$$x(t) = \frac{1}{C_\Psi}\int_0^{+\infty}\int_{-\infty}^{+\infty} W_x(u,s)\frac{1}{\sqrt{s}}\Psi\left(\frac{t-u}{s}\right)\ du\ \frac{ds}{s^2}$$

et

$$\int_{-\infty}^{+\infty} |x(t)|^2 \, dt = \frac{1}{C_\Psi} \int_0^{+\infty} \int_{-\infty}^{+\infty} |W_x(u,s)|^2 \, du \, \frac{ds}{s^2}$$

La condition

$$C_\Psi = \int_0^{+\infty} \frac{|\hat{\Psi}(\omega)|^2}{\omega} d\omega \; < +\infty$$

du théorème précédent s'appelle la *condition d'admissibilité* de l'ondelette. Pour que l'intégrale soit finie, il faut s'assurer que $\hat{\Psi}(0) = 0$, ce qui explique pourquoi les ondelettes doivent être de moyenne nulle. Cette condition est presque suffisante. Si $\hat{\Psi}(0) = 0$ avec $\hat{\Psi}(\omega)$ continûment différentiable, la condition d'admissibilité est alors satisfaite. On vérifie assez facilement que $\hat{\Psi}(\omega)$ est continûment différentiable si Ψ décroît assez vite à l'infini. C'est pourquoi on choisit aussi des ondelettes à décroissance rapide. Enfin, la dernière équation du théorème démontre la conservation de l'énergie entre le domaine temporel et le domaine des ondelettes.

Les signaux de parole et de musique sont continus mais nous travaillons sur un signal discret $f[n] = f(n)$ (de taille N). Nous utiliserons donc la version discrète de la transformée en ondelettes. La transformée en ondelettes discrète se calcule aux échelles $s = a^j$, avec $a = 2^{1/v}$, ce qui fournit v échelles intermédiaires pour chaque octave[12] $[2^j, 2^{j+1}[$. De plus, la transformée en ondelettes de f ne pourra être calculée que pour les échelles :

$$\frac{1}{N} < s \leq 1$$

3.1.3 La transformée en ondelettes discrète utilisée pour la segmentation parole/musique

Le traitement du signal basé sur les ondelettes a été utilisé avec succès pour des problèmes très variés, comme la compression d'image [Saha 00], le débruitage de la parole [Kim 01], la classification audio [Tzanetakis 02, Lin 05], la reconnaissance automatique de la parole [Sarikaya 00, Deviren 04], etc.
L'utilisation des ondelettes permet de faire une analyse multi-résolution du signal. Nous verrons l'intérêt de ce type d'analyse dans le cadre de la segmentation parole/musique. Mais tout d'abord, définissons la transformée en ondelettes discrète.

Soit un signal $f(t)$ échantillonné uniformément sur $[0, 1]$ avec un pas d'échantillonnage de $1/N$. On obtient un signal discret $f[n] = f(\frac{n}{N})$ composé de N échantillons.
Soit $\Psi(t)$ une ondelette en temps continu dont le support est inclus dans $[-K/2, K/2]$. Pour $2 \leq a^j \leq \frac{N}{K}$, on définit une ondelette discrète dilatée par a^j :

$$\Psi_j[n] = \frac{1}{\sqrt{a^j}} \Psi\left(\frac{n}{a^j}\right)$$

[12] une octave est l'intervalle séparant deux sons dont les fréquences fondamentales sont en rapport de un à deux. Il désigne ainsi un doublement de fréquence.

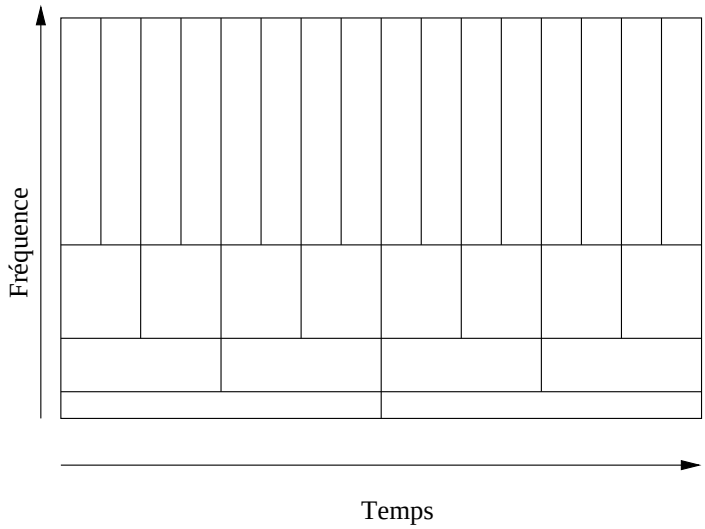

FIG. 3.4 – Décomposition temps-fréquence du signal. Une décomposition dyadique est appliquée à la fois sur l'axe du temps et l'axe des fréquences.

Elle a KNa^j valeurs non nulles sur $[-N/2, N/2]$. L'échelle a^j doit être supérieure à 2 pour que le pas d'échantillonnage soit plus petit que le support de l'ondelette. Afin d'éviter des problèmes de bords, $f[n]$ et $\Psi[n]$ sont traités comme des signaux de période N.
La transformée en ondelettes discrète peut alors s'écrire comme une convolution circulaire avec $\overline{\Psi}_j[n] = \Psi_j^*[-n]$:

$$Wf[n, a^j] = \sum_{m=0}^{N-1} f[m]\Psi_j^*[m-n] = f \circledast \overline{\Psi}_j[n]$$

où Ψ^* est le conjugué complexe de Ψ et $\circledast$ est l'opérateur de convolution circulaire.

Si nous prenons le cas où l'échelle est découpée selon une suite dyadique $\{2^j\}_{j\in\mathbb{Z}}$, c'est à dire lorsque le paramètre d'échelle est $a^j = 2^j$, alors la transformée en ondelettes discrète et dyadique s'écrit :

$$Wf[n, 2^j] = \sum_{m=0}^{N-1} f[m]\Psi_{2^j}^*[m-n] = f \circledast \overline{\Psi}_{2^j}[n]$$

avec

$$\Psi_{2^j}[n] = \frac{1}{\sqrt{2^j}}\Psi(\frac{n}{2^j}).$$

La figure 3.4 montre la décomposition temps-fréquence du signal en utilisant la transformée en ondelettes dyadique. La transformée dyadique de f ne peut être calculée que pour des échelles $1 > 2^j \geq \frac{1}{N}$. La valeur absolue de j sera utilisée par la suite pour représenter les différentes échelles dans l'analyse multi-résolution ainsi que les différentes bandes de fréquence.

L'utilisation de la transformée en ondelettes dyadique nous permet d'obtenir une partition dyadique du plan temps-fréquence de telle sorte que les basses fréquences sont représentées avec une haute résolution fréquentielle et une faible résolution temporelle alors que les hautes fréquences sont représentées avec une haute résolution temporelle et une faible résolution fréquentielle (Figure 3.4). La résolution temporelle est inversement proportionnelle à la résolution fréquentielle à cause du principe d'incertitude d'Heisenberg. Cette partition permet d'avoir une résolution fréquentielle qui se rapproche de celle de l'oreille humaine, une analyse fine des basses fréquences et qui diminue de manière logarithmique lorsque l'on monte en fréquence (Figure 3.5). C'est une approximation de l'échelle Mel, très utilisée en reconnaissance de la parole et notamment avec les MFCC.

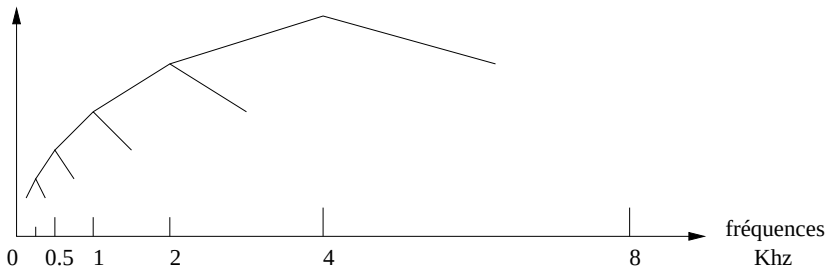

FIG. 3.5 – Résolution fréquentielle obtenue à l'aide de la décomposition en ondelettes dyadique. (arbre de décomposition dyadique avec 5 niveaux de décomposition)

3.1.4 Algorithme rapide pour la transformée en ondelettes

Mallat [Mallat 98] a montré que les coefficients de la décomposition du signal sur une base orthonormée d'ondelettes se calculent par un algorithme rapide (algorithme pyramidal) qui cascade des convolutions discrètes avec des filtres passe-bas (G) et passe-haut (H) dont les sorties sont sous-échantillonnées.
Dans notre cas, les coefficients de décomposition du signal par la transformée en ondelettes dyadique sont obtenus par filtrage successif passe-haut (H) et passe-bas (G) de la sortie du filtre passe-bas (G). Les sorties des filtres sont sous-échantillonnées par un facteur 2. L'algorithme est illustré à la figure 3.6.
Ces bancs de filtres implémentent une transformée rapide en ondelettes orthogonales, qui ne nécessite que $O(N)$ calculs pour un signal de taille N.
Le symbole "$\downarrow 2$" correspond au sous-échantillonnage par un facteur 2. La figure montre qu'à chaque niveau de décomposition j, le signal est décomposé en coefficients *d'approximation* $a_j(k)$ (la sortie du filtre passe-bas) et en coefficients de *détails* $w_j(k)$ (la sortie du filtre passe-haut (H)).
Les coefficients d'approximation correspondent à des moyennes locales du signal tandis que les coefficients de détails, aussi appelé coefficients d'*ondelettes*, dépeignent les différences entre deux moyennes locales successives, c'est-à-dire entre deux approximations

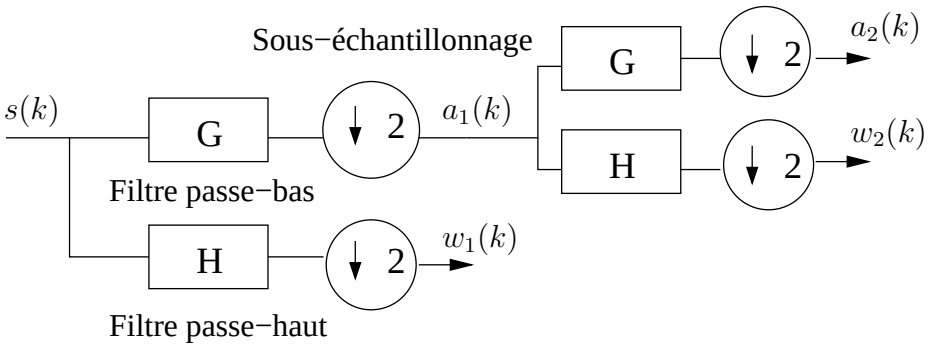

FIG. 3.6 – Transformée en ondelettes dyadique avec 2 niveaux de décomposition.

successives du signal. D'une manière plus imagée, les coefficients d'approximation donnent une représentation lissée du signal et les coefficients d'ondelettes (de détails) nous donnent les détails (le bruit) qui ont été supprimés lors du lissage. Il est tout à fait possible de reconstruire le signal de départ à partir de ces coefficients d'approximation et de détails. Pour notre tâche de discrimination parole/musique, nous proposons de n'utiliser que les coefficients d'ondelettes $w_j(k)$ pour analyser le signal acoustique. Ces coefficients correspondent à la variation du signal autour de sa valeur moyenne. Les coefficients d'ondelettes capturent ainsi les modifications soudaines du signal (les transitoires).
Enfin, nous pouvons noter que contrairement à la transformée de Fourier, la localisation temporelle des fréquences n'est pas perdue et c'est là un autre avantage de la transformée en ondelettes pour la discrimination parole/musique.

3.2 Types d'ondelettes utilisées

Il existe un nombre très important de types d'ondelettes que l'on appelle aussi familles. Cette richesse dans le choix de la base d'ondelettes, c'est-à-dire le choix des fonctions analysantes, est aussi l'un des intérêts de la transformée en ondelettes. Parmi la multitude de familles d'ondelettes qui ont été proposées, nous pouvons citer, par exemple, les coiflet, les symlet, les ondelettes de Daubechies, les ondelettes bi-orthogonales, l'ondelette de Haar, etc.
Lors de notre étude, nous nous sommes limités à trois familles d'ondelettes bien connues en traitement du signal : les ondelettes de Daubechies, les Symlets et les Coiflets (cf. Figures 3.8, 3.9, 3.10). Ces ondelettes sont toutes admissibles, selon le théorème 1, car de moyenne nulle et à décroissance rapide. De plus, elles ont déjà été étudiées en reconnaissance de la parole et ont donné de bons résultats [Deviren 03, Gemello 01]. Enfin, elles ont toutes la propriété d'avoir un support minimum pour un nombre de moments nuls donné. Avant d'aller plus loin, définissons les deux caractéristiques que nous venons de citer : le nombre de moments nuls et la taille du support d'une ondelette. Ces deux caractéristiques

importantes sont généralement prises en compte dans le choix d'une ondelette.

Les moments nuls :
Le nombre de moments nuls d'une ondelette s'exprime de la manière suivante :

$$\int_{-\infty}^{+\infty} t^k \Psi(t)\, dt = 0, \text{ pour } 0 \leq k < p.$$

Si une ondelette Ψ vérifie cette équation alors on dit que l'ondelette Ψ a p moments nuls. Cela signifie que Ψ est orthogonale à tout polynôme de degré $p-1$. L'intérêt d'avoir p moments nuls est d'obtenir des coefficients d'ondelettes w_j proches de 0 aux échelles fines 2^j (lorsque 2^j tend vers 0). En effet, si $f(t)$ est localement de classe C^k alors $f(t)$ est localement bien "approximé" par un polynôme de Taylor de degré k, et si $k < p$ alors les ondelettes seront orthogonales à ce polynôme. La transformée en ondelettes aura donc des valeurs proches de 0. A contrario, quand $f(t)$ ne pourra être approximé correctement que par des polynômes de degré supérieur à p, alors la transformée en ondelettes aura de fortes amplitudes. Cette propriété est très utile pour détecter les transitions brutales. En effet, les zones stationnaires d'un signal correspondront à de petits coefficients d'ondelettes, et les transitions brutales à de grands coefficients.

Taille du support :
Si $f(t)$ a une singularité isolée en t_0, et si t_0 est dans le support de l'ondelette Ψ_j, alors la transformée en ondelettes aura des coefficients d'ondelettes de forte amplitude autour de t_0. Si l'ondelette Ψ a un support de taille K, alors à haute résolution, c'est-à-dire aux fines échelles : lorsque l'échelle s tend vers 0, il y aura K ondelettes Ψ_j dont le support contiendra t_0. L'idée est de minimiser la taille du support de Ψ dans le but de diminuer le nombre de coefficients d'ondelettes de grande amplitude. Cela permet ainsi de faire de la détection de singularités.

Ces deux caractéristiques ne sont pas indépendantes. En effet, la taille du support et le nombre de moments nuls d'une ondelette orthogonale sont liés par le fait que si Ψ a p moments nuls alors son support est au moins de taille $2p-1$. Lors du choix d'une ondelette, on doit donc faire un compromis entre la taille du support et le nombre de moments nuls. Si $f(t)$ a peu de singularités isolées, et est très régulier entre ces singularités, il est plus approprié de choisir une ondelette ayant de nombreux moments nuls afin d'obtenir un grand nombre de coefficients d'ondelettes de petite amplitude. Lorsque la densité de singularités augmente, il vaut mieux diminuer la taille du support, quitte à avoir moins de moments nuls. En effet, les ondelettes dont le support passe par une singularité donnent des coefficients de grande amplitude.
Pour le choix des ondelettes, il faut aussi noter qu'en utilisant la transformée en ondelettes discrète, nous nous restreignons à n'utiliser que des ondelettes à filtres. En effet, seules les

ondelettes à filtres peuvent être utilisées avec la transformée discrète, alors que dans le cas continu n'importe quelle fonction d'intégrale nulle convient. Ainsi, les ondelettes utilisées sont définies directement par leurs filtres associés (filtres passe-bas et passe-haut). En fait, l'ondelette n'est pas toujours directement accessible, c'est à dire qu'aucune formule analytique ne la définit, comme par exemple l'ondelette de Daubechies. Il est toutefois possible d'obtenir d'excellentes approximations de l'ondelette définie implicitement, en utilisant un algorithme déduit de l'algorithme de reconstruction de Mallat [Mallat 00]. Les filtres correspondant aux ondelettes que nous utilisons ont été construits à l'aide du logiciel "Matlab". La figure 3.7 représente la réponse impulsionnelle des filtres associées à l'ondelette de Daubechies "db4". Tous ces éléments nous ont permis de choisir nos

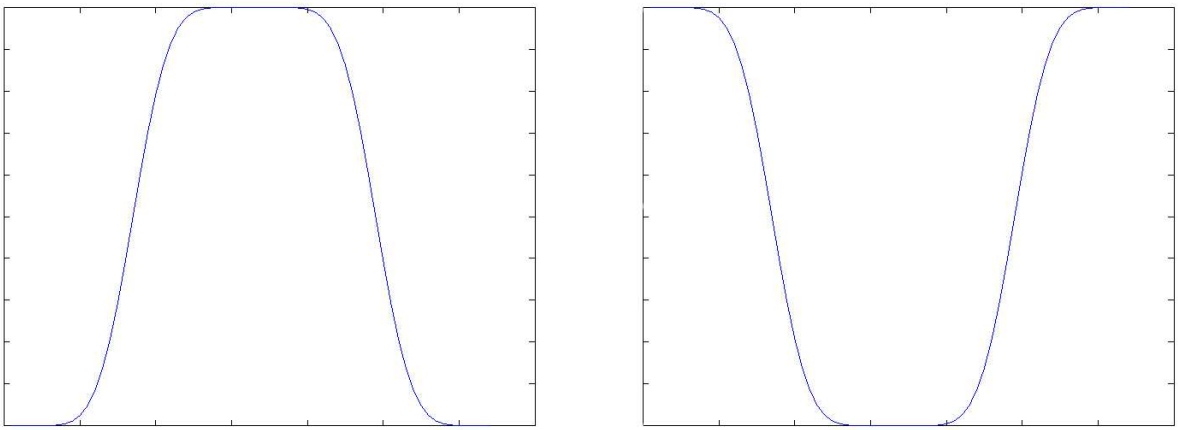

FIG. 3.7 – Représentation en module dans le domaine des fréquences des effets des filtres d'analyse passe-haut (à gauche) et passe-bas (à droite) associés à l'ondelette 'db4' (module de la transformée de Fourier).

ondelettes pour la tâche de discrimination parole/musique. Nous présentons maintenant plus en détails les trois familles d'ondelettes choisies.

3.2.1 Les ondelettes de Daubechies

Cette famille d'ondelettes a été créée par Ingrid Daubechies [Daubechies 92]. Nous noterons les ondelettes de cette famille dbN où N est l'ordre de l'ondelette. Nous retrouvons dans cette famille l'ondelette de Haar correspondant à $db1$ et qui est la plus simple et certainement la plus ancienne des ondelettes.

Exceptée $db1$, les ondelettes de cette famille n'ont pas d'expression explicite. Cette famille possède certaines propriétés intéressantes. Le nombre de moments nuls de l'ondelette dbN est N. Les ondelettes de Daubechies ont un support de taille minimale pour un nombre de moments nuls donné. Les ondelettes de Daubechies sont très asymétriques, en particulier pour les faibles valeurs de N, sauf pour $db1$.

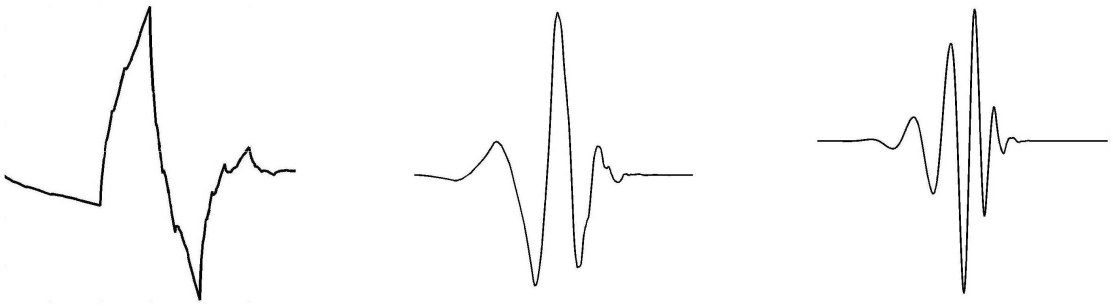

FIG. 3.8 – Exemples d'ondelettes de Daubechies : de gauche à droite nous avons 'db2', 'db4' et 'db8'.

3.2.2 Les Symlets

Les Symlets, notées $symN$, ont été proposées par Daubechies en modifiant la construction des ondelettes dbN et constituent une famille d'ondelettes presque symétrique.
A part la symétrie, les propriétés de ces deux familles sont similaires. En regardant les figures des ondelettes de Daubechies et les Symlets, nous pouvons constater que la Symlet ressemble à une ondelette de Daubechies pour un nombre de moments nuls petit, et qu'elle est plus symétrique que sa consoeur.

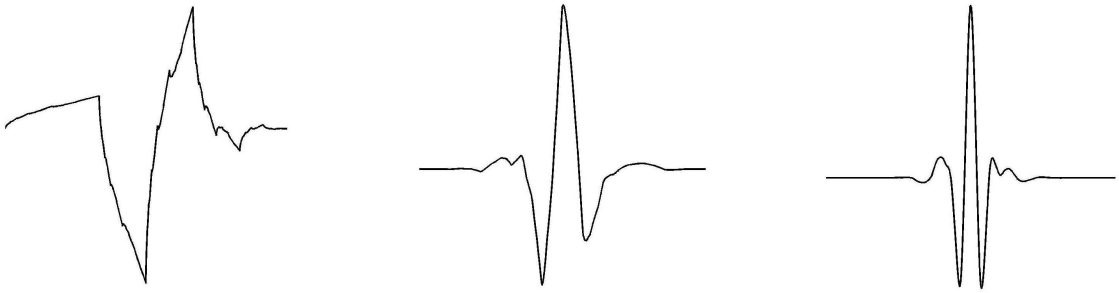

FIG. 3.9 – Exemples de Symlets : de gauche à droite nous avons 'sym2', 'sym4' et 'sym8'.

3.2.3 Les Coiflets

Les Coiflets, comme les symlets, ont été construites par Daubechies. Elles ont été créées sur la demande de R. Coifman [Daubechies 88] pour une application liée à l'analyse numérique. Nous prendrons comme notation de cette famille d'ondelettes : $coifN$.
Cette famille d'ondelettes est différente des deux précédentes, ici, l'ondelette $coifN$ aura $2N$ moments nuls. Toutefois, les Coiflets, comme nous pouvons le voir sur la figure 3.10, sont bien plus symétriques que les Symlets ou les ondelettes de Daubechies. L'intérêt principal des coiflets réside dans le fait que si nous analysons une fonction f assez régulière, alors les coefficients d'approximation (pour un nombre de niveaux de décomposition assez grand) correspondent à l'échantillonnage de f.

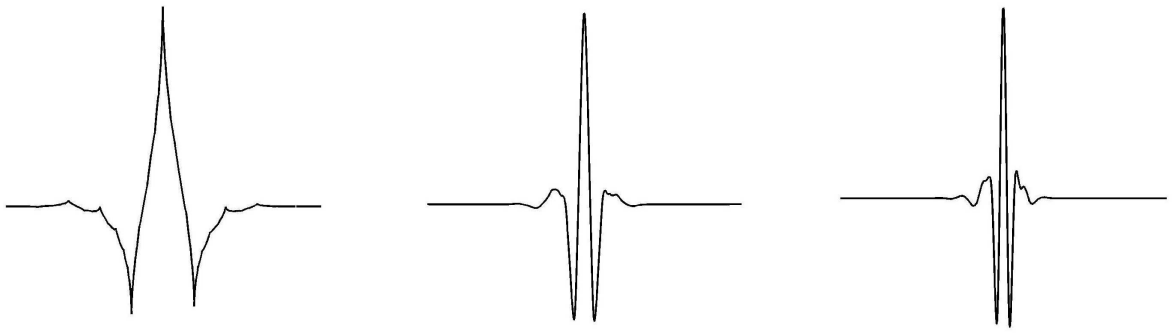

FIG. 3.10 – Exemples de Coiflets : de gauche à droite nous avons 'coif1', 'coif3' et 'coif5'.

3.3 Types d'énergies calculées sur les coefficients d'ondelettes

Que se soit en reconnaissance de la parole ou en discrimination parole/musique, l'énergie du signal est très souvent utilisée en tant que paramètre et donne de plus de bons résultats. C'est pourquoi, nous avons décidé d'utiliser l'énergie, calculée sur les coefficients d'ondelettes obtenus à partir de la transformée en ondelettes dyadique, comme paramètre pour notre tâche de discrimination. La nécessité d'utiliser l'énergie est aussi dûe au fait que les coefficients d'ondelettes sont trop nombreux dans chacune des bandes de fréquences (ou niveaux de décomposition) pour être utilisé directement.
Nous ne nous sommes pas limités à un seul type d'énergie mais nous avons choisi trois énergies aux propriétés différentes.

Dans ce qui suit, w_j^k dénote le coefficient d'ondelettes à la position temporelle k et à la bande de fréquences j. Nous rappelons que les décompositions temporelles et en bandes de fréquences suivent une échelle dyadique, c'est-à-dire que la résolution temporelle est divisée par deux alors que la résolution fréquentielle double à chaque niveau de décomposition. Le nombre de coefficients dans la bande j est noté N_j. Nous calculons finalement, à partir de l'ensemble des coefficients d'ondelettes w_k^j pour la bande de fréquences j, différents paramètres f_j pour cette bande de fréquences j en utilisant différents types d'énergie :

- *L'énergie instantanée* (notée **E**) :
  Ce type d'énergie, classiquement utilisé dans le domaine de la parole, nous donne la distribution de l'énergie dans chacune des bandes.

$$f_j = \log_{10}\left(\frac{1}{N_j}\sum_{k=0}^{N_j-1}(w_k^j)^2\right)$$

– *L'énergie de Teager* (notée **T_E**) :
Nous proposons ici d'utiliser l'opérateur discret d'énergie de Teager (*The discrete Teager Energy Operator* ou *TEO*) introduit par Kaiser [Kaiser 90]. Cet opérateur permet de calculer d'une façon simple l'énergie d'un signal et de pouvoir estimer son amplitude et sa fréquence instantanée (démodulation). Cet opérateur a été récemment utilisé en reconnaissance de la parole [Dimitriadis 05]. Il nous permet de suivre les modulations d'énergie et donne une meilleure représentation de l'information formantique du signal dans le vecteur paramètre, comparé aux MFCC [Jabloun 99]. Il permet aussi une réduction du bruit du signal en utilisant sa capacité de suivi de la modulation d'énergie.

$$f_j = \log_{10}\left(\frac{1}{N_j - 1}\sum_{k=1}^{N_j-1} |(w_k^j)^2 - w_{k-1}^j w_{k+1}^j|\right)$$

– *L'énergie hiérarchique* (notée **H_E**) :
Nous calculons ici des paramètres basés sur l'énergie mais avec une résolution temporelle hiérarchique. L'énergie hiérarchique correspond au calcul de l'énergie au centre de la fenêtre d'analyse en prenant le même nombre de coefficients dans toutes les bandes :

$$f_j = \log_{10}\left(\frac{1}{N_J}\sum_{k=(N_j-N_J)/2}^{(N_j+N_J)/2} (w_k^j)^2\right)$$

J correspond à la bande la plus basse.
Le choix de ce calcul s'explique par le fait que les coefficients d'ondelettes ont une résolution temporelle plus fine dans les hautes fréquences. Ils recouvrent des intervalles de temps de plus en plus petits lorsque l'on monte en fréquence alors que lorsque l'on descend dans les basses fréquences les coefficients d'ondelettes vont recouvrir des zones temporelles de plus en plus grandes. Le nombre de coefficients est donc différent d'une bande à l'autre, un grand nombre dans les hautes fréquences et un petit nombre dans les basses fréquences. La technique de résolution temporelle hiérarchique extrait les caractéristiques concentrées au centre de la fenêtre d'analyse, en prenant le même nombre de coefficients pour toutes les bandes.
Ce type d'énergie a été utilisé avec succès en reconnaissance automatique de la parole pour paramétriser le signal [Gemello 01].

3.4 Conclusion

Un signal audio de parole ou de musique est par essence non stationnaire, la nécessité d'une analyse temps-fréquence a été reconnue de longue date. Nous avons vu que pour analyser des signaux non stationnaires, une solution consiste à utiliser une variante de la transformée de Fourier classique : la transformée de Fourier à fenêtre aussi appelée transformée de Fourier à court terme. Cependant, cette solution a des limites, notamment dans le choix de la taille de la fenêtre d'analyse qui détermine si nous nous concentrons sur une analyse fréquentielle du signal ou sur une analyse des événements temporels. Une réponse intéressante à ce problème est la transformée en ondelettes. Comme la transformation de Fourier, la transformation en ondelettes fournit une représentation temps-fréquence du signal, mais avec une **résolution variable**. Nous pouvons ainsi effectuer une analyse multi-résolution du signal et ainsi étudier plus finement les détails du signal en l'observant à différentes échelles. Tout en respectant le principe d'incertitude d'Heisenberg, le spectrogramme en ondelettes semble plus proche de la représentation du signal vocal fournie par la cochlée [Irino 93]. Il donne à la fois des informations sur la structure harmonique du signal (dans les basses fréquences) et sur sa structure formantique (dans les fréquences plus hautes), ce qui n'est pas le cas du spectre de Fourier. Nous avons ainsi défini la méthode de décomposition du signal à l'aide de la transformée en ondelettes discrète et dyadique. Cette méthode est basée sur l'utilisation de filtres passe-haut et passe-bas en cascade. Une fois le signal décomposé en coefficients d'approximation et en coefficients de détails, plus couramment appelés coefficients d'ondelettes, nous calculons différents types d'énergie uniquement sur les coefficients d'ondelettes. Les différentes énergies calculées sur les coefficients d'ondelettes sont à la base de nos vecteurs de paramètres pour la tâche de discrimination parole/musique. Lorsqu'une énergie est choisie comme paramètre, elle est calculée sur toutes les bandes de fréquences.

4

Description des corpus utilisés

Sommaire

Dans ce chapitre, nous présentons les différents corpus utilisés pour développer et tester notre système de segmentation parole/musique. Les différentes expérimentations sont décrites au chapitre suivant. De nombreuses heures de signal audio étiquetées ont été nécessaires pour construire ces corpus.
Chaque trame de parole est étiquetée suivant le contenu du segment auquel elle appartient. Les trames peuvent ainsi être étiquetées en :

- parole (P), si le segment ne contient que de la parole,
- musique (M), si le segment ne contient que de la musique ou de la chanson,
- parole sur musique (PM), dans le cas où le segment est composé de parole sur un fond musical.

Trois types de corpus sont nécessaires au développement et au test de notre méthode. Le premier, le corpus d'apprentissage, permet d'apprendre nos modèles acoustiques pour classer les segments en parole ou en musique. Le second, le corpus de développement, permet d'évaluer notre système lors de la phase de développement et ainsi de sélectionner les meilleurs paramètres. Le dernier corpus est un corpus de validation permettant de confirmer ou d'infirmer les résultats obtenus pendant la phase de développement. Nous les présentons à présent en détails.

4.1 Corpus d'apprentissage

Le corpus d'apprentissage se divise en deux parties. Toutes les données sont échantillonnées en 16kHz mono.
La première partie, nommée "CDs Audio", est composée de différentes musiques et chansons extraites de CDs audio. Les échantillons extraits, pris en milieu de piste, ont une durée de 15 secondes. La durée totale de cette partie est de deux heures, se décomposant en trente minutes de musique instrumentale et une heure et trente minutes de chansons. La musique instrumentale provient de genres musicaux très variés tel le jazz, la musique classique, la musique électronique, le rock, etc. Les chansons sont, quant à elles, principalement de styles pop et rock.
La seconde partie du corpus d'apprentissage, intitulée "Corpus radio réel", est constituée majoritairement d'émissions radiophoniques francophones mais contient quelques éléments hispanophones (environ 60 minutes). L'ensemble de ce corpus radiophonique a une durée cumulée de 976 minutes. Les enregistrements ont été étiquetés en parole et musique dans le cadre des projets RAIVES[13] et ESTER[14]. Parmi les programmes, nous retrouvons aussi bien des journaux que des émissions thématiques avec interviews ou encore des programmes musicaux. Cette partie est très variée, aussi bien en locuteurs qu'en conditions d'enregistrement. Au niveau des locuteurs nous avons des hommes, des femmes, des enfants, de nationalités variées. Pour les conditions d'enregistrement, la parole enregistrée en studio alterne avec de la parole téléphonique, et certaines interviews ont été réalisées dans des environnements bruités (bruits de rue, de café, "cocktail party", traduction simultanée). Enfin, cette partie de corpus contient de nombreux segments de parole sur un fond musical (musique, chanson, jingle). La répartition des durées de parole et musique du corpus d'apprentissage est indiquée dans le tableau 4.1.

	Parole	Parole/musique	Musique
CDs Audio	-	-	120 mn
Corpus radio réel	846 mn	104 mn	26 mn
Total	846 mn	104 mn	146 mn

TAB. 4.1 – *Répartition des données (parole, musique, parole/musique) dans le corpus d'apprentissage.*

[13]Recherche Automatique d'Informations Verbales Et Sonores
[14]Evaluation des Systèmes de Transcription d'Emissions Radiophoniques

4.2 Corpus de développement

Le corpus de développement est composé de trois sous-corpus très différents permettant de tester notre système de discrimination parole/musique dans différentes conditions. La répartition parole, musique, parole sur musique de ces corpus ainsi que leur durée exprimée en secondes sont données dans la table 4.2.

Le premier sous-corpus, *Scheirer*, est un corpus anglais développé par E. Scheirer sous la supervision de M. Slaney [Scheirer 97]. Ce corpus est constitué de 240 fichiers de 15 secondes collectés aléatoirement à la radio. Le corpus est divisé en deux parties : "apprentissage" et "test". Nous n'utilisons ici que la partie "test" constituée de 61 fichiers homogènes : 20 fichiers contenant de la parole studio ou téléphonique, 21 fichiers contenant de la musique (sans chansons) et 20 fichiers de chansons (de la musique et des voix). Nous n'exploitons pas le fait que les fichiers soient homogènes, notre système de discrimination resegmente ces fichiers puis les classe. Ces fichiers audio ont été collectés sur des stations de radios américaines de manière aléatoire. Ils contiennent des styles d'enregistrement variés de parole, de musiques et de chansons. Nous avons ainsi, au niveau musiques et chansons, un grand éventail de styles : jazz, rock, pop, country, classic, etc.
Ce corpus nous permet d'évaluer nos nouvelles paramétrisations sur un corpus qui a été très fréquemment utilisé dans de précédentes études. L'intervalle de confiance est de $\pm 1\%$.

Le second sous-corpus, *News*, est composé de trois fichiers d'une heure provenant de programmes de radios françaises : "France-Inter" et "Radio France International". Ces programmes radios contiennent essentiellement de la parole et de la parole sur des jingles, comme c'est souvent le cas lors de la présentation des titres à développer lors du journal. Ce corpus est intéressant dans le sens où nous pouvons tester notre système de discrimination parole/musique comme si nous nous placions dans le cadre d'une application de transcription de journaux radiophoniques. L'intervalle de confiance est ici de $\pm 0.5\%$.

	Parole	Parole/musique	Musique	Durée totale
Scheirer	32%	0%	68%	15m 15s
News	86%	11%	3%	3h
Entertainment	52%	18%	30%	1h 12m
TestRTL	66.4%	9.1%	24.5%	1h

TAB. 4.2 – *Répartition des données (parole, musique, parole/musique) pour les différents corpus de développement : Scheirer, News, Entertainment, et pour le corpus de validation : TestRTL. La durée totale de chaque corpus est également donnée.*

Le troisième sous-corpus, *Entertainment*, est composé de trois émissions d'une vingtaine de minutes chacune. Ces émissions comprennent des interviews et des programmes

musicaux, ce qui fait que nous pouvons considérer ce corpus comme difficile. En effet, il y a de nombreux segments où différents événements sonores (musique, chansons) sont superposés à la parole avec des effets de fondu (*fade in-fade out*). De plus, certaines interviews sont très bruitées (enregistrement dans la rue par exemple) et une alternance de qualité de parole (téléphonique / studio) est observée dans les différents enregistrements. L'intervalle de confiance pour ce sous-corpus est de $\pm 1\%$.

4.3 Corpus de validation

Un dernier corpus dit de validation, *TestRTL*, a été créé pour tester notre meilleure paramétrisation et valider nos résultats. Ce corpus est constitué d'une heure de programmes radiophoniques, provenant de la radio française "RTL". Ce corpus est assez difficile pour la tâche de segmentation automatique en parole et musique car il contient des chansons, des publicités, des applaudissements, des rires, des interviews, etc. La répartition parole, musique, parole sur musique de ce corpus et sa durée en secondes sont données dans la table 4.2. L'intervalle de confiance pour ce dernier corpus de validation est de $\pm 1\%$.

5

Protocole expérimental et résultats

Sommaire

Ce chapitre présente les différentes expérimentations effectuées pour valider la nouvelle paramétrisation proposée dans cette thèse, pour la discrimination Parole/Musique. Nous présentons en détails l'approche choisie pour notre système de classification Parole/Musique ainsi que les mesures d'évaluation de notre système. Enfin, nous détaillons les résultats obtenus au fil des expérimentations et nous concluons sur la nouvelle paramétrisation basée sur la décomposition en ondelette du signal.

5.1 Système de classification

Cette étape de classification consiste à classer, à partir des vecteurs acoustiques obtenus lors de l'étape de paramétrisation, le signal audio en différentes catégories. Ces catégories sont ici au nombre de trois : parole (notée P), musique (notée M) et parole sur musique (notée PM). Une dernière catégorie existe toutefois lorsque ni la parole, ni la musique n'est reconnue : le bruit (B). Cette dernière catégorie ne fait pas l'objet de notre étude.
Habituellement dans les systèmes de transcription, la segmentation parole/musique permet de conserver la parole et de rejeter tout le reste sans faire de différence entre la musique et les autres événements sonores. Mais en présence de parole sur fond musical, cette approche peut être amenée à rejeter des segments de parole mal reconnus à cause de la musique. Une solution consiste à modéliser la parole, la musique et la parole sur de la musique dès le départ pour éviter de perdre ces derniers segments de parole sur musique. Une première approche consiste à modéliser précisément les événements sonores que l'on souhaite classer (parole, musique instrumentale, musique chantée, parole sur musique) et à mettre ces modèles en concurrences (en "compétition") sur le même espace de représentation. C'est ce que l'on appelle l'approche "compétitive" [Razik 04]. Notre approche est quelque peu différente. Nous utilisons une approche dite "classe/non-classe" [Pinquier 02b].

5.1.1 Approche Classe/Non Classe

L'approche de classification "classe/non-classe" consiste à assigner une classe au signal audio en comparant un modèle "classe" et un modèle "non-classe" sur le même espace de représentation. Notre système se décompose donc en deux sous-systèmes de classification

indépendants permettant de classer séparément les segments de parole et de musique (Figure 5.1). Le premier sous-système classifie le signal audio en segments de parole et de ce que l'on peut appeler de la "non-parole", c'est-à-dire tout ce qui n'est pas de la parole. Le second sous-système classe quant à lui le signal audio en segments de musique et "non-musique".

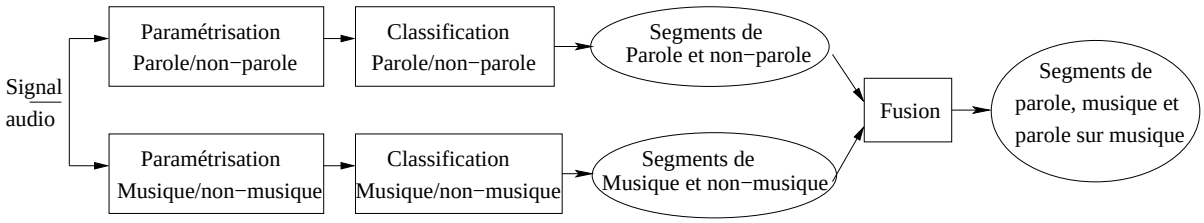

FIG. 5.1 – Système de segmentation parole/musique.

L'intérêt de cette approche par rapport à l'approche "compétitive" est qu'en séparant la classification de la parole et celle de la musique, nous pouvons utiliser des paramétrisations différentes pour la discrimination P/NP et pour la discrimination M/NM. Nous pouvons ainsi améliorer la discrimination parole/non-parole sans détériorer la discrimination musique/non-musique et inversement.

5.1.2 Modélisation à l'aide de HMMs

Pour chacun des classifieurs exposés précédemment, deux modèles de mélanges de gaussiennes (*GMMs*) sont entraînés, avec de 8 à 64 gaussiennes par état, pour modéliser la "classe" et la "non-classe". Nous avons choisi d'utiliser des GMMs car ce sont des modèles robustes très souvent utilisés dans le domaine de la reconnaissance des formes. De plus, Scheirer [Scheirer 97] a testé différents classifieurs et a montré que ce type de modèles était bien adapté à la discrimination parole/musique.
Au final, nous obtenons deux GMMs modélisant :

- la musique (M)
- la non-musique (NM)

et deux GMMs modélisant :

- la parole (P)
- la non-parole (NP)

Ces GMMs vont permettre de classer le signal trame par trame dans les différentes catégories (M, NM, P, NP). Cependant ces GMMs nous donnent une décision trame par trame,

ce qui peut mener à l'obtention de segments de parole ou de musique d'une durée de 10ms. Ceci est totalement irréaliste. En effet, si nous prenons par exemple le cas de la parole, en prononçant au minimum un mot, nous aurons une durée d'environ une demi-seconde. De même pour la musique, il est rare d'avoir des morceaux de musique de moins d'une demi-seconde, même dans le cas de jingles.
Afin d'éviter ce problème de segments trop courts, nous avons imposé une durée minimale de 0.5s pour chacun des segments reconnus. Cette contrainte de durée est directement codée dans le modèle, en concaténant 50 GMMs (si on suppose des trames de 10ms). Nous obtenons ainsi un HMM à 50 états à partir du GMM que nous avons appris, pour chacune des classes modélisées (voir Figure 5.2).

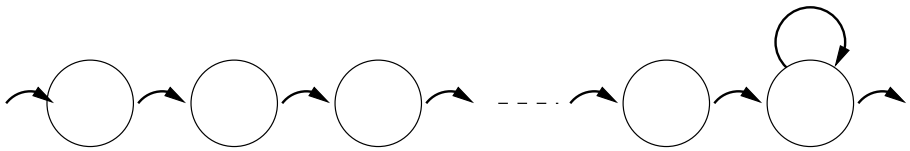

FIG. 5.2 – Topologie du HMM utilisé dans notre système de classification parole/musique

5.1.3 Décision

Pour chacun des classifieurs, l'algorithme de Viterbi permet d'obtenir la meilleure séquence de modèles décrivant le signal. La classification se fait trame par trame. Une dernière étape de fusion ou regroupement des résultats des deux systèmes de classification nous permet d'obtenir les segments de parole, musique et parole sur fond musical. Cette fusion s'effectue selon la table 5.1. Nous pouvons remarquer que des segments de bruits et de silences peuvent être obtenus lorsque les classifieurs classent les segments en "non-parole" et "non-musique".

classifieur P/NP	classifieur M/NM	étiquetage final du segment
Parole	Non-Musique	**Parole**
Parole	Musique	**Parole sur musique**
Non-Parole	Non-Musique	**Bruit/Silence**
Non-Parole	Musique	**Musique**

TAB. 5.1 – *Fusion des résultats des deux sous-systèmes de classification trame par trame pour l'étiquetage final des trames du signal audio*

5.2 Paramétrisation de référence

Pour comparer nos résultats, nous utilisons la paramétrisation la plus souvent utilisée en discrimination parole/musique dans les systèmes de transcription automatique : la paramétrisation MFCC. Nous avons donc choisi de prendre, comme paramètres de base, 12 coefficients cepstraux (dont $C0$) et leurs premières (Δ) et secondes dérivées ($\Delta\Delta$). Ces coefficients sont estimés toutes les 10 millisecondes sur une fenêtre d'analyse de Hamming de 32 millisecondes. Cette paramétrisation est très fréquemment utilisée pour séparer la parole de la musique dans les grands systèmes de transcription automatique de la parole [T.Hain 98, Gauvain 02, Fredouille 04] et donne de bons résultats dans cette tâche [Razik 04, Logan 00, Scheirer 97].

5.3 Mesures d'évaluation

Pour évaluer nos différents paramètres, trois taux d'erreurs sont calculés trame par trame de la manière suivante :

- Taux d'erreur en classification globale (TG) :

$$100 * (1 - (n_{PM}^{PM} + n_{M}^{M} + n_{P}^{P})/T) \tag{5.1}$$

- Taux d'erreur en classification Musique/Non-Musique (M/NM) :

$$100 * (1 - (n_{PM}^{M} + n_{M}^{PM} + n_{M}^{M} + n_{PM}^{PM} + n_{P}^{P})/T) \tag{5.2}$$

- Taux d'erreur en classification Parole/Non-Parole (P/NP) :

$$100 * (1 - (n_{PM}^{P} + n_{P}^{PM} + n_{M}^{M} + n_{PM}^{PM} + n_{P}^{P})/T) \tag{5.3}$$

avec n_z^y le nombre de trames reconnues comme étant de la classe z alors qu'elles sont étiquetées y, et T le nombre total de trames.

5.4 Expérimentations avec nos nouvelles paramétrisations en ondelettes

Nous décrivons ici les différentes expériences que nous avons réalisées. Le but de ces expériences est de trouver la meilleure paramétrisation pour les tâches de discrimination parole/non-parole et musique/non-musique.
Nous avons ainsi testé l'influence de différentes paramétrisations basées sur la décomposition en ondelettes du signal. Comme pour la paramétrisation MFCC, les paramètres sont extraits toutes les 10millisecondes en utilisant une fenêtre d'analyse de 32 millisecondes.

Nous détaillons ici notre démarche expérimentale.
La première expérience permet de choisir la fonction ou les fonctions d'ondelettes pour chacune des tâches de discrimination : parole/non-parole, musique/non/musique.
La seconde expérience permet de déterminer quel va être la meilleure résolution, c'est à dire le meilleur niveau de décomposition, et la meilleure énergie pour les différentes tâches de discrimination.
Les deux expériences suivantes permettent d'évaluer l'influence de la dynamique des paramètres sur la discrimination parole/non-parole et la discrimination musique non-musique. Cette dynamique est soit à court terme (dérivées premières et secondes), soit à long terme (variance sur une fenêtre d'une seconde).
Une dernière expérience a consisté à combiner les sorties de plusieurs classifieurs classe/non-classe pour tenter d'améliorer encore la discrimination parole/musique.
Toutes les expérimentations ont été réalisées sur les corpus "Scheirer", "News" et "Entertainment".
Finalement, un test de validation a été effectué sur le corpus "$TestRTL$" en utilisant les meilleures paramétrisations Parole/Non-parole et Musique/Non-musique obtenues auparavant. Nous avons aussi expérimenté une combinaison entre les paramètres en ondelettes et les paramètres MFCC et une fusion de classifieurs par vote majoritaire.

5.4.1 Choix de l'ondelette

Le but de cette expérience est de sélectionner les ondelettes les plus adéquates pour les différentes tâches de discrimination, parole/non-parole et musique/non-musique. Ce choix ne peut être fait qu'empiriquement. Nous ne pouvons pas prédire, en regardant ses propriétés mathématiques, si une ondelette est meilleure pour telle ou telle tâche. Il existe de nombreuses familles d'ondelettes. Mais nous nous sommes limités aux ondelettes utilisables par l'algorithme rapide à base de bancs de filtres : les ondelettes orthogonales. Nous avons ainsi étudié trois familles d'ondelettes, les plus connues et les plus utilisées en traitement du signal : les ondelettes de Daubechies, les Symlets et les Coiflets que nous avons décrites au chapitre 3.
Nous avons donc étudié le comportement des ondelettes de ces trois familles, aux propriétés de lissage différentes, en faisant varier le nombre de moments nuls de ces ondelettes. En effet, nous pouvons voir les ondelettes comme un outil pour lisser un signal en captant ses irrégularités. Ce sont les coefficients d'ondelettes qui correspondent aux variations captées autour du signal "lissé" (l'approximation du signal).
Aussi, nous avons vu au chapitre 3 que le nombre de moments nuls était liée à la taille du support de l'ondelette et que suivant la densité de singularité du signal, il fallait faire un compromis entre ces deux paramètres. A priori, les signaux de musique et de parole ont une grande densité de singularités et nous devrions obtenir les meilleures performances avec des ondelettes ayant très peu de moments nuls. C'est ce que nous allons investiguer.
Dans cette expérience, nous avons défini le nombre de niveaux de décomposition à 5.

Cette valeur a été choisie après l'étude de Deviren [Deviren 04] sur l'influence du nombre de niveau de décomposition avec différentes ondelettes en reconnaissance automatique de la parole. Le résultat de cette étude montrait que 4 à 6 niveaux de décomposition semblaient appropriés pour extraire du signal l'information utile à la reconnaissance. Quelques expérimentations préliminaires ont aussi été effectuées pour confirmer ce choix. Nous y avons fait varier le nombre de niveaux de décomposition de 2 à 8 et il s'est avéré que descendre en dessous de 5 niveaux de décomposition ou dépasser 7 niveaux de décomposition détériore les performances en discrimination parole/musique. Seule l'énergie instantanée calculée sur les coefficients d'ondelettes dans chaque bande est utilisée ici. L'influence des énergies est étudiée dans une seconde étape de tests. Nous avons étudié le comportement des différentes ondelettes tous corpus confondus.

Type d'ondelette	**moments nuls**	**Nb coeffs**	**M/NM**	**P/NP**	**TG**
MFCC+Δ+ΔΔ	-	36	*23.2*	*5.8*	*26.2*
db-2	2	5	11.6	4.9	18.4
db-4	4	5	15.8	**4.6**	18.8
db-8	8	5	16.8	5.0	20.5
db-12	12	5	19.0	5.6	22.4
coif-1	2	5	**11.5**	4.8	**18.0**
coif-3	6	5	16.3	4.9	19.6
coif-5	10	5	19.0	5.6	21.9
sym-2	2	5	11.6	4.9	18.4
sym-4	4	5	16.0	**4.6**	19.1
sym-8	8	5	16.7	5.2	20.7

TAB. 5.2 – *Résultats en discrimination en utilisant différentes ondelettes. Le niveau de décomposition est fixé à 5 bandes et l'énergie instantanée est utilisée. Taux d'erreurs en trames (%) pour les corpus* Scheirer, News *et* Entertainment.

Résultats : Les résultats résumés dans le tableau 5.2 sont organisés de la manière suivante. La première colonne indique le type d'ondelette utilisée : 'db' signifiant Daubechies, 'coif' signifiant Coiflet et 'sym' la famille des Symlets. La deuxième colonne indique le nombre de moments nuls des ondelettes. La troisième colonne indique le nombre de coefficients de la paramétrisation, qui correspond ici au nombre de niveaux (bandes) de décomposition. Les trois dernières colonnes correspondent aux résultats en discrimination musique/non-musique (M/NM), parole/non-parole (P/NP), et en discrimination globale (TG), c'est-à-dire lorsque nous différencions la parole, la musique et la parole sur de la musique. Le score obtenu par les MFCC est indiqué en en-tête du tableau pour avoir une référence. Nous pouvons constater que les meilleurs résultats sont obtenus avec un petit nombre de moments nuls et ceci quelque soit la famille d'ondelette utilisée. Nous

observons aussi que les résultats sont très similaires entre les différentes ondelettes lorsqu'elles ont un nombre de moments nuls égal ou très proche. Une comparaison entre les résultats obtenus avec la paramétrisation MFCC et les paramétrisations en ondelettes nous montre que les paramètres en ondelettes améliorent les performances des MFCC pour les différentes tâches. Nous obtenons un gain significatif de 50% en discrimination musique/non-musique, 20% en discrimination parole/non-parole et 30% en discrimination globale, par rapport aux MFCC.

Conclusions : Nous pouvons conclure à ce niveau que les trois familles d'ondelettes donnent de bons résultats pour les différents types de discrimination. Nous améliorons les résultats obtenus avec la paramétrisation MFCC. Comme nous l'avions prédit, le nombre de moments nuls des ondelettes doit être faible pour la discrimination musique/non-musique. En effet, nous obtenons les meilleures performances avec un petit nombre de moments nuls, c'est-à-dire lorsque les ondelettes ont une grande capacité à détecter les singularités du signal. Plus nous pouvons capter les événements transitoires observés dans les signaux de parole et de musique, plus nous sommes capables de discriminer ces deux classes. Enfin, les différentes familles d'ondelettes donnent des résultats presque identiques. Nous nous contenterons donc d'utiliser les deux ondelettes 'db-2' et 'coif-1' dans les expériences suivantes.

5.4.2 Paramètres statiques : choix du niveau de décomposition et du type d'énergie

Nous évaluons dans cette expérience des paramètres statiques basés sur les ondelettes. Nous étudions des paramètres basés sur différentes énergies. Les énergies (instantanée, Teager, hiérarchique) sont calculées sur les coefficients d'ondelettes issues de la décomposition du signal en ondelettes. Elles sont décrites en détails dans la section 3.3 du chapitre 3.
Comme nous l'avons dit dans la section précédente, nous n'utiliserons que les ondelettes Daubechies 'db-2' et Coiflet 'coif-1'. Deux niveaux de décomposition en ondelettes du signal seront testés : une décomposition à 5 niveaux et une décomposition à 7 niveaux.
Nous présentons séparément les tâches de discrimination parole/non-parole et musique/non-musique. De plus, les résultats en discrimination seront donnés corpus par corpus pour une analyse plus fine de ces résultats.

5.4.2.1 Discrimination Parole/Non-parole

Nous étudions ici différents paramètres statiques pour la tâche de discrimination parole/non-parole. Les résultats de cette expérience sont regroupés dans le tableau 5.3.

Résultats : Les résultats des paramètres basés sur les ondelettes sont comparables à ceux

Ondelette	**Nb Bandes**	**Energie**	**Scheirer**	**News**	**Entertainment**
MFCC+Δ+ΔΔ			**2.5**	**2.9**	*5.8*
db-2	5	E	3.3	3.6	4.3
db-2	5	T_E	3.3	3.2	**4.2**
db-2	5	H_E	3.2	4.6	4.3
db-2	7	E	3.3	6.5	6.9
db-2	7	T_E	3.3	6.4	5.9
db-2	7	H_E	3.3	7.6	5.9
coif-1	5	E	3.3	3.7	**4.2**
coif-1	5	T_E	3.3	3.2	**4.2**
coif-1	5	H_E	3.3	4.4	4.3
coif-1	7	E	3.3	7.4	6.8
coif-1	7	T_E	3.6	6.4	6.1
coif-1	7	H_E	3.3	7.6	6.6

TAB. 5.3 – *Résultats en discrimination parole/non-parole en utilisant différentes énergies, les ondelettes* db-2 *et* coif-1 *ainsi que 5 et 7 niveaux de décomposition en ondelettes. Les résultats sont donnés en taux d'erreurs en trames (%).*

obtenus avec les MFCC. Nous n'obtenons pas d'améliorations sur les corpus "Scheirer" et "News". La seule amélioration apportée par les ondelettes est obtenue sur le corpus "Entertainment" avec un gain significatif de près de 30%. Ce résultat est obtenu avec 'db-2' et 'coif-1' lorsque nous avons une décomposition en 5 bandes, ce qui confirme l'étude de Deviren [Deviren 04]. De plus, nous observons que l'utilisation de différentes énergies n'apporte rien sur le corpus "Scheirer". Par contre, sur les corpus "News" et "Entertainement", l'énergie Teager, en association avec une décomposition en 5 bandes, donne les meilleurs résultats.

Conclusions : La performance des paramètres statiques basés sur les ondelettes pour la tâche de discrimination parole/non-parole est similaire à celle des MFCC. Mais les paramètres basés sur les ondelettes ont une représentation plus compacte, en effet, nous avons des vecteurs à 5 ou 7 composantes pour les paramètres en ondelettes alors que le vecteur de paramètres MFCC a 36 composantes. Le fait que nous n'améliorons les résultats que pour le corpus "Entertainement" peut s'expliquer par l'excellente capacité qu'ont les MFCC pour représenter la parole. La discrimination parole/non-parole sur les corpus "Scheirer", où la parole est nette et bien différenciée de la musique et "News", principalement composé de parole, est en effet très bonne avec les MFCC. Par contre, sur le corpus "Entertainment", où la parole est plus bruitée et où les zones de parole sur un fond musical sont plus présentes, la paramétrisation en ondelettes surpasse la paramétrisation MFCC. C'est donc sur ce type de corpus que notre paramétrisation est la mieux adaptée.

Enfin, les différentes énergies donnent des résultats similaires, contrairement au nombre de bandes. En effet, excepté sur le corpus Scheirer, nous observons une dégradation des performances en discrimination parole/non-parole lorsque nous augmentons le nombre de bandes. Il semble donc que l'utilisation de 5 niveaux de décomposition soit la meilleure solution pour la discrimination parole/non-parole.

5.4.2.2 Discrimination Musique/Non-musique

Nous testons ici différents paramètres statiques pour la tâche de discrimination musique/non-musique de la même manière que pour la tâche de discrimination parole/non-parole. Les résultats de cette expérimentation sont regroupés dans le tableau 5.4.

Ondelettes	**Nb Bandes**	**Energie**	**Scheirer**	**News**	**Entertainment**
MFCC+Δ+ΔΔ			*6.5*	*13.1*	*23.1*
db-2	5	E	5.3	8.3	15.9
db-2	5	T_E	5.4	7.9	17.0
db-2	5	H_E	5.1	7.2	19.2
db-2	7	E	4.3	11.4	**13.3**
db-2	7	T_E	**3.7**	10.1	14.0
db-2	7	H_E	**3.7**	10.8	13.8
coif-1	5	E	5.3	7.8	16.5
coif-1	5	T_E	5.6	8.0	17.0
coif-1	5	H_E	5.3	**7.0**	18.5
coif-1	7	E	4.3	11.4	14.5
coif-1	7	T_E	**3.7**	10.1	14.6
coif-1	7	H_E	**3.7**	10.9	14.8

TAB. 5.4 – *Résultats en discrimination musique/non-musique en utilisant différentes énergies, les ondelettes* db-2 *et* coif-1 *ainsi que 5 et 7 niveaux de décomposition en ondelettes. Les résultats sont donnés en taux d'erreurs en trames (%).*

Résultats : Nous constatons que les paramètres basés sur les ondelettes sont bien meilleurs que les MFCC en discrimination musique/non-musique. L'amélioration se retrouve pour tous les corpus.
Par rapport à la paramétrisation MFCC nous avons un gain relatif de 43% sur le corpus "Scheirer", 46% sur le corpus "News" et 42% sur "Entertainement", en prenant à chaque fois la meilleure paramétrisation en ondelettes.
Nous observons que l'énergie hiérarchique et l'énergie Teager donnent de bons résultats. Par contre, d'après ces résultats, les deux types d'ondelette, Daubechies et Coiflet, donnent des résultats similaires. La différence entre les résultats n'est pas significative.

Une dernière observation sur ces résultats nous permet de dire qu'une décomposition en 7 bandes est plus adéquate pour les corpus "Scheirer" et "Entertainment" alors qu'une décomposition en 5 bandes semble plus adéquate pour le corpus "News". Ceci peut s'expliquer par le fait que le corpus News est composé en majorité de parole (86%) et contient très peu de musique (14%), comparé aux deux autres corpus, où la musique (musique et parole sur musique) représente 68% du corpus "Scheirer" et 48% du corpus "Entertainment".

Conclusions : Ces résultats confirment notre hypothèse que les paramètres basés sur les coefficients d'ondelettes sont plus adaptés que les MFCC pour traiter des signaux non-stationnaires tels que la musique. De plus, l'énergie Teager, comme pour la discrimination parole/non-parole, donne de bons résultats. L'énergie hiérarchique obtient ici, pour la tâche de discrimination musique/non-musique, de très bons résultats.

5.4.3 Paramètres dynamiques à court terme : Δ et $\Delta\Delta$

D'après les résultats obtenus dans la section précédente, nous nous limiterons à l'étude de la Coiflet 'coif-1' avec une décomposition en 5 bandes pour la tâche de discrimination parole/non-parole et 'coif-1' avec une décomposition en 7 bandes pour la discrimination musique/non-musique.
Nous étudions dans cette section l'influence de la dynamique à court terme de nos paramètres sur les performances en discrimination parole/non-parole et musique/non-musique. En effet, plusieurs études [Scheirer 97, Umapathy 05] ont montré que la dynamique des paramètres permet de prendre en compte les spécificités de la structure de la parole et de la musique.
La dynamique à court terme des paramètres se caractérise ici par l'ajout des dérivées premières (Δ) et secondes ($\Delta\Delta$) des paramètres. Les dérivées premières et secondes sont calculées avec HTK [Young 95b]. Les coefficients delta (Δ) sont obtenus à l'aide de la formule de régression suivante :

$$\Delta_t = \frac{\sum_{\theta=1}^{\Theta} \theta(c_{t+\theta} - c_{t-\theta})}{2\sum_{\theta=1}^{\Theta} \theta^2}$$

où Δ_t est le coefficient delta (la dérivée première) au temps t calculé sur les coefficients statiques : $c_{t-\theta}$ à $c_{t+\theta}$. Θ correspond à la taille de la fenêtre utilisée pour calculer nos dérivées. Pour le calcul de nos dérivées, nous utilisons $\Theta = 2$.
La même formule est appliquée aux coefficients Δ pour obtenir les coefficients d'accélération ($\Delta\Delta$). Enfin, lorsque l'on se trouve en début ou en fin de signal, HTK va par défaut, répliquer le premier ou le dernier vecteur autant de fois que nécessaire pour remplir la fenêtre de régression.

5.4.3.1 Discrimination Parole/Non-parole

Nous utilisons pour ce test en discrimination parole/non-parole, l'ondelette 'coif-1' et 5 niveaux de décomposition. Nous testons dans un premier temps l'ajout des dérivées premières (Δ), puis dans un second temps, l'ajout des dérivées premières et secondes (Δ et $\Delta\Delta$).

Paramètres	**Nb Param.**	**Scheirer**	**News**	**Entertainment**
MFCC+Δ+$\Delta\Delta$	36	*2.5*	*2.9*	*5.8*
E	5	3.3	3.7	4.2
ΔE	5	**1.7**	3.5	3.4
E+Δ	10	3.0	2.7	3.0
E+Δ+$\Delta\Delta$	15	**1.7**	**2.6**	3.2
T_E	5	3.3	3.2	4.2
ΔT_E	5	**1.7**	3.8	3.3
T_E+Δ	10	**1.7**	2.7	2.9
T_E+Δ+$\Delta\Delta$	15	**1.7**	2.7	**2.8**
H_E	5	3.3	4.4	4.3
ΔH_E	5	**1.7**	3.2	3.4
H_E+Δ	10	**1.7**	2.8	3.2
H_E+Δ+$\Delta\Delta$	15	**1.7**	2.9	3.3

TAB. 5.5 – *Résultats en discrimination parole/non-parole avec l'ajout de paramètres dynamiques (Δ, $\Delta\Delta$). L'ondelette* coif-1 *avec une décomposition en 5 bandes est utilisée. Taux d'erreurs en trames (%).*

Résultats : Le tableau 5.5 regroupe les résultats. Nous avons laissé en première ligne la paramétrisation MFCC de référence et les différentes paramétrisations statiques pour chaque énergie afin de comparer l'influence des paramètres dynamiques. Une première constatation est que les coefficients dynamiques seuls sont déjà plus performants que les paramètres statiques sur les différents corpus et toutes énergies confondues. L'ajout des premières dérivées (Δ) aux paramètres statiques améliorent encore les performances. Nous avons par exemple, pour l'énergie Teager, un gain relatif de 32% sur le corpus "Scheirer", 7% sur le corpus "News" et 50% sur le corpus "Entertainement", par rapport aux paramètres statiques. En revanche, l'ajout des dérivées secondes n'apporte aucune amélioration significative. Finalement, le résultat que nous devons retenir ici, est que l'ajout des paramètres dynamiques a permis de dépasser les résultats obtenus avec les MFCC en discrimination parole/non-parole, et ceci sur tous les corpus. Avec nos meilleurs paramètres, nous avons un gain relatif significatif de 32% sur le corpus "Scheirer", 10% sur le corpus "News" et 52% sur le corpus "Entertainement", par rapport aux paramètres MFCC. De plus, nos paramètres sont plus compacts que les MFCC. En effet, nous obtenons de meilleurs résultats avec des vecteurs à 10 composantes alors que les MFCC en ont 36.

5.4.3.2 Discrimination Musique/Non-musique

Nous utilisons pour ce test l'ondelette 'coif-1' et 7 niveaux de décomposition. Comme précédemment, nous regardons l'influence des paramètres dynamiques sur les résultats en discrimination musique/non-musique.

Paramètres	Nb Param.	Scheirer	News	Entertainment
MFCC+Δ+ΔΔ	36	*6.5*	*13.1*	*23.1*
E	7	4.3	11.4	**14.5**
ΔE	7	**1.8**	8.1	18.1
E+Δ	14	**1.8**	7.9	15.2
E+Δ+ΔΔ	21	**1.8**	9.5	17.4
T_E	7	3.7	10.1	14.6
ΔT_E	7	3.4	**6.3**	18.2
T_E+Δ	14	**1.8**	7.2	15.0
T_E+Δ+ΔΔ	21	**1.8**	9.7	17.4
H_E	7	3.7	10.9	14.8
ΔH_E	7	3.4	8.8	20.4
H_E+Δ	14	**1.8**	7.2	14.8
H_E+Δ+ΔΔ	21	**1.8**	8.6	18.3

TAB. 5.6 – *Résultats en discrimination musique/non-musique avec l'ajout de paramètres dynamiques (Δ, ΔΔ). L'ondelette* coif-1 *avec une décomposition en 7 bandes est ici utilisée. Taux d'erreurs en trames (%).*

Résultats : Le tableau 5.6 regroupe les différents résultats obtenus. Nous observons une nouvelle fois que l'ajout des dérivées premières (Δ) améliore les performances en discrimination par rapport aux paramètres statiques sur les corpus "Scheirer" et "News". Nous constatons aussi que les coefficients dynamiques seuls donnent de meilleurs résulats que les paramètres statiques sur ces deux corpus. Nous obtenons par exemple, en prenant l'ajout des Δ aux paramètres basés sur l'énergie hiérarchique (H_E), un gain relatif significatif de 51% sur le corpus "Scheirer" et 34% sur le corpus "News", par rapport aux paramètres statiques. Par rapport aux paramètres MFCC, le gain relatif est encore plus important : 72% sur le corpus "Scheirer" et 45% sur le corpus "News". En revanche, il n'y a aucune amélioration et même une légère détérioration des performances sur le corpus "Entertainment" lorsque l'on utilise les coefficients dynamiques et lorsque l'on ajoute les coefficients dynamiques aux paramètres statiques. Une raison pour expliquer cela pourrait venir du fait que le corpus "Entertainment" contient plus de musique et de parole sur de la musique. Les coefficients dynamiques seraient alors moins performants en présence de parole sur un fond musical. Enfin, nous remarquons que l'ajout des dérivées secondes (ΔΔ) n'apporte pas de gain de performance par rapport à l'ajout des dérivées premières. Nous constatons

même une détérioration des performances sur les corpus “News” et “Entertainment”. Nous attribuons cette baisse de performance aux propriétés des dérivées secondes. En effet, elles permettent de capturer l'accélération de nos paramètres, c'est-à-dire la vitesse de variation à court terme de nos paramètres. Cette information ne semble pas très utile, voire même source de confusion en discrimination parole/musique.

5.4.3.3 Conclusions

En conclusion de cette section, nous pouvons dire que la dynamique à court terme des paramètres, avec l'ajout de leurs dérivées premières (Δ) apporte un réel gain de performance et nous permet de dépasser les résultats obtenus avec la paramétrisation MFCC de référence, que ce soit en discrimination parole/non-parole ou musique/non-musique.
En revanche en ce qui concerne l'ajout des dérivées secondes, les résultats sont sans appels. Cela ne sert à rien si ce n'est à détériorer nos résultats. Enfin, l'énergie Teager semble une nouvelle fois sortir du lot, peut-être que cela est dû à ses capacités en débruitage. En effet, l'énergie Teager amplifie les variations et nous savons que le cerveau est sensible aux variations. C'est donc les variations du signal qui semble être l'élément le plus discriminant pour séparer la parole de la musique.

5.4.4 Paramètres dynamiques à long terme : Variance sur une seconde

Nous voulons ici vérifier l'hypothèse de Sheirer et Slaney [Scheirer 97] : l'utilisation de la variance des paramètres au lieu des paramètres eux-mêmes donne de meilleurs résultats en discrimination parole/musique. Aussi, d'après [Umapathy 05, Williams 99], l'étude de

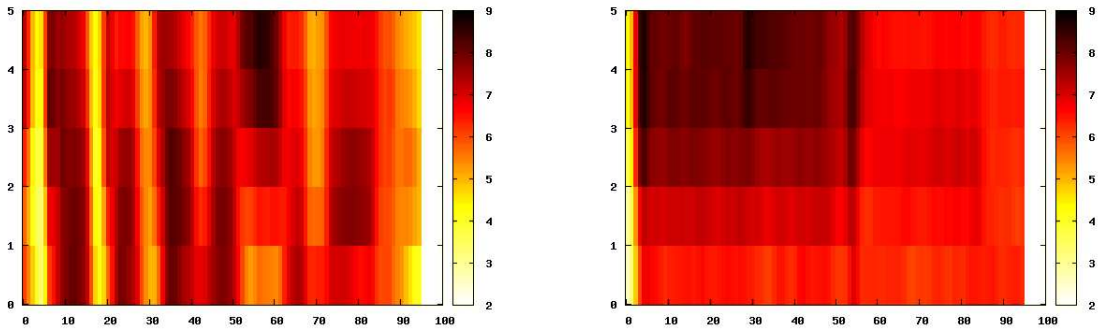

FIG. 5.3 – Spectrogrammes représentant l'énergie instantanée des coefficients d'ondelettes sur deux signaux d'une durée d'une seconde : un signal de parole (à droite) et un signal de musique (à gauche). Il y a ici 5 niveaux de décomposition en ondelettes.

la variance sur une grande fenêtre d'analyse, c'est-à-dire d'une durée allant de 1 à 2.5 secondes, semble intéressante. Enfin, la figure 5.3 nous permet de nous convaincre que la variance des paramètres en ondelettes doit donner de bons résultats en discrimination parole/musique. En effet, nous observons une forte variation des paramètres pour le signal de parole (image de gauche), alors que les paramètres varient très peu pour le signal de musique (image de droite).

Après une étude préliminaire pour déterminer la taille de notre fenêtre, nous avons choisi d'utiliser une fenêtre d'une durée d'une seconde pour calculer la variance des paramètres. Nous calculons la variance sur nos paramètres statiques, et pour pouvoir comparer les résultats, nous calculons aussi la variance sur les paramètres de référence MFCC.

5.4.4.1 Discrimination Parole/Non-parole

Comme pour les paramètres dynamiques à court terme, nous utilisons l'ondelette 'coif-1' et 5 niveaux de décomposition pour la tâche de discrimination parole/non-parole. Nous analysons ensuite l'influence de l'utilisation de la variance des paramètres à la place des paramètres eux-mêmes, comme le préconise Scheirer [Scheirer 97]. Nous étudions aussi l'influence de l'ajout aux paramètres statiques de leur variance calculée sur une seconde. Tous les résultats sont résumés dans la table 5.7.

Paramètres	**Nb Param.**	**Scheirer**	**News**	**Entertainment**
MFCC+Δ+ΔΔ	36	*2.5*	*2.9*	*5.8*
Variance des MFCC	12	*2.2*	*4.1*	*8.1*
MFCC + Variance des MFCC	24	*3.4*	*4.3*	*10.4*
E+Δ	10	3.0	2.7	3.0
T_E+Δ	10	**1.7**	**2.7**	**2.9**
H_E+Δ	10	**1.7**	2.8	3.2
Variance de E	5	**1.7**	3.9	3.7
Variance de T_E	5	**1.7**	4.0	3.7
Variance de H_E	5	**1.7**	4.2	4.1
E+(Variance de E)	10	2.1	4.2	4.2
T_E+(Variance de T_E)	10	**1.7**	4.1	4.1
H_E+ (Variance de H_E)	10	2.1	4.5	5.1

TAB. 5.7 – *Résultats en discrimination parole/non-parole en utilisant : la variance des paramètres statiques calculée sur une fenêtre d'une seconde ou cette variance ajoutée aux paramètres statiques. L'ondelette* coif-1 *et 5 bandes de décomposition sont ici utilisées. Les résultats sont donnés en taux d'erreurs en trames (%).*

Résultats : Nous observons que la variance de nos paramètres, quelle soit combinée à nos paramètres statiques ou seule, donne de bons résultats. L'utilisation de la variance seule donne les meilleures performances. On constate toutefois que si les résultats sont similaires

à ceux obtenus lors de l'ajout des Δ pour le corpus "Scheirer", il n'en est pas de même pour les deux autres corpus. Par rapport à la paramétrisation MFCC, nous obtenons un gain plus faible que lors de l'ajout des Δ, pour le corpus "Entertainment" mais pas pour le corpus "News". Encore une fois, ce corpus étant constitué principalement de parole et très peu de musique, il est très difficile de faire mieux que les MFCC. Nous constatons aussi que la variance sur une seconde, calculée sur la paramétrisation MFCC de référence et utilisée pour la tâche de discrimination parole/non-parole, améliore les performances de la paramétrisation de référence sur tous les corpus excepté "News". Nous revenons ainsi à la même conclusion que précédemment : il est très difficile de faire mieux que les MFCC lorsque le corpus est essentiellement composé de parole.

5.4.4.2 Discrimination Musique/Non-musique

Nous utilisons pour cette expérience l'ondelette 'coif-1' et 7 niveaux de décomposition. Nous analysons, comme précédemment, l'influence de l'utilisation de la variance des paramètres à la place des paramètres eux-mêmes, la variance étant toujours calculée sur une fenêtre d'une seconde. De même, nous étudions l'influence de l'adjonction de cette variance aux paramètres statiques. Les résultats des tests se trouvent dans le tableau 5.8.

Paramètres	**Nb Param.**	**Scheirer**	**News**	**Entertainment**
MFCC+Δ+$\Delta\Delta$	36	*6.5*	*13.1*	*23.1*
Variance des MFCC	12	*3.1*	*7.7*	*25.1*
MFCC + Variance des MFCC	24	*4.7*	*9.4*	*22.5*
E+Δ	14	1.8	7.9	15.2
T_E+Δ	14	1.8	7.2	15.0
H_E+Δ	14	1.8	7.2	**14.8**
Variance de E	7	**1.7**	7.5	16.3
Variance de T_E	7	1.8	**7.1**	16.4
Variance de H_E	7	1.8	7.3	16.7
E + (Variance de E)	14	1.8	8.3	18.4
T_E + (Variance de T_E)	14	1.8	9.2	19.2
H_E + (Variance de H_E)	14	1.8	8.6	19.1

TAB. 5.8 – *Résultats en discrimination musique/non-musique en utilisant : la variance des paramètres statiques calculée sur une fenêtre d'une seconde ou cette variance ajoutée aux paramètres statiques. L'ondelette* coif-1 *et 7 bandes de décomposition sont ici utilisées. Les résultats sont donnés en taux d'erreurs en trames (%).*

Résultats : En regardant le tableau 5.8, nous pouvons dire que la variance des paramètres calculée sur une fenêtre d'une seconde donne de très bons résultats comparée à la paramétrisation MFCC de référence. Nous avons un gain relatif significatif de 74% pour "Scheirer", 46% pour "News" et 29% pour "Entertainement", comparé à la paramétrisa-

tion MFCC de référence. Nous observons aussi que la variance sur une seconde, calculée sur la paramétrisation MFCC de référence et utilisée pour la tâche de discrimination musique/non-musique, améliore les performances de la paramétrisation de référence sur tous les corpus excepté "Entertainment". Même si nous nous comparons à la variance des paramètres MFCC, nous obtenons un gain relatif significatif de 50% pour "Scheirer", 22% pour "News" et 30% pour "Entertainement". Enfin, en comparant les tableaux 5.7 et 5.8, nous constatons que l'utilisation de la variance des paramètres calculée sur une seconde donne des résultats similaires à l'adjonction de la dérivée première (Δ) aux paramètres statiques. Mais un avantage revient à la variance qui permet une représentation encore plus compacte : un vecteur de 7 composantes pour la variance au lieu d'un vecteur de 14 composantes pour l'utilisation des Δ.

5.4.4.3 Conclusions

A la vue des différents résultats en discrimination parole/non-parole et musique/non-musique, nous pouvons dire que l'hypothèse de Scheirer et Slaney [Scheirer 97] s'avère vraie. L'utilisation de la variance des paramètres donne de meilleurs résultats que l'utilisation des paramètres eux-mêmes. La figure 5.4 nous permet de nous en convaincre.

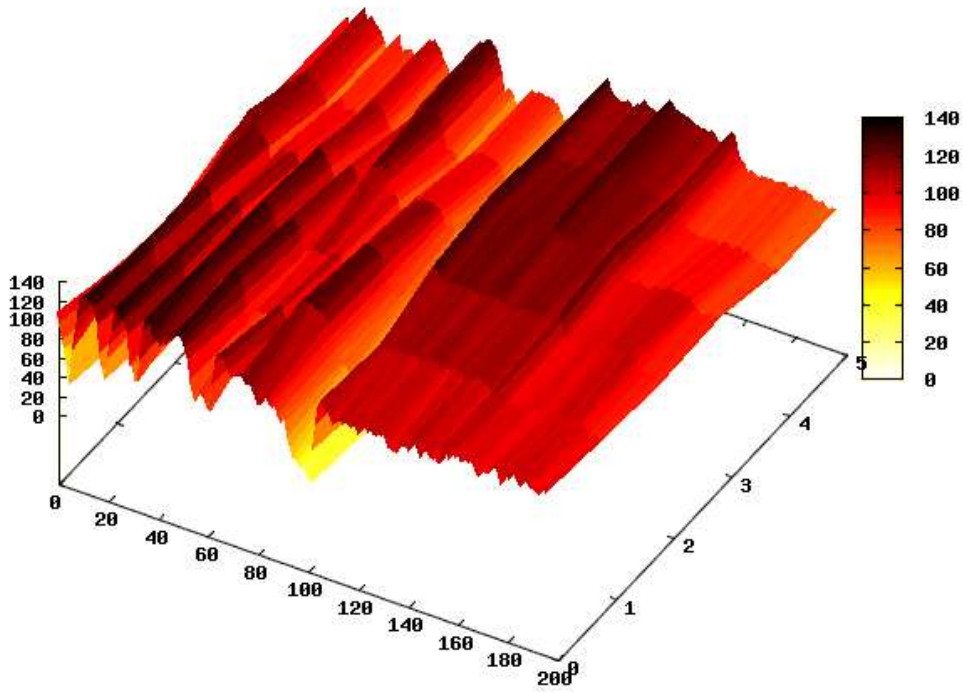

FIG. 5.4 – Spectrogramme représentant l'énergie instantanée des coefficients d'ondelettes sur un signal d'une durée de 2 secondes : 1s de parole suivie par 1s de musique. Nous avons ici 5 niveaux de décomposition.

La combinaison de la variance et des paramètres statiques n'apportent quant à elle aucune amélioration, comparée à l'utilisation de la variance seule. Nous observons même une dégradation des performances pour le corpus "Entertainement". Enfin, les paramètres basés sur l'ajout des dérivées premières et ceux basés sur l'utilisation de la variance donnent des performances similaires en discrimination parole/non-parole et musique/non-musique. La variance sur une seconde a un avantage supplémentaire par rapport à l'ajout des dérivées premières, c'est qu'elle permet une paramétrisation plus compacte du signal. En effet, nous obtenons des résultats similaires avec moitié moins de paramètres en utilisant la variance sur une seconde pour les différentes tâches de discrimination. Nous utilisons un vecteur a 5 composantes en utilisant la variance sur une seconde au lieu de 10 avec l'ajout des Δ en discrimination parole/non-parole et un vecteur de 7 composantes au lieu de 14 en discrimination musique/non-musique.

5.4.5 Discrimination globale : Parole/Musique

La discrimination globale correspond à la séparation de la parole, de la musique et de la parole sur un fond musical. Elle consiste ici à effectuer séparément les étapes de discrimination parole/non-parole et musique/non-musique puis à regrouper les résultats obtenus par les deux sous-systèmes de classification, comme nous l'avons expliqué dans la section 5.1.3 (tableau 5.1).

5.4.5.1 Regroupement des sorties des meilleurs classifieurs P/NP et M/NM

Dans cette expérience, nous effectuons une fusion simple entre le meilleur classifieur parole/non-parole et le meilleur classifieur musique/non-musique. Cette fusion s'effectue selon la table 5.9.

Sortie du classifieur P/NP	Sortie du classifieur M/NM	étiquetage final du segment
Parole	Non-Musique	**Parole**
Parole	Musique	**Parole sur musique**
Non-Parole	Non-Musique	**Bruit/Silence**
Non-Parole	Musique	**Musique**

TAB. 5.9 – *Fusion des résultats des meilleurs sous-systèmes de classification P/NP et M/NM pour l'étiquetage final des trames du signal audio. La classification s'effectue trame par trame.*

Dans les expériences précédentes, nous avons conclu que la meilleure paramétrisation parole/non-parole est basée sur l'ondelette 'coif-1' et 5 bandes de décomposition et que celle musique/non-musique est basée sur l'ondelette 'coif-1' mais avec une décomposition en 7 bandes. En ce qui concerne l'énergie, nous utilisons l'énergie de Teager. En effet, cette énergie a donné de très bons résultats dans toutes nos expériences précédentes.

Nous avons donc choisi de garder les paramètres dynamiques à court terme basés sur l'ajout des Δ et la paramétrisation à long terme basée sur l'utilisation de la variance des paramètres. Ces deux paramétrisations ont donné d'excellents résultats que ce soit en discrimination parole/non-parole ou en discrimination musique/non-musique. Finalement, nous calculons, pour évaluer les performances, le taux d'erreurs en trames pour la tâche de discrimination globale. Ce score est présenté dans la section 5.3 (voir équation 5.1). Les différents résultats sont présentés dans le tableau 5.10.

Param.M/NM	**Nb**	**Param.P/NP**	**Nb**	**Scheir.**	**News**	**Enter.**
MFCC+Δ+$\Delta\Delta$	36	*MFCC+Δ+$\Delta\Delta$*	36	*8.1*	*15.0*	*26.3*
T_E(7bandes)+Δ	10	T_E(5bandes)+Δ	14	**3.4**	**9.0**	**18.4**
Var de T_E(7bandes)	5	Var de T_E(5bandes)	7	**3.4**	10.7	19.6

TAB. 5.10 – *Discrimination globale, sur les différents corpus de test, avec les meilleurs paramètres : l'ondelette* coif-1 *avec 7 bandes et les Δ pour la discrimination musique/non-musique et l'ondelette* coif-1 *avec 5 bandes et les Δ pour la discrimination parole/non-parole. Taux d'erreurs en trames (%).*

Résultats : Nos deux paramétrisations améliorent les performances obtenues avec la paramétrisation MFCC de référence. Les meilleurs résultats sont toutefois obtenus avec les paramètres à court terme, c'est-à-dire avec l'ajout des dérivées premières. Nous obtenons ainsi avec la meilleure paramétrisation, un gain relatif de 58% sur le corpus "Scheirer", 40% sur "News" et 30% sur "Entertainment", comparé à la paramétrisation de référence.

5.4.5.2 Conclusions

Notre paramétrisation en ondelettes donne les résultats escomptés après les différents tests que nous avons effectués précédemment. Le gain obtenu pour les différents corpus est très significatif. De plus, nous avons un autre avantage pour notre paramétrisation : sa compacité. Nous avons au maximum : un vecteur à 14 composantes contre un vecteur à 36 composantes pour la paramétrisation MFCC.

5.4.6 Test de validation sur le corpus *TestRTL*

Les expériences précédentes nous ont permis de trouver les meilleures paramétrisations en discrimination parole/non-parole et en discrimination musique/non-musique. Ces paramétrisations sont obtenues en utilisant l'ondelette 'coif-1', l'énergie de Teager et une décomposition en 5 bandes pour la discrimination parole/non-parole et en 7 bandes pour la discrimination musique/non-musique. Les meilleurs résultats ont été obtenus à la fois lors de l'adjonction des Δ aux paramètres statiques et lors de l'utilisation de la variance des paramètres calculée sur une seconde. Nous avons décidé d'évaluer ces deux paramétri-

sations sur un corpus de validation : $TestRTL$. Ce dernier test va permettre de confirmer ou d'infirmer nos résultats.

5.4.6.1 Tests de discrimination sur $TestRTL$

Nous avons regroupé dans le tableau 5.11 les résultats des tests en discrimination parole/non-parole (P/NP), musique/non-musique (M/NM) et globale (TG) en utilisant nos deux meilleures paramétrisations ainsi que la paramétrisation MFCC de référence. La deuxième colonne du tableau nous indique le nombre de paramètres utilisés par les classifieurs. Le premier chiffre correspond au nombre de paramètres utilisés par le classifieur P/NP et le second chiffre au nombre de paramètres utilisés par le classifieur M/NM.

Paramètres	**Nb Param.**	**M/NM**	**P/NP**	**TG**
MFCC+Δ+ΔΔ	36-36	*24.9*	**8.1**	*30.9*
T_E + Δ	10-14	30.8	10.4	37.6
Variance de T_E	5-7	**9.0**	10	**15.5**

TAB. 5.11 – *Résultats en discrimination pour le corpus TestRTL en utilisant coif-1 et 5 bandes (pour la détection parole/non-parole) et avec 7 bandes (pour la détection musique/non-musique). Taux d'erreurs en trames (%).*

Résultats : Les résultats de la table 5.11 montrent que nos paramètres basés sur la variance sur une seconde donnent toujours de très bons résultats. Nous obtenons un gain de performance en discrimination musique/non-musique et globale, mais pas en discrimination parole/non-parole. Nous avions déjà observé ce résultat sur les corpus précédemment testés. Cela s'explique par le fait que les MFCC sont connus pour être de très bons paramètres pour représenter la parole. Le gain relatif obtenu est significatif : 64% en discrimination musique/non-musique et 50% en discrimination globale.
En revanche, les paramètres, composés des paramètres statiques basés sur l'énergie Teager et de leur premières dérivées (Δ), donnent de moins bons résultats par rapport à la paramétrisation de référence. Une explication à ce phénomène pourrait venir de la composition du corpus $TestRTL$. En effet, celui-ci est particulièrement difficile. Il est composé de rires, de publicités, d'applaudissements, d'interviews, etc.
Une analyse approfondie des segments mal classés en musique et non-musique nous indique que ce sont bien les bruits (rires et applaudissements) ainsi que certaines musiques qui sont à l'origine des erreurs. En effet, il y a beaucoup d'extraits d'émissions de télévision ou de spectacles qui sont reconnus en musique à cause des réactions du public (rires, applaudissements, huées, etc.). Mais des erreurs sont aussi commises à cause de certaines voix. Nous avons par exemple des extraits du spectacle de "Shirley et Dino", où leurs voix sont vraiment très spéciales. Nous avons aussi découvert que lorsque plusieurs personnes

parlaient en même temps, c'est-à-dire lorsque nous sommes en présence de parole concurrente, le système de classification M/NM a tendance de classer le signal en musique. Ainsi les voix de "Nagui" et de "Omar et Fred" sont reconnues en musique lorsqu'ils parlent en même temps et avec divers bruits (petits rires). Pour ce dernier cas, "Nagui" précise qu'il n'a plus de voix et à l'écoute on remarque bien sa voix cassée. De plus, si nous sommes très attentifs, nous pouvons entendre, par moment, tout au long de l'émission une musique de fond (musique et parfois chanson) qui n'a pas été entendue par la personne étiquetant le corpus. Tout ceci montre que ce corpus de validation est extrêmement difficile à étiqueter, même pour un être humain.

Conclusions : les paramètres à long terme basés sur les ondelettes, c'est-à-dire la variance calculée sur une seconde, donnent de très bons résultats et sont constants quelque soit le corpus. Ce n'est pas le cas pour les paramètres à court termes utilisant les dérivées premières des paramètres. Ils pêchent par leur inconstance entre les différents corpus, bien que ce soient les particularités (rires, applaudissements, publicités, etc.) du corpus $TestRTL$ qui seraient à l'origine de la mauvaise performance des paramètres à court terme. Les bons résultats de la variance sur ce dernier corpus confirment une fois de plus l'hypothèse de Scheirer [Scheirer 97]. La variance calculée sur le long terme est donc un paramètre robuste pour discriminer la parole de la musique. C'est pareil pour l'homme : nous avons besoin d'un certain temps pour faire la différence entre parole, musique et parole sur musique. Enfin, le test de validation montre la limite des expérimentations pour trouver la meilleure paramétrisation. En effet, il existe une multitude de styles musicaux et une grande diversité de bruits. Nous ne pouvions pas tous les retrouver dans nos corpus d'apprentissage et lors de nos tests de développement. C'est là un des problèmes de l'apprentissage statistique. Nous nommerons dans la suite "1PNP-1MNM" la fusion des sorties des classifieurs parole/non-parole et musique/non-musique basés sur la variance de l'énergie Teager calculée sur une seconde.

5.4.6.2 Analyse approfondie des résultats

Pour savoir à quel niveau nos paramètres basés sur les ondelettes améliorent les résultats sur le corpus $TestRTL$ pour la tâche de discrimination globale, nous donnons la répartition des trames reconnues en parole, musique et parole sur musique. Le tableau 5.12 nous donne cette répartition avec la paramétrisation MFCC de référence. Le tableau 5.13 donne la répartition des trames reconnues avec la meilleure paramétrisation en ondelettes sur ce corpus, c'est-à-dire la paramétrisation utilisant la variance calculée sur une seconde des paramètres basés sur l'ondelette 'coif-1' et l'énergie Teager.

Ces tableaux nous permettent de faire ressortir la difficulté de discriminer la parole de la musique, spécialement lorsque nous sommes en présence de segments où la parole et la musique sont superposées. Nous pouvons voir que, comparé aux paramètres MFCC (tableau 5.12), nous avons une réduction significative du nombre de segments mal classés

étiquetées \ reconnues	**M**	**PM**	**P**
M	98.1	1.9	0.0
PM	61.3	34.6	4.2
P	3.1	33.8	63.1

TAB. 5.12 – *Répartition des trames (%) pour la tâche de discrimination globale en utilisant la paramétrisation de référence : 12 coefficients MFCC avec leur premières et secondes dérivées. (Corpus TestRTL).*

étiquetées \ reconnues	**M**	**PM**	**P**
M	95.1	2.6	1.8
PM	63.3	22.9	13.7
P	4.6	6.2	89.1

TAB. 5.13 – *Répartition des trames (%) pour la tâche de discrimination globale en utilisant la meilleure paramétrisation en ondelette, i.e. l'ondelette coif-1, la variance de l'énergie de Teager avec 5 bandes de décomposition pour la discrimination parole/non-parole et la variance de l'énergie de Teager avec 7 bandes de décomposition pour la discrimination musique/non-musique. (Corpus TestRTL).*

pour la parole. Dans les émissions en public, nous avons remarqué que beaucoup de segments de parole contiennent du bruit de fond et que ces derniers sont souvent mal classés lorsque nous utilisons la paramétrisation MFCC : ils sont reconnus comme de la parole sur musique. Notre paramétrisation en ondelettes réussit, quant à elle, à bien identifier ces segments. Cela peut s'expliquer par l'utilisation conjointe des ondelettes et de l'énergie de Teager. En effet, les ondelettes et l'énergie de Teager sont souvent utilisées lorsque l'on est en présence de bruit.

5.5 Combinaison de paramètres et fusion de classifieurs

Nous présentons dans cette section des résultats de tests complémentaires sur nos différents corpus. Nous avons voulu tester deux stratégies pour combiner nos paramètres : d'une part la combinaison de la paramétrisation MFCC avec nos meilleurs paramètres en ondelettes sous la forme d'un grand vecteur et d'autre part, la combinaison de plusieurs classifieurs parole/non-parole et musique/non-musique basés sur différentes paramétrisations.

5.5.1 Combinaison des MFCC et des paramètres en ondelettes

Les MFCC donnant de bons résultats en discrimination, nous avons voulu les combiner à nos paramètres en ondelettes. Pour cela, nous avons créé un grand vecteur de paramètres en concaténant les paramètres MFCC à nos paramètres en ondelettes. Les résultats de ce test se trouvent dans la table 5.14.

Param.M/NM	**Param.P/NP**	**Nb Par.**	**Scheir.**	**News**	**Enter.**
MFCC+Δ+ΔΔ	*MFCC+Δ+ΔΔ*	36-36	*8.1*	*15.0*	*26.3*
T_E(7bandes)+Δ	T_E(5bandes)+Δ	10-14	**3.4**	**9.0**	**18.4**
(MFCC+Δ+ΔΔ)+ T_E(7bandes)+Δ	(MFCC+Δ+ΔΔ)+ T_E(5bandes)+Δ	46-50	5.0	9.6	22.5

TAB. 5.14 – *Discrimination globale en utilisant les paramètres MFCC couplés aux meilleurs paramètres en ondelettes sur les corpus "Scheirer", "News" et "Entertainment" : Δ avec l'ondelette* coif-1 *et une décomposition en 7 bandes pour la discrimination musique/non-musique ; Δ avec l'ondelette* coif-1 *et une décomposition en 5 bandes pour la discrimination parole/non-parole. Taux d'erreurs en trames (%). La deuxième ligne du tableau correspond à la meilleure paramétrisation basée sur les ondelettes pour la discrimination parole/musique.*

Résultats : A la vue du tableau 5.14, nous pouvons conclure que la combinaison des paramètres avec nos paramètres en ondelettes dans un même vecteur n'apporte rien. Nous dégradons même les performances de nos paramètres. Nous pouvons expliquer ce phénomène par ce que l'on appelle communément la "malédiction de la dimension" (*"curse of dimensionality"* en anglais) [Bellman 61]. En effet, plus on augmente la dimension de l'espace par l'intermédiaire de la taille du vecteur de paramètres, plus nos vecteurs d'observation vont correspondre à des points isolés dans un espace immense. Il est alors très difficile de faire de l'analyse statistique et donc d'apprendre quelque chose à partir de nos vecteurs de paramètres.

Conclusions : La conclusion qui s'impose ici est qu'il n'est pas judicieux de construire des vecteurs de paramètres de très grande taille. Le problème de la "malédiction de la dimension" nous le rappelle. L'approche consistant à utiliser plusieurs classifieurs basés sur différentes paramétrisations et à fusionner leurs sorties semble être une meilleure idée. Nous allons le voir dans la section suivante.

5.5.2 Fusion de classifieurs par vote majoritaire

Dans le but d'améliorer les performances des différentes tâches de discrimination, nous combinons les sorties de plusieurs classifieurs. Les classifieurs diffèrent par leur paramétrisation et nous les définirons donc par l'intermédiaire de leurs paramètres. Cette étude

est une première approche de la fusion des sorties de classifieurs, c'est pourquoi nous nous limitons à la fusion par vote majoritaire. Nous prendrons comme référence la fusion des sorties des meilleurs classifieurs parole/non-parole et musique/non-musique ("1PNP-1MNM"). Elle correspond à la discrimination globale que nous avons étudiée à la section 5.4.5.1. Nous avons appelé cette première fusion : **"fusion 1PNP-1MNM"**. Nous avons vu dans la section 5.4.5 qu'elle donnait déjà de bons résultats en discrimination globale. Les paramètres définissant les classifieurs pour la fusion 1PNP-1MNM sont donnés ci-dessous.

Fusion 1PNP-1MNM :

* Paramètres utilisés pour la tâche de discrimination parole/non-parole :
  - la variance sur 1 seconde calculée sur l'ondelette 'coif-1' associée à l'énergie Teager et avec une décomposition en 5 bandes.

* Paramètres utilisés pour la tâche de discrimination musique/non-musique :
  - la variance sur 1 seconde calculée sur l'ondelette 'coif-1' associée à l'énergie Teager et avec une décomposition en 7 bandes.

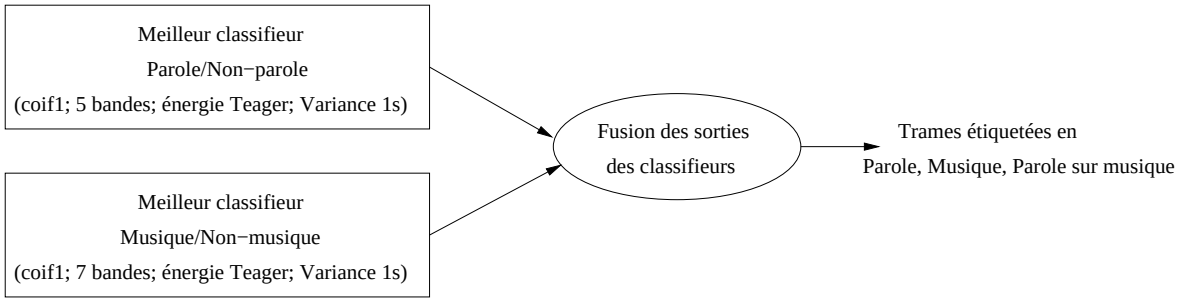

FIG. 5.5 – Fusion de classifieurs consistant à regrouper la sortie du meilleur classifieur parole/non-parole et la sortie du meilleur classifieur musique/non-musique. (Fusion 1PNP-1MNM).

Pour le second type de fusion, que nous appellerons **"fusion 3PNP-3MNM"**, nous utilisons 3 classifieurs basés sur des paramètres différents pour chacune des tâches de discrimination (parole/non-parole et musique/non-musique). Les sorties de ces classifieurs sont alors fusionnées en utilisant la stratégie du vote majoritaire.
Nous supposons que les paramétrisations choisies donnent de bons résultats, apportent une certaine diversité et produisent différents types d'erreurs. La combinaison de tels experts devrait ainsi réduire globalement l'erreur en classification.

Fusion 3PNP-3MNM :

Pour chacune des tâches de discrimination, les trois paramétrisations sont choisies de la manière suivante : nous sélectionnons les meilleurs paramètres statiques, les meilleurs paramètres dynamiques à court terme (paramètres statiques plus leur dérivées), et les meilleurs paramètres à long terme (variance sur une seconde). D'après les expériences effectuées précédemment, nous obtenons :

* Paramètres utilisés pour la tâche de discrimination parole/non-parole :
  - l'ondelette 'coif-1' associée à l'énergie Teager et à une décomposition en 5 bandes ,
  - l'ondelette 'coif-1' associée à l'énergie Teager et à une décomposition en 5 bandes avec leurs premières dérivées (Δ),
  - la variance sur 1 seconde calculée sur l'ondelette 'coif-1' associée à l'énergie Teager et à une décomposition en 5 bandes.

* Paramètres utilisés pour la tâche de discrimination musique/non-musique :
  - l'ondelette 'coif-1' associée à l'énergie Teager et à une décomposition en 7 bandes ,
  - l'ondelette 'coif-1' associée à l'énergie hiérarchique et à une décomposition en 7 bandes avec leurs premières dérivées (Δ),
  - la variance sur 1 seconde calculée sur l'ondelette 'coif-1' associée à l'énergie Teager et à une décomposition en 7 bandes.

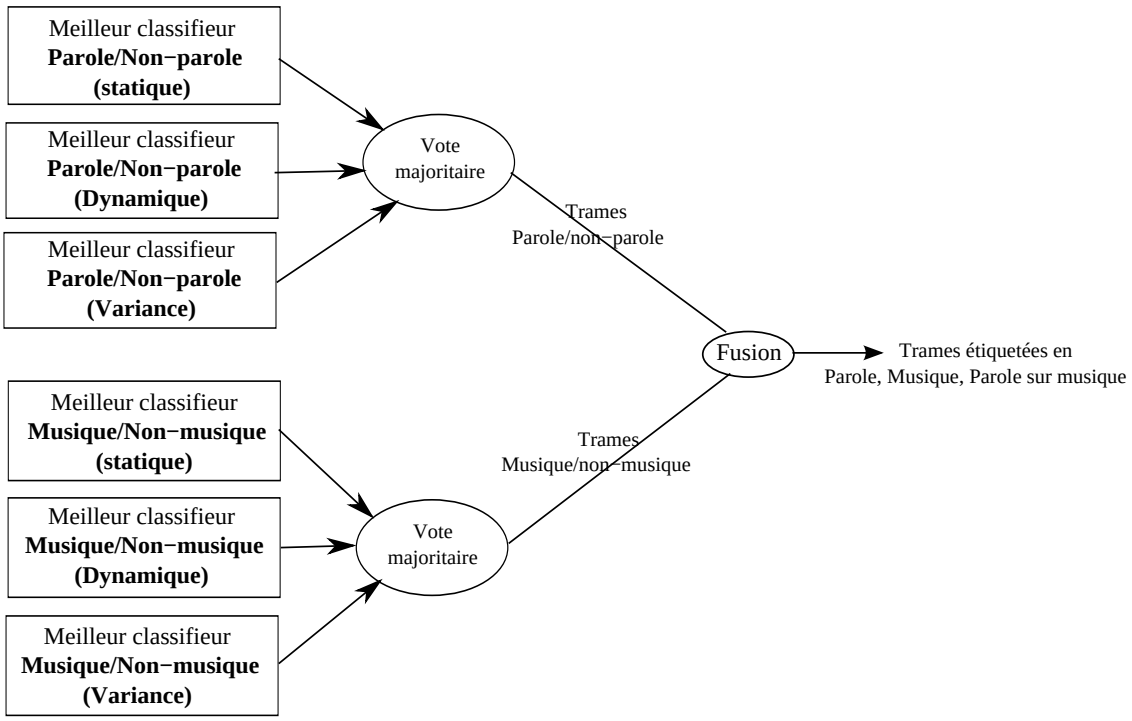

FIG. 5.6 – Fusion de classifieurs par vote majoritaire en utilisant 3 classifieurs parole/non-parole et 3 classifieurs musique/non-musique. (Fusion 3PNP-3MNM).

Les résultats des tests sur la fusion de classifieurs pour les différents corpus *Scheirer*, *News*, *Entertainment* et *TestRTL* sont donnés dans les tables 5.15, 5.16, 5.17 et 5.18 respectivement.

Param.	**M/NM**	**P/NP**	**TG**
Fusion 1PNP-1MNM	*1.8*	*1.7*	*3.4*
Fusion 3PNP-3MNM	1.8	1.7	**3.3**

TAB. 5.15 – *Taux d'erreurs (%) pour les 3 tâches de discrimination sur le corpus* **Scheirer**, *en utilisant la fusion de classifieurs.*

Résultats Scheirer : Le tableau 5.15 regroupe les résultats des différentes fusions pour le corpus Scheirer. Nous constatons que la combinaison des meilleurs classifieurs Parole/non-parole et Musique/Non-musique (*Fusion 1PNP-1MNM*) et la fusion par vote majoritaire de trois classifieurs Parole/non-parole et de trois classifieurs Musique/non-musique (*Fusion 3PNP-3MNM*)donnent des résultats similaires. La légère amélioration en discrimination globale pour la *Fusion 3PNP-3MNM* n'est pas significative.

Param.	**M/NM**	**P/NP**	**TG**
Fusion 1PNP-1MNM	*7.1*	*4.0*	*10.7*
Fusion 3PNP-3MNM	**6.7**	**2.8**	**8.6**

TAB. 5.16 – *Taux d'erreurs (%) pour les 3 tâches de discrimination sur le corpus* **News**, *en utilisant la fusion de classifieurs.*

Résultats News : Pour le corpus *News* nous remarquons, dans la table 5.16, que pour toutes les tâches de discrimination, nous obtenons avec la *Fusion 3PNP-3MNM* une amélioration significative des performances par rapport à la *Fusion 1PNP-1MNM*. Nous avons ainsi un gain relatif de 5.6% en Musique/Non-musique, 30% en Parole/non-parole et 19.6% et discrimination globale.

Param.	**M/NM**	**P/NP**	**TG**
Fusion 1PNP-1MNM	*16.4*	*3.7*	*19.6*
Fusion 3PNP-3MNM	**14**	**2.3**	**16.1**

TAB. 5.17 – *Taux d'erreurs (%) pour les 3 tâches de discrimination sur le corpus* **Entertainment**, *en utilisant la fusion de classifieurs.*

Résultats Entertainment : La table 5.17 nous montre les résultats pour les trois tâches de discrimination en utilisant les deux approches de fusion pour le corpus *Entertainment*. Nous pouvons noter une amélioration significative pour chacune des tâches de discrimination. Même si l'amélioration en passant de la *Fusion 1PNP-1MNM* à la *Fusion 3PNP-3MNM* n'est que de 14.6% dans le cas de la discrimination Musique/non-musique, elle

est de 37.8% dans le cas de la discrimination parole/non-parole. De plus, le passage de la *Fusion 1PNP-1MNM* à la *Fusion 3PNP-3MNM* apporte une diminution significative du taux d'erreurs en classification globale (TG) : un gain relatif de 17.6%.

Param.	**M/NM**	**P/NP**	**TG**
Fusion 1PNP-1MNM	*9.0*	*10.0*	*15.5*
Fusion 3PNP-3MNM	**8.6**	**9.3**	**14.5**

TAB. 5.18 – *Taux d'erreurs (%) pour les 3 tâches de discrimination sur le corpus* **TestRTL**, *en utilisant la fusion de classifieurs.*

Résultats TestRTL : Sur le dernier corpus (*TestRTL*), nous pouvons constater que là encore la fusion de classifieur par vote majoritaire (*Fusion 3PNP-3MNM*) améliore les performances déjà très bonnes de la fusion simple de classifieurs (*Fusion 1PNP-1MNM*). Nous obtenons ainsi, avec la *Fusion 3PNP-3MNM*, un gain relatif significatif de 7% en parole/non-parole, 4.5% en musique/non-musique et 6.5% en discrimination globale, comparé à la *Fusion 1PNP-1MNM*.

Pour conclure cette partie sur la fusion, nous avons tenu à mettre les tableaux 5.19 et 5.20 correspondant à la répartition des trames en parole, musique, parole sur musique sur le corpus *TestRTL*, pour la *Fusion 1PNP-1MNM* et la *Fusion 3PNP-3MNM* respectivement. Nous pouvons observer, en comparant ces tableaux, une réduction significative des segments mal classés en parole et en parole sur musique en utilisant la *Fusion 3PNP-3MNM*.

reconnues / étiquetées	**M**	**PM**	**P**
M	95.1	2.6	1.8
PM	63.3	22.9	13.7
P	4.6	6.2	89.1

TAB. 5.19 – Répartition des trames (%) pour la tâche de discrimination globale en utilisant la Fusion 1PNP-1MNM

reconnues / étiquetées	**M**	**PM**	**P**
M	92.2	5.3	2.4
PM	49.8	37.6	12.6
P	4.2	6.0	89.6

TAB. 5.20 – *Répartition des trames (%) pour la tâche de discrimination globale en utilisant la Fusion 3PNP-3MNM*

En effet, nous avons mieux étiqueté 14.7% de trames de parole sur musique et 0.5% de trames de parole en passant de la fusion du meilleur classifieur P/NP et du meilleur classifieur M/NM à la fusion par vote majoritaire des 3 meilleurs classifieurs P/NP et des 3 meilleurs classifieurs M/NM. Cette dernière amélioration du nombre de trames de parole correctement étiquetées n'est pas négligeable sachant que la parole est prédominante sur ce corpus. De plus, pour des applications comme la détection de mots clés, la transcription automatique, etc., il est très important de ne pas perdre de segments de parole.
On note toutefois, une augmentation du nombre de segments mal classés en musique lorsque nous utilisons le vote majoritaire de six classifieurs (*Fusion 3PNP-3MNM*) au lieu de la simple fusion des meilleurs classifieurs parole/non-parole et musique/non-musique (*Fusion 1PNP-1MNM*).

5.6 Conclusions

Tout au long de ce chapitre, nous avons présenté nos expériences en discrimination parole/musique. Notre approche étant basée sur l'utilisation de classifieurs 'classe/non-classe', nous avons pu séparer les différentes tâches de discrimination et ainsi rechercher les meilleurs paramètres pour chacune des tâches de discrimination : parole/non-parole et musique/non-musique.
Les expériences réalisées portent sur l'utilisation de la décomposition du signal en ondelettes et sur l'utilisation de différentes énergies. Nous avons aussi évalué l'influence de l'ajout des dérivées (premières et secondes) ainsi que l'influence de la variance des paramètres sur les performances en discrimination.
Enfin, plusieurs fusions ont été réalisées :

- au niveau des paramètres, en combinant MFCC et paramètres en ondelettes
- au niveau des classifieurs, en combinant d'une part le meilleur classifieur parole/non-parole et le meilleur classifieur musique/non-musique et d'autre part en combinant par vote majoritaire les 3 meilleurs classifieurs parole/non-parole et les 3 meilleurs classifieurs musique/non-musique.

Nous avons pris comme base de comparaison la paramétrisation MFCC, largement utilisée en discrimination parole/musique.
La première conclusion que nous pouvons établir est que notre paramétrisation basée sur les ondelettes nous a permis d'obtenir une représentation plus compacte du signal que les MFCC. Nous sommes passés d'une paramétrisation MFCC à 36 composantes à des paramétrisation en ondelettes comprenant entre 5 et 14 composantes.
Aussi sur les différents corpus : $Scheirer$, $News$, $Entertainment$ et $TestRTL$, l'utilisation des ondelettes a permis une amélioration significative des performances en discrimination, que ce soit en musique/non-musique ou en discrimination globale.

Cette amélioration n'a pas été observée pour la tâche de discrimination parole/non-parole. Nous avons seulement réussi à approcher les résultats obtenus par la paramétrisation MFCC. Cela peut s'expliquer par le fait que les MFCC sont déjà très performants dans ce domaine. En effet, les MFCC sont quasiment incontournables en reconnaissance de la parole car ils modélisent très bien la parole.
En ce qui concerne le choix des meilleurs paramètres en ondelettes, nous avons observé au cours du processus de sélection que :

- au niveau du **choix de l'ondelette** :
  Il existe une grande variété de familles d'ondelettes et le seul moyen de trouver l'ondelette adéquate est d'expérimenter. Cependant nous n'avons pas observé de grandes différences sur les résultats en discrimination avec les différentes familles d'ondelettes utilisées.
  Nous avons tout de même observé que les meilleurs résultats ont été obtenus avec des ondelettes possédant un petit nombre de moments nuls. Lorsque l'ondelette a peu de moments nuls, la taille de son support est lui aussi réduit. L'ondelette est alors capable de détecter les variations brutales du signal, même de très courte durée. Pour rappel, nous n'utilisons, pour nos paramètres en ondelettes, que les coefficients d'ondelettes (ou de détails). Ces coefficients correspondent aux variations captées par l'ondelette.

- au niveau du **nombre de bandes de décomposition** :
  Nous avons établi que la discrimination parole/non-parole nécessitait un nombre plus petit de décomposition que la discrimination musique/non-musique. Ainsi, les meilleurs résultats en parole/non-parole ont été réalisés avec 5 niveaux de décomposition en ondelettes contre 7 niveaux de décomposition pour la discrimination musique/non-musique. Il avait déjà été établi dans [Deviren 04] que le nombre de niveaux de décomposition en ondelettes pour la reconnaissance de la parole était compris entre 4 et 6.

- au niveau des **énergies calculées sur les coefficients d'ondelettes** :
  C'est l'énergie de Teager qui apparaît comme étant la plus adéquate à la fois en discrimination parole/non-parole et en discrimination musique/non-musique. En effet, c'est l'énergie qui s'est révélée la plus stable au cours des différentes expériences, bien que l'énergie hiérarchique ait donné de bons résultats en discrimination musique/non-musique.

- au niveau de **l'ajout des dérivées premières et secondes des paramètres** :
  L'adjonction de la dynamique des paramètres par l'intermédiaire des dérivées premières (Δ) s'est avérée payante. Nous avons réussi à dépasser les performances des

MFCC en discrimination parole/musique. Cependant, si l'ajout des dérivées premières (Δ) a apporté un réel gain de performance, il n'en a pas été de même lorsque nous avons ajouté les dérivées secondes ($\Delta\Delta$). Le gain était alors très faible voir inexistant. Nous avons même parfois constaté une dégradation des performances dans le cas de la discrimination musique/non-musique.

- au niveau de **l'utilisation de la variance des paramètres** :
  Nous confirmons l'hypothèse de Scheirer [Scheirer 97] selon laquelle l'utilisation de la variance des paramètres donne de meilleurs résultats en discrimination parole/musique que les paramètres eux-mêmes. Avec l'utilisation de la variance des paramètres calculée sur une fenêtre d'une seconde, nous avons obtenu des résultats équivalents à ceux obtenus lors de l'ajout de la dynamique à court terme (ajout des Δ). Mais si les résultats obtenus avec la variance sont légèrement moins bons que ceux obtenus avec l'ajout des Δ, il s'avère que les paramètres basés sur la variance sont plus stables sur les différents corpus. En effet, lors du test de validation que nous avons effectué sur le corpus $TestRTL$, la variance sur une seconde des paramètres a confirmé ses bons résultats alors que les paramètres basés sur l'ajout des dérivées premières ont donné de mauvais résultats en discrimination parole/non-parole. La variance de nos paramètres en ondelettes semble être la meilleure paramétrisation pour la discrimination parole/musique.

Enfin nous avons réalisé deux expériences de fusion. La première consistait à combiner simplement les sorties du meilleur classifieur parole/nom-parole et du meilleur classifieur musique/non-musique obtenus auparavant.
La seconde expérience de fusion a consisté à prendre, pour chaque tâche de discrimination parole/non-parole et musique/non-musique, les trois meilleurs classifieurs et à les fusionner en faisant un vote majoritaire. Pour sélectionner ces trois classifieurs, nous les avons pris de manière à ce qu'ils soient assez hétérogènes pour donner des résultats différents et ainsi ne pas commettre les mêmes erreurs.
Nous avons donc pris pour la discrimination parole/non-parole, un classifieur basé sur les paramètres statiques, un basé sur les paramètres dynamiques à court terme et un basé sur la variance des paramètres. Nous avons fait cette même sélection pour la discrimination musique/non-musique. Au final, chacune de ces expériences a permis d'améliorer les performances en discrimination parole/musique.

Le vote majoritaire, bien qu'étant une fusion très simple, a ainsi permis un gain relatif de 53% en discrimination globale sur le corpus très difficile $TestRTL$, comparé à la paramétrisation MFCC de référence (cf. tableau 5.21). La fusion de classifieurs est donc une voie à explorer pour encore améliorer la discrimination parole/musique.

Paramétrisation	**M/NM**	**P/NP**	**TG**
MFCC+Δ+Δ	*24.9*	**8.1**	*30.9*
Fusion 3PNP-3MNM	**8.6**	9.3	**14.5**

TAB. 5.21 – *Taux d'erreurs (%) pour les 3 tâches de discrimination sur le corpus TestRTL.*

6

Le système de détection de mot clés

Sommaire

Durant cette thèse financée par une une bourse CIFRE, nous avons été amenés à travailler avec l'entreprise TNS-Média Intelligence (TNS-MI). TNS-MI est une des branches européennes du grand groupe anglais Taylor Nelson Sofres. Cette entreprise internationalement reconnue est le leader de la veille media en Europe. Une de leur activité consiste à faire de la veille pluri-média : presse écrite, audiovisuelle et Internet. Ils scrutent ainsi l'actualité pour leurs clients, suivent l'évolution d'une nouvelle intéressante ou qui peut avoir un impact sur le client. Une fois les informations recueillies et jugées intéressantes d'après les critères du client, elles sont regroupées puis transmises au client.
Ce travail fastidieux est actuellement réalisé par des personnes qui écoutent ou regardent

les différents médias. L'objectif que nous avions établi avec TNS-MI était de réaliser un système permettant de détecter les différents mots-clés donnés par leurs clients. Réaliser un tel système permet ainsi d'apporter une aide aux personnes dans leur travail de veille radiophonique.

6.1 Présentation de l'application

Actuellement, la détection de mots clés dans les flux audios se fait manuellement chez TNS-MI. Plusieurs personnes sont chargées d'écouter les flux audio. Chaque personne peut écouter deux émissions à la fois, une pour chaque oreille. Ces personnes doivent aussi apprendre une liste de mots clés, redéfinie quasi-quotidiennement, afin de pouvoir détecter ces mots clés dans les flux et générer des alertes. Les alertes sont ensuite vérifiées et si elles s'avèrent intéressantes, la partie de l'émission correspondante, voire l'émission entière, est retranscrite.
Notre travail consiste à remplacer la tâche fastidieuse d'écoute et de détection de mots clés faîte par plusieurs personnes, par un système automatique de détection de mots clés. De plus, l'utilisation d'un système automatique de détection de mots permettra d'étendre la veille radiophonique. Cette veille est actuellement limitée à quelques radios à cause de la limitation humaine : une personne ne peut être chargée que d'au plus deux radios. Nous pouvons à présent définir ce qu'est un système automatique de détection de mots clés.

Un système automatique de détection de mots clés est un système permettant de rechercher des mots spécifiques dans un flux audio. Le flux audio est ici continu et correspond aux émissions de radios françaises. Le système nécessite une liste de mots clés qui est définie préalablement par l'utilisateur. Cette liste peut être de grande taille et dépasser plusieurs milliers de mots. De plus, elle n'est pas statique et l'utilisateur (TNS-MI) peut la modifier à tout moment en fonction des demandes de ses clients. Le système de détection de mots clés génère, pour chaque mot clé détecté, une alarme que l'opérateur de l'application pourra ensuite vérifier et traiter ultérieurement. Il pourra par exemple, si l'information autour du mot clé s'avère pertinente, retranscrire l'émission en partie ou entièrement suivant la demande du client.
Cette utilisation des systèmes de détection de mots clés n'est qu'une application parmi tant d'autres. Par exemple, la détection de mots clés peut être utilisée dans des systèmes à commandes vocales utilisant la parole spontanée [Riccardi 97, Foote 95], dans des applications d'indexation [Gelin 97, Gelin 96], de reconnaissance par le contenu ou de détection de thème [Gorin 97].
Les systèmes de détection de mots-clés, que ce soit pour de l'indexation de documents audio, pour la gestion d'automates par commandes vocales, ont une structure générale qui peut être vue comme une combinaison de mots clés et de séquences de parole (ou de bruit) constituées de mots non-clés. Un système de détection doit d'une part modéliser les mots non-clés afin de réduire les fausses acceptations, c'est-à-dire lorsque un mot non-clé

est reconnu comme un mot clé, et d'autre part modéliser les mots clés afin de pouvoir les détecter au milieu des autres mots. L'objectif du système est de réaliser un taux élevé de détection et de minimiser le nombre de fausses acceptations. Les applications de détection de mots clés sont principalement basées sur deux méthodes.
La première méthode consiste à associer à chaque mot clé, un modèle et un anti-modèle [BenAyed 03, Sukkar 96]. L'anti-modèle, aussi appelé modèle poubelle (*garbage model*) modélise tout excepté le mot clé considéré. La détection s'effectue en recherchant le modèle de chacun des mots clés dans le signal de parole. Ici, la transcription de tout le signal de parole n'est pas nécessaire, la recherche s'effectuant localement tout au long du signal. Une dernière comparaison entre les mots clés retenus et leurs modèles poubelles associés permet de valider chacun des mots clés ou de les rejeter. La figure 6.1 décrit un exemple d'architecture d'un système de détection de mots clés basé sur des modèles poubelles.
Les systèmes utilisant cette méthode ont l'avantage d'être moins sensibles à certains

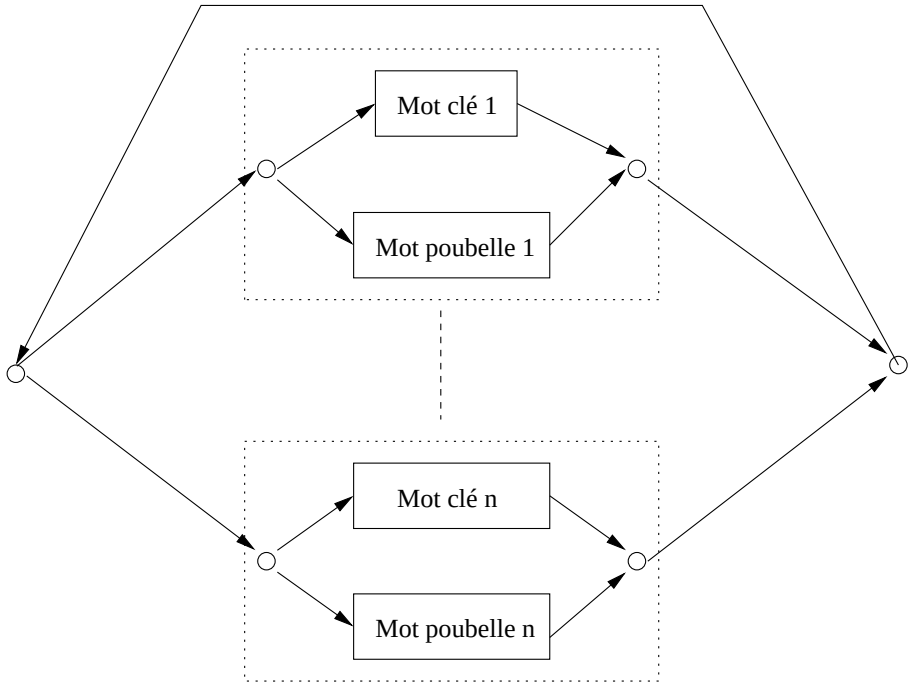

FIG. 6.1 – Description d'un système de détection de mots clés basé sur l'utilisation d'un réseau de mots clés et de modèles poubelles.

problèmes tels que des hésitations et des reprises que l'on retrouve souvent en parole spontanée [Wilpon 90].
La seconde méthode consiste à effectuer une reconnaissance grand vocabulaire du signal de parole. Nous obtenons ainsi une transcription du flux audio dans laquelle sont recherchés les différents mots clés. Cette méthode facilite la recherche de mots clés, une simple recherche textuelle peut suffire. Elle permet aussi de diminuer le nombre de fausse détection de mot clé. En effet, les mots clés correspondant à des sous-parties de mots et étant détectés comme des mots clés par la méthode précédente (par exemple : "thèse"

dans "hypothèse"), ne le sont plus avec cette méthode. Cette méthode nécessite tout de même la mise en place d'un système de reconnaissance automatique de la parole grand vocabulaire [Rose 95]. De plus, ce système de reconnaissance doit être performant afin d'obtenir la meilleure transcription possible. La performance de la détection de mots clés dépend de la qualité de cette transcription.
C'est cette seconde méthode que nous avons choisie d'utiliser pour notre système de détection de mots clés. En effet, la méthode à base de modèles poubelles n'est utilisé que dans des systèmes ne comprenant qu'un petit nombre de mots clés (moins d'une centaine). Ici, avec un nombre de mots clés très important, de l'ordre de plusieurs milliers de mots clés, et avec de la parole spontanée non contrainte (émissions radios), la méthode basée sur un système de reconnaissance grand vocabulaire s'avère plus adéquate.
Le système que nous proposons, illustré par la figure 6.2, possède une architecture modulaire. Il est lui même basé sur ANTS (*Automatic News Transcription System*), le système de transcription automatique du loria [Brun 04]. Cette architecture nous permet de développer et d'améliorer chaque module séparément. Comme le montre la figure 6.2, le système prend en entrée le flux audio de la radio que nous devons surveiller et donne en sortie les mots clés qu'il a détectés ainsi que les positions de ces mots clés dans le flux audio. L'utilisateur peut ensuite récupérer la transcription donnée par le système et écouter la bande sonore pour vérifier que le mot-clé a bien été prononcé.

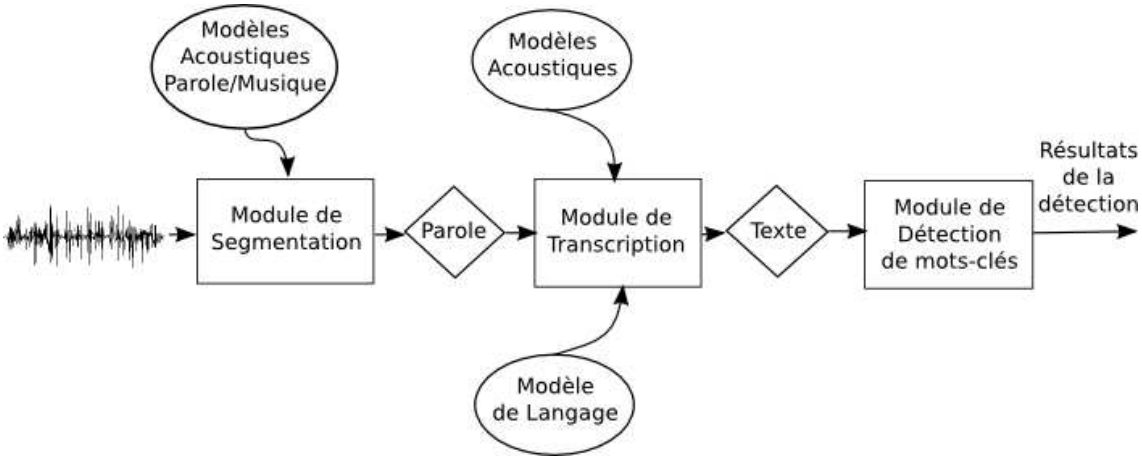

FIG. 6.2 – Description de notre système de détection de mots clés. Le flux est tout d'abord segmenté et la parole est séparée de la musique. Le module de transcription nous fournit le texte correspondant à la parole en entrée. Enfin le module de détection recherche dans le texte les différents mots-clés et génère des alarmes.

Le système complet comprend trois modules :

– le **module de segmentation** permettant de récupérer la parole du flux audio.

- le **module de transcription automatique** qui génère le texte correspondant aux segments de parole provenant du module de segmentation.

- le **module de détection de mots-clés** qui, comme son nom l'indique, extrait les mots clés du texte provenant du module de transcription.

Nous décrivons maintenant ces différents modules constitutifs du système. Pour chacun d'eux nous établissons les besoins de l'application et les contraintes que nous nous sommes donné ou qui nous ont été imposées par l'entreprise ainsi que les solutions que nous avons proposé.

Avant de commencer la description des différents modules, nous pouvons donner quelques informations sur la paramétrisation du signal. La paramétrisation utilisée dans les différents modules et sous-modules sera parfois différente d'un module à l'autre. Elle est basée sur les MFCC excepté dans le cas de notre module de segmentation parole/musique. La fenêtre d'analyse est partout de 32ms et les paramètres sont extraits toutes les 10ms.

6.2 Description des différents modules

6.2.1 Le module de segmentation

Le but de la segmentation est de découper le signal audio en segments acoustiquement homogènes. Ce premier module prend donc en entrée le flux audio et nous renvoie, après segmentation, les segments contenant de la parole. Tout les segments correspondant à de la musique sont rejetés.

Un autre intérêt de cette étape de segmentation est de permettre l'utilisation de méthodes ou d'algorithmes spécifiques en fonction de la nature des segments dans les modules suivants. C'est pourquoi dans ce module, une première étape de segmentation est effectuée pour séparer la parole téléphonique de la parole non-téléphonique. Comme l'illustre la figure 6.3, le module de segmentation est composé de plusieurs sous-modules se succédant et permettant :

- la séparation de la parole large bande, aussi appelée parole studio, de la parole téléphonique,
- la détection et le rejet des parties musicales, pour ne conserver que la parole studio,
- la détection des silences et des respirations pour obtenir des segments de parole acceptables pour le moteur de reconnaissance dans le module de transcription,
- la séparation des locuteurs masculins et féminins,
- le regroupement des segments de parole prononcés par des locuteurs ayant des voix similaires.

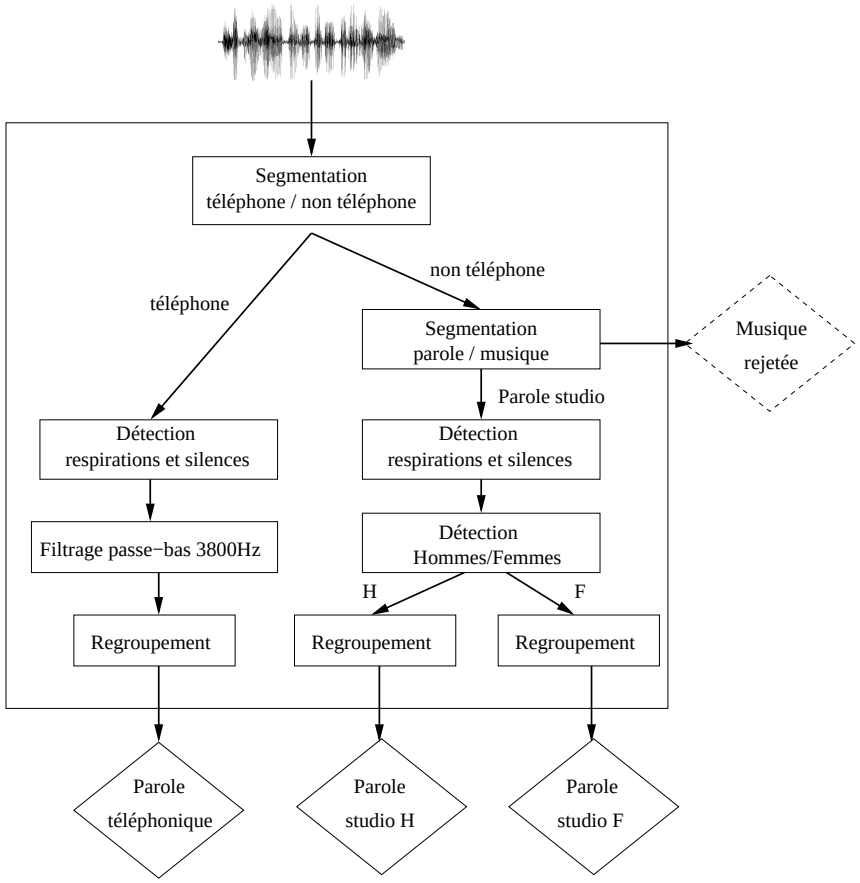

FIG. 6.3 – Description du module de segmentation du système de détection de mots clés.

Nous allons maintenant décrire les approches mises en oeuvre dans ces différents sous-modules.

6.2.1.1 Segmentation téléphone/non téléphone

Cette segmentation du signal audio sépare le signal en deux types de segments : des segments de signal correspondant à un locuteur au téléphone et le reste. Comme nous pouvons le constater sur la figure 6.4, l'énergie se retrouve en dessous de 4kHz pour la parole téléphonique.

L'intérêt de séparer la parole téléphonique de la non-téléphonique est donc évident : nous pouvons construire des modèles de parole spécifiques au signal téléphonique et ainsi améliorer la reconnaissance. De plus, la musique est la plupart du temps absente sur un signal téléphonique. Nous n'aurons donc pas à séparer la parole de la musique dans les segments téléphoniques.
La segmentation téléphone/non-téléphone est réalisée à l'aide de deux mélanges de lois gaussiennes (GMM) à 32 gaussiennes :

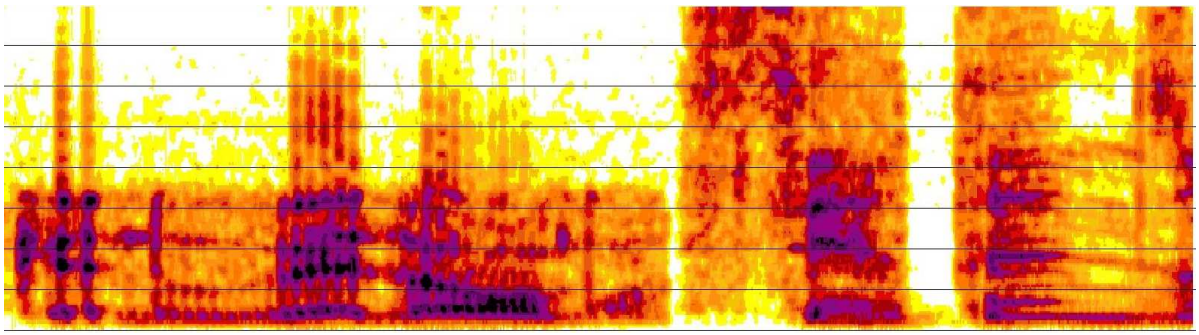

FIG. 6.4 – Spectrogramme représentant un signal correspondant à de la parole téléphonique suivi par de la parole non-téléphonique. La largeur de bande du signal est de 8kHz. On remarque qu'il n'y a pratiquement pas d'énergie au dessus de 4kHz pour le locuteur au téléphone.

- 1 GMM appris sur de la parole "bande étroite", c'est-à-dire de la parole prononcée au téléphone.
- 1 GMM appris sur de la parole "bande large", aussi appelée parole studio.

Pour paramétrer le signal audio, 13 coefficients MFCC sans CMR sont utilisés. Ne pas normaliser en appliquant la soustraction de la moyenne cepstrale (CMR) permet de conserver l'influence du canal de transmission. La segmentation est fondée sur l'algorithme de Viterbi. Aussi, afin d'assurer une durée minimale d'une demi-seconde pour chaque segment détecté, chaque modèle utilisé lors de la reconnaissance est la concaténation de 50 GMMs. Enfin, un filtrage passe-bas à 3800 Hertz est réalisé sur les segments de parole téléphonique afin de supprimer le bruit du canal de transmission pouvant se trouver au dessus de 3800 Hz.

6.2.1.2 Segmentation parole/musique

Ce module prend en entrée les segments de signal non téléphonique, effectue une resegmentation du signal en parole et musique afin de ne conserver que les morceaux du signal correspondant à de la parole. Ces segments seront alors transmis au module de détection de silences et respirations (cf. Figure 6.3). La segmentation parole/musique n'est pas effectuée sur le signal téléphonique car nous considérons que ces parties du signal ne comportent que de la parole, la largeur de bande du téléphone n'étant pas généralement utilisée pour transmettre de la musique. Pour ce sous-module du module de segmentation, nous avons mis en place deux types de segmentation.

Segmentation parole/musique de référence :

La première est celle déjà utilisée dans le système ANTS [Brun 04]. Elle est fondée

sur l'utilisation de quatre modèles constitués de GMMs mis en concurrence. Ces GMMs permettent de modéliser :

- la parole
- la musique instrumentale
- les chansons
- la parole sur fond musical

Chaque modèle a un état, comporte 16 gaussiennes et a été appris avec la boîte à outils HTK [Young 95a]. La paramétrisation du signal est constituée de 39 coefficients : 13 MFCC avec leurs dérivées premières (Δ) et secondes ($\Delta\Delta$), sans normalisation CMR (*Cepstral Mean Removal*). Ces coefficients sont calculés toutes les 10ms sur des fenêtres de 32ms. Nous avons défini une durée minimale d'une demi-seconde qui est codée directe-

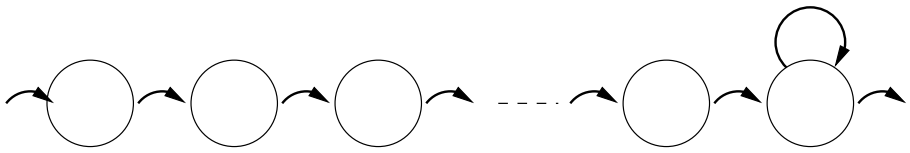

FIG. 6.5 – Topologie des HMMs utilisés pour l'approche de mise en compétition de 5 modèles dans le module de classification parole/musique. 50 GMMs à un état sont concaténés pour définir une durée minimale d'une demi-seconde.

ment dans les modèles afin d'éviter des segments irréalistes de 10ms. Comme le montre la figure 6.5, chacune des classes est donc modélisée par un modèle à 50 états. L'alignement sur le signal audio de la meilleure séquence des modèles en compétition est réalisé par de l'algorithme de Viterbi. Cet alignement nous fournit la segmentation du flux en parole et musique. Les segments de parole et parole sur fond musical sont regroupés sous une seule étiquette : "parole" et les segments de musique instrumentale et de chansons sont eux regroupés sous l'étiquette "musique" et définitivement éliminés.

Notre segmentation parole/musique :

La seconde méthode de segmentation consiste à utiliser notre système de segmentation parole/musique basée sur une paramétrisation en ondelettes. Cette méthode utilise une toute autre approche : une approche dite "Classe/Non Classe". L'intérêt est d'ici séparer la parole de la non-parole, la "Classe" et la "Non Classes" est donc ici la parole et la non-parole. Nous avons modélisé ces classes et non-classes à l'aide de 2 GMMs :

- 1 GMM pour la parole,
- 1 GMM pour la non-parole.

Comme pour le système précédent, pour éviter d'avoir des segments trop courts, nous avons modélisé chacune des classes comme la succession de 100 modèles GMMs. Le modèle parole est mis en compétition avec son modèle non-classe. Une décision est alors prise pour savoir si le segment contient ou non de la parole. Pour plus de détails, le lecteur peut se référer à la section 5.1 où notre système de segmentation parole/non-parole est décrit de manière précise.

6.2.1.3 Détection des respirations et des silences

Le sous-module de détection des respirations et des silences prend en entrée les segments de signal correspondant à de la parole, qu'elle soit téléphonique ou non. Ce module nous permet d'une part de réduire la taille des segments préalablement obtenus et d'autre part de trouver les groupes de souffle qui correspondent souvent à des entités syntaxiques ou sémantiques. Ce sous-module renvoie donc des segments de parole découpés en segments plus petits qui pourront être pris en charge par le module de transcription. Pour réaliser cette segmentation, une reconnaissance phonétique du signal est effectuée en utilisant des modèles de phonèmes monophones, un modèle de respiration et un modèle de silence. Ces modèles sont appris à l'aide de l'outil HTK.
Comme ce sous-module est utilisé pour les deux types de parole téléphonique et non-téléphonique, nous devons apprendre des modèles spécifiques à chaque type de parole. Nous avons ainsi :

- 36 modèles de phonèmes, un modèle de silence et un modèle de respiration appris sur du corpus "bande large".
- 36 modèles de phonèmes, un modèle de silence et un modèle de respiration appris sur du corpus "bande étroite".

Chacun des modèles est un HMM gauche-droit à trois états sans saut avec 2 gaussiennes par état.
La paramétrisation du signal pour cette reconnaissance phonétique est basée sur les MFCC : 13 MFCC, 13 Δ et 13 $\Delta\Delta$. Une normalisation est effectuée en utilisant la soustraction de la moyenne cepstrale (*CMR : Cepstral Mean Removal*).
La grammaire utilisée pendant cette reconnaissance attribue la même probabilité de transition entre les modèles.
Les segments sont alors découpés lorsque l'on trouve une respiration d'une durée minimale de 150ms ou un silence d'au moins 300ms.

6.2.1.4 Segmentation hommes/femmes

Le sous-module de segmentation hommes/femmes resegmente le signal suivant le genre du locuteur. Cette segmentation permet d'utiliser des modèles de parole "hommes" et des

modèles de parole "femmes" lors de la phase de transcription qui suit celle de segmentation. Ceci afin d'améliorer la reconnaissance de la parole et par conséquent la transcription fournie par le système de reconnaissance.
Ce sous-module est similaire au sous-module de segmentation "téléphone/non-téléphone". Nous avons ici :

- 1 modèle GMM appris sur de la parole prononcée par des locuteurs masculins
- 1 modèle GMM appris sur de la parole prononcée par des locuteurs féminins

Chaque GMM est un mélange de 256 gaussiennes. De plus, lors de la reconnaissance à l'aide de l'algorithme de Viterbi, afin d'assurer une durée minimale d'une demi-seconde par segment pour éviter une alternance homme/femme irréaliste, chaque modèle va être la concaténation de 50 GMMs. Enfin, la paramétrisation utilisée pour séparer les locuteurs hommes et femmes est la suivante : 13 MFCC + Δ + $\Delta\Delta$ avec normalisation CMR (*Cepstral Mean Removal*).

6.2.1.5 Regroupement par locuteurs

Ce module regroupe les segments de parole prononcés par des locuteurs ayant des voix similaires afin d'effectuer une adaptation des modèles phonétiques dans une étape ultérieure.
L'algorithme permettant de regrouper deux segments est fondé sur le critère d'information bayésien (BIC pour *Bayesian Information Criterion*) et s'inspire d'une méthode proposée par Perrine Delacourt [Delacourt 00] :

- chaque segment est représenté par une loi gaussienne,
- pour chaque couple de segments, le BIC est calculé,
- le couple de segments qui maximise le BIC est fusionné en un segment unique.

Le processus est itéré tant qu'il existe des valeurs de BIC positives. Chaque segment ainsi obtenu est considéré comme ayant été prononcé par un seul locuteur (ou par des locuteurs ayant des voix proches).
La paramétrisation du signal utilisée pour le regroupement est la suivante : 13 MFCC + Δ + $\Delta\Delta$ sans CMR.

6.2.2 Le module de transcription

Pour obtenir la transcription automatique des segments de parole issus du module de segmentation décrit précédemment, il nous faut un système complet de transcription automatique de la parole. Nous avons besoin d'un moteur de reconnaissance automatique grand vocabulaire de la parole, de corpus audio et textuel pour construire les modèles

acoustiques et un modèle de langage. Nous avons besoin également de définir un lexique qui correspondra aux mots pouvant être reconnus par le système. Nous décrivons tout ceci dans cette section.

6.2.2.1 Le moteur de reconnaissance : Julius

Pour l'application d'aide à la veille radiophonique envisagée, le moteur de reconnaissance doit pouvoir traiter très rapidement un très grand vocabulaire. En effet, la reconnaissance automatique des segments de parole devra se faire au final en temps réel pour notre application de détection de mots clés permettant d'aider à la veille radiophonique. Pour le choix du moteur de reconnaissance, nous avons à notre disposition plusieurs moteurs performants. Nous pouvons citer par exemple :

- le système *ISIP* [ISIP] (*Mississipi State University*),
- le système *Sphinx* [Sphinx] (*Carnegie Mellon University*),
- le système *Julius* (*Interactive Speech Technology Consortium*) [Julius , Lee 01],
- etc.

Nous avons choisi d'utiliser le moteur de reconnaissance gratuit *Julius*. *Julius* est un moteur de reconnaissance grand vocabulaire développé à l'origine par des chercheurs de l'université de Kyoto et en particulier Akinobu Lee [Lee 01]. Le projet est maintenant poursuivi par l'*Interactive Speech Technology Consortium* qui regroupe différents universitaires japonais. Julius est un système très performant, paramétrable et peut être temps réel. Dans Julius, la reconnaissance est effectuée en deux passes.

La première passe consiste à construire un graphe d'exploration, aussi appelé graphe de mots, correspondant au décodage de la phrase considérée.
Cette première passe ou "passe avant" parce qu'elle s'effectue dans le sens normal de lecture, effectue quelques approximations pour accélérer le décodage :

- utilisation d'un modèle de langage bigramme au lieu du modèle trigramme,
- possibilité de sélectionner différentes techniques d'élagage des fonctions de densité gaussienne,
- limitation de la largeur du faisceau de recherche à un nombre maximum d'hypothèses,
- approximation de la dépendance au contexte du mot suivant.

Le décodage repose sur l'algorithme de Viterbi. Le moteur de reconnaissance procède trame par trame. La première passe génère finalement un graphe de mots contenant un ensemble restreint d'hypothèses parmi lesquelles le système de reconnaissance effectuera la recherche de la solution. Ce graphe de mots interne au moteur de reconnaissance est

généré de manière trame-synchrone. Pour chaque trame, le graphe contient l'ensemble des mots du lexique qui peuvent finir à cette trame après élagage. Pour chaque mot, différentes informations sont accessibles :

- les temps de début et de fin du mot,
- la probabilité acoustique du mot,
- la probabilité bigramme du mot,
- un lien vers le mot prédécesseur au sens de Viterbi,
- le score cumulé depuis le début de la phrase du meilleur chemin menant à ce mot.

La seconde passe de Julius, aussi appelée "passe arrière", utilise le graphe de mots généré lors de la première passe. C'est la dernière étape de la reconnaissance avant de donner la phrase correspondant à la meilleure hypothèse à l'utilisateur. Cette seconde passe a comme particularité de se dérouler en sens inverse de lecture : de la fin de la phrase vers le début, d'où son nom de "passe arrière". La reconnaissance se fait à l'aide d'un algorithme à pile de type A^* [Nilson 71]. La phase de recherche du meilleur chemin est basée sur le graphe de mots généré lors de la première passe. Pendant la deuxième passe, les probabilités acoustiques et linguistiques sont recalculées avec des modèles plus fins (modèles trigrammes).

6.2.2.2 Paramétrisation

La paramétrisation du signal utilisée par le moteur de reconnaissance grand vocabulaire est basée sur l'utilisation des coefficients cepstraux à échelle Mèl (MFCC). Le signal est échantillonné à 16kHz et nous utilisons une fenêtre d'analyse de type Hamming d'une durée de 32ms. Les paramètres sont calculés toutes les 10ms.
Le vecteur d'observation associé à la fenêtre d'analyse est constitué de 13 coefficients cepstraux, C_0 inclus, ainsi que des dérivées premières et secondes de ces 13 coefficients. Enfin, une normalisation CMS, c'est-à-dire une soustraction de la moyenne cepstrale, est effectuée. Cette normalisation est nécessaire pour réduire les différences entre les enregistrements dûes aux canaux de transmission et aux microphones utilisés. En effet, les émissions radios comportent une grande diversité d'environnements (enregistrements dans la rue, dans un studio d'enregistrement, etc.) et de microphones (téléphone, micro de rue, micro de studio, etc.).

6.2.2.3 Lexique

Le lexique est constitué de 60000 mots. 55000 correspondent aux mots les plus fréquents du corpus textuel du journal "Le Monde" de 1987 à 2003. Les 5000 autres mots sont les mots les plus fréquents des corpus d'apprentissage et de développement d'ESTER. Après phonétisation, le lexique contient 112000 prononciations.

Pour la détection de mots clés, il faudra bien sûr que les différents mots clés recherchés fassent partie du lexique pour pouvoir être reconnus.

6.2.2.4 Modèle de langage

Le modèle de langage utilisé est obtenu par interpolation linéaire de deux modèles trigrammes. Ces modèles sont générés :

- à partir du corpus textuel du journal "Le Monde" pour le premier modèle trigramme.
- à partir des transcriptions des corpus d'apprentissage et de développement d'ESTER pour le second modèle trigramme.

Les poids pour l'interpolation sont de 0.55 pour le modèle de langage "Le Monde" et de "0.45" pour l'autre modèle de langage. Au final, le modèle de langage contient 7.4 millions de bigrammes et 25.4 millions de trigrammes.

6.2.2.5 Modèles acoustiques pour la transcription

A l'aide du corpus ESTER, nous avons appris des modèles HMM à l'aide du logiciel HTK. Nous avons au final 40 modèles :

- 36 modèles phonétiques à trois états,
- 1 modèle correspondant à une courte pause,
- 1 modèle de respiration,
- 1 modèle de "bruit de bouche",
- 1 modèle correspondant à tous les autres bruits.

Tous ces modèles ont trois états, sauf le modèle de pause courte qui n'en comporte qu'un seul. Ces modèles triphones sont ceux utilisés par Julius pour la reconnaissance. Cependant nous avons vu que le signal a été segmenté en :

- parole téléphonique (parole bande étroite) et parole studio (parole bande large),
- genre (homme ou femme).

Nous devons donc effectuer différentes phases d'adaptation de nos modèles acoustiques triphones.

Adaptation bande large et genre

Les 40 modèles sont appris sur la partie du corpus d'apprentissage correspondant à de la parole large bande. Cette partie du corpus d'apprentissage a été extraite en utilisant le module de segmentation automatique téléphone/non-téléphone. Les modèles dépendant du genre sont ensuite obtenus en utilisant 8 itérations d'une adaptation SMAP (*Structural*

Maximum A Posteriori).

Adaptation bande étroite

Afin d'augmenter artificiellement la quantité de données audio "bande étroite", l'intégralité du corpus d'apprentissage a été filtré entre 300 et 3800Hz. Les modèles triphones ont été appris sur ce corpus filtré. Une adaptation MAP a ensuite été effectuée sur la partie "bande étroite" du corpus d'apprentissage.

Adaptation au locuteur

Nous adaptons ici au locuteur, de manière non supervisée, les modèles phonétiques qui sont utilisés par le moteur de reconnaissance. Pour cela, on effectue d'abord une première reconnaissance des segments obtenus à l'étape de regroupement en locuteurs, en utilisant des modèles phonétiques non adaptés. Grâce à cette reconnaissance, ces modèles sont ensuite adaptés à l'aide d'une adaptation SMLLR[15] "*block-diagonal*". Pour ceci nous utilisons trois blocs pour la matrice de transformation qui correspondent aux coefficients statiques, aux dérivées premières et aux dérivées secondes.

Travail sur la modélisation acoustique

Les modèles acoustiques doivent être appris sur un corpus d'apprentissage aussi proche que possible des conditions de test. En d'autres termes, nous avons besoin suivant la tâche à réaliser, d'avoir un corpus d'apprentissage comprenant des conditions d'enregistrement, des bruits, des locuteurs, etc. qui recouvre au maximum ce que nous pourrons retrouver dans le corpus de test. Il faut également que le signal audio du corpus d'apprentissage et celui du corpus de test aient les mêmes caractéristiques : taux de compression, fréquence d'échantillonnage, etc. Nous travaillons ici sur la transcription d'émissions radiophoniques. Nous avons donc besoin d'un corpus audio d'émissions de radio. Dans le cadre de cette thèse CIFRE, pour construire des modèles acoustiques propre à l'entreprise, nous avons mis en place un système d'acquisition d'émissions de radio. En parallèle, nous avons récupéré de nombreuses transcriptions manuelles d'émissions de radio correspondant aux périodes d'acquisition audio. Pour certaines transcriptions manuelles, nous avons ainsi le signal audio correspondant. A partir de cela, il nous est possible de faire un alignement forcé entre le texte et l'audio afin d'apprendre de nouveaux modèles acoustiques. L'alignement réalisé est basé sur la programmation dynamique. Nous utilisons l'algorithme DTW (*Dynamic Time Warping*). Cependant, par manque de temps et de corpus, nous n'avons pas pu terminer la construction de ces nouveaux modèles acoustiques. Nous utiliserons donc les modèles acoustiques précédemment décrits.

[15]SMLLR signifie *Structural Maximum Likelihood Linear Regression*

6.2.3 Le module de détection de mots-clés

Ce module permet, à partir d'une liste de mots clés et de la transcription du signal, de détecter les mots clés et leurs positions dans le signal. Il donne aussi en sortie la transcription du segment audio dans lequel se trouve le mot clé. Le fichier créé par ce module contient ainsi des entrées de la forme :

mot clé - position - transcription du segment dans lequel se trouve le mot clé

Au stade actuelle, le module de détection est assez basique. Il consiste en une simple recherche textuelle dans la transcription fournie par Julius.

6.3 Conclusions

Nous venons de présenter notre système de détection de mots clés. Ce système est basique, et nécessite de nombreuses améliorations. La première modification que nous avons apportée se situe au niveau de la segmentation parole/musique. Nous avons remplacé l'approche "compétitive" utilisant 5 modèles GMMs représentant la parole téléphonique, la parole non-téléphonique, la musique instrumentale, la musique chantée et la parole sur fond musical et basée sur la paramétrisation MFCC par notre approche "classe/non-classe" basée sur notre paramétrisation en ondelettes.
Nous n'avons pas encore entièrement évalué l'impact du nouveau module de segmentation parole/musique sur les performances en transcription. Cependant, nous pouvons déjà penser à quelques perspectives d'améliorations au niveau :

- du moteur de reconnaissance : le module de segmentation parole/musique que nous avons réalisé permet de reconnaître la parole sur fond musical. Il serait donc intéressant de l'utiliser pour découper le corpus d'apprentissage et apprendre des modèles acoustiques spécifiques à la parole en présence de musique. Les segments détectés comme de la parole en présence de musique pourraient alors être mieux transcrits. La deuxième amélioration au niveau du moteur de reconnaissance se situe au niveau de la modélisation du langage. En effet, notre collaboration avec l'entreprise TNS nous a permis d'obtenir de nombreuses transcriptions d'émissions radiophoniques. Ces transcriptions sont d'excellente qualité et ont la particularité de retranscrire exactement ce qui a été prononcé, c'est-à-dire qu'elle ne sont pas réécrites avec un vocabulaire plus soigné. Ainsi, les hésitations, les reprises, les coupures entre interlocuteurs, la façon de parler sont conservées et se retrouvent dans la transcription de l'émission. A partir de ces transcriptions, nous pouvons construire une nouvelle modélisation du langage plus proche de la parole spontanée que l'on retrouve dans la plupart des émissions de radio.

– de la détection de mots clés : de nombreuses améliorations peuvent être apportées à ce module. Parmi les améliorations, une semble rapidement applicable : l'ajout d'une mesure de confiance. En effet, le moteur de reconnaissance Julius comprend une mesure de confiance intégrée dans l'étape de reconnaissance [Lee 04]. Cette mesure de confiance sur les mots est calculable au cours de la phase de décodage de la deuxième passe de Julius. Ajouter une mesure de confiance dans notre système nous permettrait de pouvoir valider les mots clés détectés ou tout du moins donner une information supplémentaire à la personne chargée de vérifier la présence des mots clés détectés [Razik 07a, Razik 07b].

7

Conclusions et perspectives

Les travaux présentés dans ce mémoire concernent la segmentation parole/musique. Cette tâche est essentielle dès que l'on veut traiter des documents audio. En effet, que ce soit pour des applications d'indexation, de transcription ou d'aide au sous-titrage, la nécessité de pouvoir discriminer la parole de la musique est évidente.

Dans le cadre de cette thèse CIFRE, nous avons été amenés à développer une application de détection de mots clés basée sur un système complet de reconnaissance automatique de la parole. L'application est décrite au chapitre 6. Dans ce système, notre travail s'est plus particulièrement porté sur le module de segmentation parole/musique.

Ce système de détection de mots clés dans les émissions radiophoniques est basé sur la transcription automatique du flux audio. Or le flux audio n'est pas uniquement composé de parole. En effet, dans les émissions de radios généralistes, sur lesquelles l'entreprise TNS veut appliquer le système de détection de mots clés, la part de musique n'est pas négligeable. De plus, dans les interviews, les émissions ou les journaux, un fond musical est souvent présent. La transcription de ces zones devient alors difficile avec les modèles appris sur de la parole sans fond musical. Tout ceci a motivé notre choix de nous concentrer sur la segmentation parole/musique afin de détecter les zones de parole, de musique mais aussi de parole sur de la musique. Dans ce dernier cas, il est alors possible d'utiliser des modèles spécifiques appris sur de la parole en présence de musique pour améliorer la transcription et par conséquent la détection de mots clés.

Nous avons étudié de nombreux paramètres et classifieurs au chapitre 1 ainsi que différents modules de segmentation parole/musique dans plusieurs systèmes de transcription automatique de la parole dans le chapitre 2. Nous en avons conclu que pour améliorer la segmentation parole/musique, une solution est de trouver de nouveaux paramètres permettant de mieux discriminer la parole et la musique.

Notre principale contribution pour la segmentation parole/musique s'est donc faite au niveau de la paramétrisation du signal.

Comme nous l'avons vu au chapitre 1, les paramètres habituels sont basés pour la plupart sur les caractéristiques soit temporelles soit fréquentielles du signal. Certains pa-

ramètres, comme la modulation à 4Hz [Pinquier 04a], sont basés sur une analyse temps-fréquence du signal. Ces paramètres temps-fréquences ont d'ailleurs donnés de très bons résultats en discrimination parole/musique. Enfin les paramètres MFCC, souvent utilisés en reconnaissance automatique de la parole, ont montré qu'ils étaient capables de modéliser la musique [Logan 00] et sont désormais utilisés en discrimination parole/musique [Carey 99, Williams 99, West 04, Scheirer 97]. Les paramètres cepstraux ont permis d'obtenir de bonnes performances en segmentation parole/musique, c'est pourquoi nous les retrouvons très souvent comme paramètres pour segmenter le signal en parole et musique dans les systèmes de transcription automatique de flux audio. La paramétrisation MFCC est dans notre étude la paramétrisation de référence. Nous avons choisi d'étudier des paramètres temps-fréquences du signal pour discriminer la parole de la musique. Notre nouvelle paramétrisation pour la discrimination est basée sur une décomposition en ondelettes du signal. L'utilisation des ondelettes permet d'extraire des paramètres temps-fréquence mais aussi de traiter les signaux non-stationnaires. Ce dernier point est important. En effet, au quotidien, notre attention est attirée par le mouvement et les phénomènes transitoires, au contraire des stimuli stationnaires qui sont vite ignorés. Notre cerveau donne la priorité aux phénomènes transitoires et permet ainsi de sélectionner les informations importantes de notre environnement. On peut alors penser retrouver ce phénomène lorsque nous distinguons la parole de la musique. L'étude des phénomènes transitoires des signaux de parole et de musique, en utilisant la transformée en ondelettes, doit permettre de discriminer la parole de la musique.

Finalement, notre paramétrisation est basée sur une décomposition multi-échelles du signal à l'aide d'ondelettes et sur le calcul de différentes énergies sur les coefficients d'ondelettes obtenus. Ces coefficients d'ondelettes représentent les variations du signal et nous permettent de détecter les singularités du signal, c'est-à-dire les brusques variations du signal. Les énergies calculées sur ces coefficients sont :

- l'énergie instantanée,
- l'énergie de Teager,
- l'énergie hiérarchique.

La méthode de classification utilisée dans notre système de discrimination parole/ musique est basée sur une approche "classe/non-classe". L'approche considérée nous permet de modéliser les classes de parole et de musique ainsi que leur non-classe : non-parole et non-musique. Il est alors possible de séparer la tâche de discrimination parole/musique en deux sous-tâches : la discrimination parole/non-parole (P/NP) et la discrimination musique/non-musique (M/NM). Nous pouvons ainsi utiliser des paramétrisations différentes pour chacun des sous-systèmes de discrimination afin d'obtenir une meilleure discrimination parole/musique.

Nous avons réalisé de nombreuses expériences pour déterminer les meilleurs paramètres pour les tâches de discrimination P/NP et M/NM. Pour ces différentes expérimentations, nous avons utilisé un important corpus de développement composé de trois sous-corpus

hétérogènes : “Scheirer”, “News” et “Entertainment”.

Au cours de ce processus de sélection, nous avons pu tirer plusieurs conclusions.
La première conclusion se situe au niveau de la décomposition du signal en ondelettes. Nous avons testé différentes familles d'ondelettes : Daubechies, Coiflets, Symlets. Pour chacune de ses familles, nous avons fait varier le nombre de moments nuls des ondelettes. Le nombre de moments nuls d'une ondelette conditionne sa capacité à détecter les singularités dans un signal, c'est à dire les transitions brutales dans le signal. Nous avons observé d'une part que les différentes familles d'ondelettes donnaient des résultats similaires et d'autre part que plus le nombre de moments nuls de l'ondelette était petit, plus les résultats en discrimination étaient bons.
Au niveau du nombre de niveaux de décomposition de la transformée en ondelettes, 5 bandes de fréquence semble être le meilleur choix pour la discrimination P/NP alors qu'une décomposition en 7 bandes semble être optimale pour la discrimination M/NM.

Seconde conclusion : nos paramètres statiques basés sur les ondelettes et différentes énergies améliorent les performances en discrimination M/NM mais pas celles en discrimination P/NP, comparé à la paramétrisation MFCC de référence. Cependant, dès l'utilisation de la dynamique des paramètres, nous obtenons des meilleurs résultats à la fois en discrimination P/NP et M/NM, par rapport aux MFCC. Ainsi l'ajout aux paramètres de leurs dérivées premières a permis d'obtenir un gain relatif de 52% en discrimination P/NP et 72% en discrimination M/NM. De plus, l'utilisation de la variance des paramètres à la place des paramètres eux-mêmes a permis d'obtenir des résultats similaires à ceux obtenus avec l'ajout des dérivées premières. Enfin, au cours de tous ces tests, l'énergie Teager est apparue comme étant la plus adéquate à la fois en discrimination parole/non-parole et en discrimination musique/non-musique. C'est l'énergie qui s'est révélée être la plus stable au niveau des performances au cours des différentes expériences.

Au cours d'un dernier test, dit de validation, nous avons voulu confirmer les résultats obtenus avec nos meilleurs paramétrisation P/NP et M/NM. Nous avons donc sélectionner les paramétrisations suivantes pour la discrimination P/NP :

- les paramètres construits à partir de l'énergie Teager calculée sur les coefficients d'ondelettes obtenus en utilisant l'ondelette Coif1 et 5 niveaux de décomposition ainsi que leurs dérivées premières.
- la variance calculée sur une fenêtre d'une seconde des paramètres construits à partir de l'énergie Teager calculée sur les coefficients d'ondelettes obtenus en utilisant l'ondelette Coif1 et 5 niveaux de décomposition.

, et ceux-ci pour la discrimination M/NM :

- les paramètres construits à partir de l'énergie Teager calculée sur les coefficients d'ondelettes obtenus en utilisant l'ondelette Coif1 et 7 niveaux de décomposition

ainsi que leurs dérivées premières.
- la variance calculée sur une fenêtre d'une seconde des paramètres construits à partir de l'énergie Teager calculée sur les coefficients d'ondelettes obtenus en utilisant l'ondelette Coif1 et 7 niveaux de décomposition.

La paramétrisation basée sur l'utilisation de paramètres dynamiques à long terme, c'est-à-dire sur l'utilisation de la variance calculée sur une seconde, a confirmé ces bons résultats sur le corpus de validation. Ce ne fut pas le cas de la paramétrisation à court terme, basé sur l'ajout des dérivées premières, qui a réalisée une contre-performance en discrimination musique/non-musique. La variance sur une seconde de nos paramètres en ondelettes est donc un paramètre robuste et performant pour la tâche de discrimination parole/musique. Sur ce dernier test de validation, nous avons obtenu un taux d'erreurs en trame de 9% en discrimination M/NM, 10% en discrimination P/NP et 15.5% en discrimination parole/musique avec notre meilleure paramétrisation basée sur la variance, au lieu de 24.9%, 8.1%, 30.9% respectivement en utilisant la paramétrisation MFCC de référence. (cf. tableau 7.1). Le gain relatif obtenu par rapport à la paramétrisation MFCC est donc de 64% en musique/non-musique et 50% en discrimination globale. Les meilleurs résultats obtenus en discrimination parole/non-parole restent ceux obtenus avec les MFCC. Ceci peut s'expliquer par le fait que les MFCC sont connus, dans le domaine de la reconnaissance de la parole, pour bien modéliser la parole.

Pour conclure sur la partie expérimentale, une dernière expérience a été réalisée. Nous avons réalisé une fusion des sorties de trois classifieurs P/NP et de trois classifieurs M/NM. Cette fusion s'est faite en regroupement le résultat du vote majoritaire entre les sorties des classifieurs P/NP et du vote majoritaire entre les sorties des classifieurs M/NM. Cette fusion utilisant un simple vote majoritaire nous a permis d'obtenir un gain relatif de 6.5% en discrimination parole/musique par rapport à la simple fusion des sorties du meilleur classifieur P/NP et du meilleur classifieur M/NM (cf. tableau 7.1).

Param.	**TG**
MFCC+Δ+ΔΔ	*30.9*
Fusion 1PNP-1MNM	*15.5*
Fusion 3PNP-3MNM	**14.5**

TAB. 7.1 – *Taux d'erreurs (%) pour la tâche de discrimination parole/musique sur le corpus de validation* **TestRTL.**

La segmentation parole/musique est une tâche indispensable dans de nombreuses applications indexation, reconnaissance de la parole, transcription, aide au sous-titrage, etc.. Nous avons développé ici une nouvelle approche pour paramétrer le signal afin d'améliorer la segmentation du flux audio en parole, musique et parole sur musique. Nos perspectives

pour la segmentation parole/musique sont multiples. D'une part, nous n'avons pas utilisé tout le potentiel de la transformée en ondelettes. En effet, nous n'avons calculé que des paramètres basés sur différentes énergies calculées sur les coefficients d'ondelettes. D'autres paramètres basés sur les coefficients d'ondelettes devraient être étudiés pour la tâche de discrimination parole/musique :

- la variance, la moyenne des coefficients d'ondelettes,
- l'extraction de pics dans les coefficients d'ondelettes pour détecter différents rythmes dans le signal,
- etc.

Nous pourrions aussi ne sélectionner que certaines bandes de fréquences pour la parole et d'autres pour la musique afin de rechercher des variations dans certaines zones fréquentielles uniquement.

D'autre part, une autre direction à explorer est l'utilisation, non plus de la transformée en ondelettes, mais de la transformée en paquets d'ondelettes pour extraire de nouveaux paramètres du signal. La transformée en paquets d'ondelettes permet d'obtenir de nouveaux découpages temps-fréquence du signal et ainsi d'analyser plus finement l'évolution temporelle de certaines zones fréquentielles (Figure 7.1).

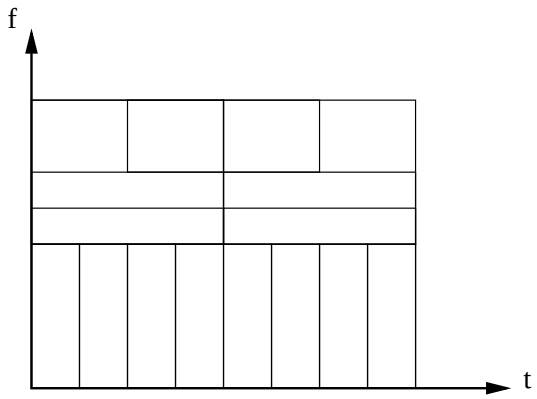

FIG. 7.1 – Exemple de couverture temps-fréquence avec la transformée en paquets d'ondelettes.

Nous pourrions alors sélectionner uniquement certains atomes temps-fréquence à partir de différentes caractéristiques *a priori* des signaux de parole et de musique, afin d'obtenir de nouveaux paramètres pour la tâche de discrimination parole/musique.

Enfin, dans ce mémoire nous nous sommes concentrés sur le problème de la segmentation parole/musique. Il existe de nombreux autres problèmes en segmentation du signal, comme par exemple la segmentation en locuteurs, la segmentation hommes/femmes, etc. Nous pourrions donc, par la suite, nous orienter vers ces autres problèmes de segmentation qui peuvent se retrouver en indexation de documents audio ou encore en transcription automatique de la parole.

Glossaire

4Hz ASD : 4Hz Amplitude and Standard Deviation

ANTS : Automatic News Transcription System

BIC : Bayesian Information Criterion

CIFRE : Convention Industrielle de Formation par la Recherche
CMR : Cepstral Mean Removal
CMS : Cepstral Mean Substraction

DARPA : Defense Advanced Research Projects Agency
DCT : Discrete Cosine Transform
DFB : Divergence Forward-Backward
DSP : Densité Spectrale de Puissance
DTW : Dynamic Time Warping

EM : Expectation Maximisation
ESTER : Evaluation des Systèmes de Transcription Enrichie d'Emissions Radiophoniques

FFT : Fast Fourier Transform
FIR : Finite Impulse Response

GMM : Gaussian Mixture Model

HMM : Hidden Markov Model
HTK : Hidden Markov Model Toolkit

IBM : International Business Machines corporation

kNN : k Nearest Neighbors
kppv : k plus proches voisins

LFCC : Linear Frequencies Cepstrum Coefficients
LFMAD : Low Frequency Modulation Amplitude and Deviation
LPC : Linear Predictive Coding

MAP : Maximum A Posteriori
MFCC : Mel Frequency Cepstral Coefficient
MLLR : Maximum Likelihood Linear Regression
MLP : Multi-Layer Perceptron
MSVM : Multi-Class Support Vector Machine

NIST : National Institute of Standards and Technology

PLP : Perceptual Linear Prediction
PMC : Perceptron Multi-Couches

RAIVES : Recherche Automatique d'Informations Verbales Et Sonores
RIF : Réponse Impulsionnelle Finie
RMS : Root Mean Square
RN : Réseau de Neurones

SMAP : Structural Maximum A Posteriori
SMLLR : Structural Maximum Likelihood Linear Regression
SVM : Support Vector Machine

TEO : Teager Energy Operator
TNS : Taylor Nelson Sofres

ZCR : Zero Crossing Rate

Bibliographie

[Ajmera 02] J. Ajmera, I. McCowan & H. Bourlard. *Robust HMM-Based Speech/Music Segmentation.* ICASSP-02, vol. 1, pages 297–300, 2002.

[Ajmera 03] J. Ajmera, I. McCowan & H. Bourlard. *Speech/Music Discrimination using Entropy and Dynamism Features in a HMM Classification Framework.* Speech Communication, vol. 40, pages 351–363, 2003.

[André-Obrecht 88] R. André-Obrecht. *A new statistical approach for automatic speech segmentation.* IEEE Transactions on Audio, Speech, and Signal Processing, vol. 36, 1988.

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

[Baum 72] L. Baum. *An inequality and associated maximization technic in statistical estimation for probabilistic functions of Markov processes.* Inequalities, vol. 3, pages 1–8, 1972.

[Bellman 61] R. Bellman. Adaptative control processes. Princeton University Press, 1961.

[BenAyed 03] Y. BenAyed. *Détection de mots clés dans un flux de parole.* PhD thesis, Ecole Nationale Supérieure des Télécommunications (Paris), 2003.

[Boser 92] B. Boser, I. Guyon & V. Vapnik. *A training algorithm for optimal margin classifiers.* In COLT'92, pages 144–152, 1992.

[Bourlard 90] H. Bourlard & C. Wellekens. *Links between Markov models and multilayer perceptrons.* In Trans. PAMI, volume 12, pages 1167–1178, 1990.

[Breebaart 03] J. Breebaart & M. McKinney. *Features for audio classification.* In International Symposium on Music Information Retrieval (ISMIR), 2003.

[Brun 04] A. Brun, C. Cerisara, D. Fohr, I. Illina, D. Langlois, O. Mella &

K. Smaïli. *ANTS : le système de transcription automatique du LORIA*. In Journées d'Etude sur la Parole - JEP'04, 2004.

[Caran 96] G. Caran & Y. Lechevallier. *Règles de décision de Bayes et méthodes statistiques de discrimination*. Revue d'intelligence artificielle, vol. 10, no. 2-3, pages 219–283, 1996.

[Carey 99] M.J. Carey, E.S. Parris & H. Lloyd-Thomas. *A Comparison of Features for Speech, Music Discrimination*. In ICASSP-99, 1999.

[Cho 98] Y.D. Cho, M.Y. Kim & S.R. Kim. *A spectrally mixed excitation (SMX) vocoder with robust parameter determination*. In ICASSP-98, pages 601–604, 1998.

[Chou 01] W. Chou & L. Gu. *Robust singing detection in speech/ music discriminator design*. In ICASSP-01, pages 865–868, 2001.

[Cornuéjols 02] A. Cornuéjols & L. Miclet. Apprentissage artificiel concepts et algorithmes. Eyrolles, 2002.

[Daubechies 88] I. Daubechies. *Orthonormal bases of compactly supported wavelets*. Communications on Pure and Applied Mathematics, vol. 41, no. 7, pages 909–996, 1988.

[Daubechies 92] I. Daubechies. Ten lectures on wavelets. Society for Industrial and Applied Mathematics, 1992.

[Davis 52] K.H. Davis, R. Biddulph & S. Balashek. *Automatic recognition of spoken digits*. Journal of the Acoustical Society of America, pages 637–642, 1952.

[Davis 80] S.B. Davis & P. Mermelstein. *Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences*. IEEE Trans. ASSP, vol. 28, pages 357–366, 1980.

[Delacourt 00] P. Delacourt. *La segmentation et le regroupement par locuteurs pour l'indexation de documents audio*. PhD thesis, Ecole Nationale Supérieure des Télécommunications, 2000.

[Dempster 77] A.P. Dempster, N.M. Laird & D.B. Rubin. *Maximum likelihood from incomplete data via the em algorithm (with discussion)*. Journal of the Royal Statistical Society, vol. 39, no. 1, pages 1–38, 1977.

[Devijver 82] P. Devijver & J. Kittler. Pattern recognition : a statistical approach. Prentice Hall, 1982.

[Deviren 03] M. Deviren & K. Daoudi. *Frequency filtering or wavelet filtering ?* In Joint Intl. Conf. on Artificial Neural Networks and Neural Information Processing (ICANN/ICONIP, 2003.

[Deviren 04] M. Deviren. *Revisiting speech recognition systems : dynamic Bayesian networks and new computational paradigms*. PhD thesis, Université Henri Poincaré, Nancy, France, 2004.

[Dimitriadis 05] D. Dimitriadis, P. Maragos & A. Potamianos. *Auditory Teager energy cepstrum coefficients for robust speech recognition.* In European Conference On Speech Communication and Technology, pages 3013–3016, 2005.

[Duda 73] R. Duda & P. Hart. Pattern classification and scene analysis. Wiley, 1973.

[Foote 95] J.T. Foote, G.J.F. Jones, K. Spärck Jones & S.J. Young. *Talker-independent keyword spotting for information retrieval.* In European Conference On Speech Communication and Technology, pages 2145–2148, 1995.

[Fredouille 04] C. Fredouille, D. Matrouf, G. Linares & P. Nocera. *Segmentation en macro-classes acoustiques d'émissions radiophoniques dans le cadre d'ESTER.* In JEP04, 2004.

[Gabor 46] D. Gabor. *Theory of communication.* J. IEEE, vol. 93, pages 429–457, 1946.

[Ganapathiraju 00] A. Ganapathiraju & J. Picone. *Hybrid SVM/HMM Architectures for Speech Recognition.* In Neural Information Processing Systems, 2000.

[Gauvain 98] J.L. Gauvain, L. Lamel & G. Adda. *The LIMSI 1997 Hub-4E Transcription System.* In Proc. DARPA Broadcast News Transcription & Understanding Workshop, pages 75–79, February 1998.

[Gauvain 99] J.L. Gauvain, L. Lamel, G. Adda & M. Jardino. *The LIMSI 1998 Hub-4E Transcription System.* In Proc. DARPA Broadcast News Transcription Workshop, pages 99–104, February 1999.

[Gauvain 02] J.L. Gauvain, L. Lamel & G. Adda. *The LIMSI Brodcast News Transcription System.* Speech Communication, vol. 37, no. 1–2, pages 89–108, 2002.

[Gelin 96] P. Gelin & C.J. Wellekens. *Keyword spotting for video soundtrack indexing.* In International Conference on Spoken Language Processing, pages 299–302, 1996.

[Gelin 97] P. Gelin & C.J. Wellekens. *Keyword spotting for multimedia document indexing.* In International Symposium and Education Program on Voice, Video, and Data Communications, 1997.

[Gemello 01] R. Gemello, D. Albesano, L. Moisa & R. De Mori. *Integration of Fixed and Multiple Resolution Analysis in a Speech Recognition System.* In ICASSP-01, 2001.

[Gerhard 02] D.B. Gerhard. *Perceptual features for a fuzzy speech-song classification.* In ICASSP-02, pages 4160–4163, 2002.

[Gorin 97] A.L. Gorin, G. Riccardi & J.H. Wright. *How May I Help You ?* Speech Communication, vol. 23, pages 113–127, 1997.

[Gravier 04] G. Gravier, J.F. Bonastre, E. Geoffrois, S. Galliano, K. Mc Tait & K. Choukri. *ESTER, une campagne d'évaluation des systèmes d'indexation automatique d'émissions radiophoniques en français.* In JEP04, 2004.

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

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

[Hopfield 87] J.J. Hopfield. *Learning algorithms ans probability distributions in fee-forward networks.* In Nat. Acad. Sci., pages 8429–8433, 1987.

[Hornik 89] K. Hornik, M. Stinchcombe & H. White. *Multilayer feedforward networks are universal approximators.* Neural Networks, vol. 2, 1989.

[Houtgast 85] T. Houtgast & J. M. Steeneken. *A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria.* Journal of the Acoustical Society of America, vol. 77, no. 3, pages 1069–1077, 1985.

[Huang 01] X. Huang, A. Acero & H. Hon. Spoken language processing : A guide to theory, algorithm and system development. Prentice Hall, 2001.

[Irino 93] T. Irino & H. Kawahara. *Signal reconstruction from modified auditory wavelet transform.* IEEE Transaction on Signal Processing, vol. 41, pages 3549–3554, 1993.

[ISIP] ISIP. *Le système ISIP,*
http ://www.ece.msstate.edu/research/isip/projects/speech/.

[Jabloun 99] F. Jabloun & A. Enis Cetin. *The Teager Energy based Feature Parameters for Robust Speech Recognition in Car Noise.* In ICASSP-99, 1999.

[Jelinek 76] F. Jelinek. *Continuous speech recognition by statistical models.* In IEEE Transaction on ASSP, volume 64, pages 532–566, 1976.

[Julius] Julius. *Le système Julius,*
http ://julius.sourceforge.jp/en/.

[Kaiser 90] J.F. Kaiser. *On a Simple Algorithm to Calculate the 'Energy'of a Signal.* In ICASSP-90, 1990.

[Karneback 01] S. Karneback. *Discrimination between Speech and Music based on a Low Frequency Modulation Feature.* In European Conf. on Speech Comm. and Technology, 2001.

[Karneback 02] S. Karneback. *Expanded Examination of a Low Frequency Modulation Feature for Speech/Music Discrimination.* In International Conference of Spoken Language Processing, pages 2009–2012, 2002.

[Karneback 04] S. Karneback. *Speech/Music Discrimination using discrete Hidden Markov Models.* TMH-QPSR, vol. 46, no. 1, pages 41–59, 2004.

[Kim 01] I. J. Kim, S. I. Yang & Y. Kwon. *Speech enhancement using adaptive wavelet shrinkage.* In ISIE-2001, volume 1, pages 501–504, 2001.

[LeCun 85] Y. LeCun. *Une procédure d'apprentissage pour réseau à seuil asymétrique.* In Proc. Cognitiva, pages 599–604, 1985.

[Lee 01] A. Lee, T. Kawahara & K. Shikano. *Julius - an open source real-time large vocabulary recognition engine.* In European Conference On Speech Communication and Technology (Eurospeech), pages 1691–1694, 2001.

[Lee 04] A. Lee, T. Kawahara & K. Shikano. *Real-time word confidence scoring using local posterior probabilities on tree trellis search.* In ICASSP-04, pages 793–796, 2004.

[Lin 05] C.-C. Lin, S.-H. Chen, T.-K. Truong & Y. Chang. *Audio Classification and Categorization Based on Wavelets and Support Vector Machine.* IEEE Transactions on Speech and Audio Processing, vol. 13, pages 644–651, 2005.

[Logan 00] B. Logan. *Mel Frequency Cepstral Coefficients for Music Modeling.* In International Symposium on Music Information Retrieval (ISMIR), 2000.

[Lu 01] L. Lu, H. Jiang & H. Zhang. *A Robust Audio Classification and Segmentation Method.* In ACM International Conference on Multimedia, 2001.

[Mallat 98] S. Mallat. A Wavelet Tour of Signal Processing. Academic Press, 1998.

[Mallat 00] S. Mallat. Une exploration des signaux en ondelettes. Editions de l'Ecole polytechnique, 2000.

[Mauclair 04] J. Mauclair & J. Pinquier. *Fusion de paramètres en classification Parole/Musique.* In JEP2004, 2004.

[Mesgarani 06] N. Mesgarani, M. Slaney & S.A. Shamma. *Discrimination of Speech from Nonspeech based on multiscale spectro-temporal modulations.* IEEE trans. on Audio, Speech, and Language Processing, vol. 14, no. 3, pages 920–930, 2006.

[Moore 92] B.C.J Moore. An introduction to the psychology of hearing. Academic Press, 1992.

[Nilson 71] N.J. Nilson. Problem-solving methods in artificial intelligence. McGraw-Hill, 1971.

[Peeters 03] G. Peeters. *A Large Set of Audio Features for Sound Description.* Rapport technique, IRCAM, 2003.

[Pinquier 02a] J. Pinquier, J-L. Rouas & R. André-Obrecht. *Robust speech/music classification in audio documents.* In International Conference on Spoken Language Processing, pages 2005–2008, 2002.

[Pinquier 02b] J. Pinquier, C. Senac & R. André-Obrecht. *Speech and music classification in audio documents.* In ICASSP-02, 2002.

[Pinquier 04a] J. Pinquier. *Indexation sonore : recherche de composantes primaires pour une structuration audiovisuelle.* PhD thesis, Université Paul Sabatier (Toulouse III), 2004.

[Pinquier 04b] J. Pinquier, J. Arias & R. André-Obrecht. *Audio classification by search of primary components.* In MIVARM'2004, 2004.

[Rabiner 78] L.R. Rabiner & R.W. Schafer. Digital processing of speech signals. Prenctice Hall, 1978.

[Rabiner 93] L.R. Rabiner & B.H. Juang. Fundamentals of speech recognition. Prenctice Hall, 1993.

[Razik 04] J. Razik, D. Fohr, O. Mella & N. Parlangeau-Vallès. *Segmentation Parole/Musique pour la transcription automatique.* In JEP04, 2004.

[Razik 07a] J. Razik, O. Mella, D. Fohr & J.P. Haton. *Frame-Synchronous And Local Confidence Measures For On-The-Fly Keyword Spotting.* In International Symposium on Signal Processing and its Applications - ISSPA 2007, 2007. 4 pages.

[Razik 07b] R. Razik. *Mesures de confiance trame-synchrones et locales en reconnaissance automatique de la parole.* PhD thesis, Université Henri Poincaré, Nancy, France, 2007.

[Riccardi 97] G. Riccardi, A. L. Gorin, A. Ljolje & M. Riley. *A spoken language system for automated routing.* In International Conference on Acoustics, Speech and Signal Processing, pages 1143–1146, 1997.

[Rose 95] R.C. Rose. *Keyword detection in conversational speech utterances using hidden Markov model based continuous speech recognition.* Computer, Speech and Language, vol. 9, pages 309–333, 1995.

[Rosenblatt 62] F. Rosenblatt. Principles of neurodynamics. Spartan Books, 1962.

[Rossignol 00] S. Rossignol. *Segmentation et Indexation des signaux sonores musicaux.* PhD thesis, Université Paris IV, IRCAM, 2000.

[Saha 00] S. Saha. *Image Compression from DCT to Wavelets : a Review.* ACM Crossroads, vol. 6, no. 3, pages 644–651, 2000.

[Sarikaya 00] R. Sarikaya & J.H.L. Hansen. *High Resolution Speech Feature Parameterization for Monophone-based Stressed Speech Recognition.* IEEE Signal Processing Letters, vol. 7, no. 7, pages 182–185, 2000.

[Saunders 96] J. Saunders. *Real-Time Discrimination of Broadcast Speech/Music.* In ICASSP-96, 1996.

[Scheirer 97] E. Scheirer & M. Slaney. *Construction and Evaluation of a Robust Multifeature Speech/Music Discriminator.* In ICASSP-97, 1997.

[Sphinx] Sphinx. *Le système Sphinx, http ://cmusphinx.sourceforge.net/html/cmusphinx.php.*

[Sukkar 96] R.A. Sukkar & C.H. Lee. *Vocabulary independent discriminative utterance verification for nonkeyword rejection in subword based speech recognition.* IEEE Transactions on Speech and Audio Processing, vol. 4, pages 420–429, 1996.

[T.Hain 98] T.Hain, S.E.Johnson, A.Tuerk, P.C. Woodland & S.J.Young. *Segment Generation and Clustering in the HTK Broadcast News Transcription System.* In Proc. 1998 DARPA Broadcast News Transcription and Understanding Workshop, pages 133–137, 1998.

[Tzanetakis 02] G. Tzanetakis & P. Cook. *Musical Genre Classification of Audio Signals.* IEEE Transaction on Speech and Audio Processing, vol. 10, no. 5, pages 293–302, 2002.

[Umapathy 05] K. Umapathy, S. Krishman & S. Jimaa. *Multigroup classification of audio signlas using time-frequency parameters.* IEEE Transaction on Multimedia, vol. 7, pages 308–315, 2005.

[Vapnik 82] V.N. Vapnik. Estimation of dependences based on empirical data. Springer-Verlag, N.Y, 1982.

[Vapnik 98] V.N. Vapnik. Statistical learning theory. John Wiley & Sons, Inc., N.Y., 1998.

[Viterbi 67] A. Viterbi. *Error bounds for convolutional codes and an asymptotically optimum decoding algorithm.* IEEE transaction on Information Theory, vol. 13, no. 2, pages 260–269, 1967.

[Wan 05a] V. Wan & J. Carmichael. *Polynomial dynamic time warping kernel support vector machines for dysarthric speech recognition with sparse training data.* In INTERSPEECH, 2005.

[Wan 05b] V. Wan & S. Renals. *Speaker verification using sequence discriminant support vector machines.* IEEE Transaction on Speech and Audio Processing, vol. 13, 2005.

[Wan 07] V. Wan. Kernel methods in bioengineering, signal and image processing, chapitre Building sequence kernels for speaker verification and word recognition. Idea Group Publishing, 2007.

[Wang 03] W.Q. Wang, W. Gao & D.W. Ying. *A fast and robust speech/music discrimination approach.* In ICIS-PCM 2003, pages 1325–1329, 2003.

[West 04] K. West & S. Cox. *Features and classifiers for the automatic classification of musical audio signals.* In ISMIR'04, pages 531–536, 2004.

[Williams 99] G. Williams & D. Ellis. *Speech/Music Discrimination based on posterior probability features.* In European Conference on Speech Communication and Technology, 1999.

[Wilpon 90] J.G. Wilpon, L.R. Rabiner, C. Lee & E.R. Goldman. *Automatic recognition of keywords in unconstrained speech using hidden Markov models.* IEEE Transactions on Acoustics, Speech and Signal Processing, vol. 38, pages 1870–1878, 1990.

[Woodland 98] P.C. Woodland, T. Hain, S.E. Johnson, T.R. Niesler, A. Tuerk & S.J. Young. *Experiments in Broadcast News Transcription.* In Proc. ICASSP'98, 1998.

[Young 95a] S.J. Young & al. The htk book. Entropic Ltd., Cambridge, England, 1995.

[Young 95b] SJ. Young & al. The HTK Book. Cambridge, England, Entropic Ltd., 1995.

[Zhan 99] P. Zhan, S. Wegmann & L. Gillick. *DRAGON Systems' 1998 Broadcast News Transcription System for Mandarin.* In DARPA Broadcast News Workshop, 1999.

[Zhang 98] T. Zhang & C.C.J. Kuo. *Hierarchical System for Content-Based Audio Classification and Retrieval.* In Conference on Multimedia Storage and Archiving Systems III, tome 3527 de SPIE, 1998.

Bibliographie personnelle

[Didiot 03] E. Didiot. *Conception et mise en oeuvre de M-SVM dédiées au traitement de séquences biologiques.* Rapport technique, INRIA, 2003.

[Didiot 06a] E. Didiot, I. Illina, O. Mella, D. Fohr & J.-P. Haton. *Speech/music discrimination based on wavelets for broadcast programs.* In International Conference on Signal Processing and Multimedia Applications, pages 151–156, Août 2006.

[Didiot 06b] E. Didiot, I. Illina, O. Mella, D. Fohr & J.-P. Haton. *Une nouvelle approche fondée sur les ondelettes pour la discrimination parole/musique.* In XXVIes Journées d'Etude sur la Parole, pages 209–212, Juin 2006.

[Didiot 06c] E. Didiot, I. Illina, O. Mella, D. Fohr & J.-P. Haton. *A wavelet-based parameterization for Speech/Music segmentation.* In International Conference on Spoken Language Processing, pages 653–656, Septembre 2006.

[Didiot 07] E. Didiot, I. Illina, O. Mella & D. Fohr. *A Wavelet-based Time-Frequency Parameterization for Speech/Music Discrimination.* Speech Communication, 2007. En révision.

Résumé

Dans cette thèse, nous étudions la segmentation d'un flux audio en parole, musique et parole sur musique (P/M). Cette étape est fondamentale pour toute application basée sur la transcription automatique de flux radiophoniques et plus généralement multimédias. L'application visée ici est un système de détection de mots clés dans les émissions radiophoniques. Les performances de ce système dépendront de la bonne segmentation du signal de parole fournie par le système de discrimination parole/musique. En effet, une mauvaise classification du signal audio peut provoquer des omissions de mots clés ou des fausses alarmes. Afin d'améliorer la discrimination parole/musique, nous proposons une nouvelle méthode de paramétrisation du signal. Nous utilisons une décomposition en ondelettes du signal qui permet une analyse des signaux non stationnaires dont la musique est un exemple. Nous calculons différentes énergies sur les coefficients d'ondelettes pour construire nos vecteurs de paramètres. Le signal est alors segmenté en quatre classes : parole (P), non-parole (NP), musique (M) et non-musique (NM) à l'aide de deux systèmes disjoints de classification HMM classe/non-classe. Cette architecture a été choisie car elle permet de trouver les meilleurs paramètres indépendamment pour chaque tâche P/NP et M/NM. Une fusion des sorties des classifieurs est alors effectuée pour obtenir la décision finale : parole, musique ou parole sur musique. Les résultats obtenus sur un corpus réel d'émissions de radio montrent que notre paramétrisation en ondelettes apporte une nette amélioration des performances en discrimination M/NM et P/M par rapport à la paramétrisation de référence fondée sur les coefficients cepstraux.

Mots-clés: discrimination parole/musique, ondelettes, paramètres statiques et dynamiques, paramètres à long-terme, fusion

Abstract

In this thesis, we study the segmentation of an audio stream in speech, music and speech on music (S/M). This is a fundamental step for all application based on automatic transcription of radiophonic stream and most commonly multimedia. The target application here is a keyword detection system in broadcast programs. The application performance depends on the quality of the signal segmentation given by the speech/music discrimination system. Indeed, bad signal classification can give miss-detections or false alarms. To improve the speech/music discrimination task, we propose a new signal parameterization method. We use the wavelet decomposition which allows an analysis of non-stationary signal like music for instance. We compute different energies on wavelet coefficients to construct our feature vectors. The signal is then segmented in four classes : speech (S), non-speech (NS), music (M) and non-music (NM), thanks to two apart class/non-class classification systems. These classification systems are based on HMM. We chose a class/non-class architecture because it allows to find independently the best parameters for each S/NS and P/NP tasks. A fusion of the classifier ouputs is then performed to obtain the final decision : speech, music or speech on music. The obtained results on a real broadcast program corpus show that our wavelet-based parameterization gives a significant improvement in performance in both M/NM and S/M discrimination tasks compared to the baseline parameterization using cepstral coefficients.

Keywords: speech/music discrimination, wavelets, static and dynamic parameters, long-term parameters (variance), classifiers fusion