Indexation sonore : recherche de composantes primaires pour une structuration audiovisuelle

Julien Pinquier

UNIVERSITE PAUL SABATIER
TOULOUSE III

Université Toulouse III **École Doctorale Informatique et Télécommunications**

Formation Doctorale Informatique de l'Image et du Langage

Indexation sonore : recherche de composantes primaires pour une structuration audiovisuelle

THÈSE

présentée et soutenue publiquement le 20 décembre 2004

pour l'obtention du

Doctorat de l'Université Toulouse III – Paul Sabatier

(spécialité Informatique)

par

Julien PINQUIER

Composition du jury

Rapporteurs :	M. Paul DELÉGLISE	Université du Maine
	M. Patrick GROS	IRISA Rennes
Examinateurs :	M. Daniel DOURS	Université Toulouse III
	M. Jean CARRIVE	Institut National de l'Audiovisuel
	M. Dominique FOHR	LORIA Nancy
Directeur de thèse :	Mme Régine ANDRÉ-OBRECHT	Université Toulouse III

Institut de Recherche en Informatique de Toulouse — UMR 5505

Mis en page avec la classe thloria.

Si vous ne pouvez expliquer un concept à un enfant de six ans, c'est que vous ne le comprenez pas complètement.

$Albert Einstein$

Remerciements

Je ne peux pas commencer mes remerciements sans évoquer ma directrice de thèse, Régine André-Obrecht, qui m'a donné le « virus » de la recherche !
Durant mes années de doctorat, elle m'a laissé prendre des responsabilités administratives, de recherche et d'enseignement et m'a donné les moyens dont j'avais besoin. Je lui en suis très reconnaissant.

Mes remerciements s'adressent également aux membres du jury, Daniel Dours, président, Patrick Gros, et Paul Deléglise, rapporteurs, Dominique Fohr et Jean Carrive, examinateurs. Ils ont accepté d'évaluer mon travail et m'ont prodigué de bons conseils que j'ai essayés de suivre.

Je remercie également Nathalie Vallès-Parlangeau et Philippe Daubias car leurs remarques ont permis d'améliorer la qualité de mon rapport.

Je remercie Martine et Jean-Luc pour leurs conseils qui m'ont été très utiles pour la préparation de ma soutenance.

Les travaux de ma thèse ont pu être appliqués dans le cadre du projet FERIA : je remercie notamment Brigitte, Claire-Hélène et Jean d'avoir rendu ce projet si passionnant.

Je remercie Agathe, Annie, Brigitte, Jean-Claude et Jean-Pierre de régler tous les tracas administratifs et techniques.

Un grand merci à l'ensemble des membres de l'équipe SAMOVA pour leur accueil et leurs conseils tout au long de cette thèse. Merci à Christine pour sa bonne humeur, Isabelle pour sa méthodologie de travail et Philippe pour sa disponibilité.

Je tiens à remercier Jérôme d'avoir partager son bureau avec moi. Son aide quasi quotidienne et ses invitations culinaires m'ont permis de tenir bon jusqu'à la fin !

Je remercie Xavier, mon camarade d'étude, qui me supporte depuis le DEUG !
C'est un collègue de sorties sportives, culturelles ou de jeux mais c'est surtout un grand ami.

Un grand merci à « Anatole », alias « Mister JP », « magouille.com », « JP Mc Donald ». Il n'est pas que le roi du bon plan c'est aussi un ami sur qui je peux compter.

Je n'oublie pas mes anciens collègues de DEA avec lesquels je passe encore de bons moments : Amélie, Bertrand, Farah, Fred, Hamid, Julie...

Pour arriver à être performant dans son travail, il faut être studieux mais aussi savoir profiter des pauses cafés, des repas, des soirées et des week-ends pour faire le vide. Avec des amis comme Bassem (l'espion de la zone café), Eric (Rac), Jean-Luc Roucas (le guide du routard US), le chef Jérôme, Jéromo, Jorge (Root beer), José (Tequila), Julie (la boute-en-train), Mathieu (coach Vahid), Momo (la anglaise), Samuel (la poche), Sara (la TV's star), Thomas (le responsable du salon de l'auto) et Véronique (le ch'ti parpaing), ce sont des moments de plaisir garanti !

Je remercie Jean-Claude, Siba, Gaël et Laurent pour leur bonne humeur quotidienne au travers des *****.

Dans un autre registre :

- je tiens à remercier l'ensemble des joueurs de la JSE pour les folles journées de football : AJY, Ben, Bibi, Canto, Custry, la Fadeur, Fent la bise, Friton, Jeannot, Nono, RV, Sac à vin, Sac à gnole, la Savate et le petit Guérin... Petites précisions : notre couleur, c'est le jaune et je ne signe pas les licences !
- je remercie également tous mes collègues pompiers du centre de secours d'Entraygues et notamment Christopher et Riri qui me remplacent toujours au pied levé.

Un petit coucou aux Millavois : Cathy, Del, Manu, Patricia, Richard, Sandra...

Depuis les folles années de lycée, certains n'ont pas changé et c'est très bien ainsi. Quel plaisir de se retrouver pour se fendre l'orteil avec Paulou, Marie, Grosse mule et Tête d'ail.

Merci à Colette et Claude de m'inviter si souvent chez eux. Je passe toujours un agréable moment surtout en parlant d'invitation au restaurant !

Je n'oublie pas la famille Marty qui me loue leur dernier boulet (à prix d'or ! !) et qui me supporte lorsque je rentre au pays.

Mes plus grands remerciements vont bien évidemment vers Claire qui me soutient depuis des années : je t'aime bien petit Nook...

Enfin, un grand merci à mes parents de m'avoir soutenu durant ces « quelques » années d'études !

PS : Salut à Béa « le Ratafia » (qui ne perd pas le Cap !) et aux Espeyracois : Pat, Chiassite et la famille Raynal (surtout Antoine et...).

À Nookie,

Table des matières

Partie II Les sons clés

Table des figures

Liste des tableaux

Introduction générale

1 Indexation

L'**AFNOR** (**A**ssociation **F**rançaise de **NOR**malisation) définit l'indexation comme une représentation de résultats d'une analyse en langue naturelle ou dans un langage normalisé d'un document. Une définition plus classique, non contradictoire, est l'identification et la localisation de séquences pertinentes ou de thèmes majeurs au sein d'un document par une analyse de son contenu.

L'objectif est, à l'aide de ses index, de classer ultérieurement le document parmi un ensemble de documents d'une collection donnée, d'extraire le contexte de cet index au sein du document lui-même. Ce type d'indexation a pour but l'optimisation de l'accès aux données dans de grandes bases.

À l'heure actuelle, les méthodes d'indexation en audio et vidéo sont principalement manuelles : un opérateur humain doit lire, écouter et/ou regarder le document numérique de façon à sélectionner les informations recherchées.

Or les possibilités technologiques offertes en matière de conservation, de communication, et d'accès aux données numériques incitent de nombreuses actions de production massive d'archives numériques d'images, de sons et de vidéos aussi bien dans le domaine de l'information et de l'audiovisuel (agences de presse, INA), que de la culture (musées), des transports (surveillance), de l'environnement (images satellitaires), ou de l'imagerie médicale (dossiers médicaux en milieux hospitaliers).

Compte tenu de l'accroissement gigantesque du volume de données à traiter, la tâche d'indexation devient extrêmement fastidieuse, et l'automatisation semble désormais indispensable. Afin de **synthétiser l'information pertinente**, naviguer ou rechercher efficacement dans les bases de données multimédia, des systèmes d'indexation doivent être développés pour pouvoir exploiter ces nouvelles technologies numériques. Ces systèmes pourront s'appuyer sur la norme MPEG7 qui permet de faciliter la recherche, le filtrage, l'organisation et la navigation sur de larges collections en faisant le lien entre l'extraction de paramètres et le moteur de recherche.

Par **analogie avec les documents textuels** qui sont faciles à manipuler (stockage, manipulation et recherche d'information étant devenus des opérations abordables par le grand public), le traitement des documents multimédia n'en est qu'à son balbutiement. Par exemple, trouver la vidéo contenant les premiers pas d'Armstrong sur la Lune (sans information a priori) est

pour l'instant assez critique si l'on ne traite que les documents multimédia. Il serait souhaitable, comme en indexation textuelle, que l'on puisse utiliser des moteurs de recherche via des mots clés. Cela nécessite d'extraire du sens de la vidéo et/ou de l'audio et de les utiliser conjointement.

2 Indexation sonore

Un document sonore, c'est-à-dire la bande sonore d'un document multimédia ou enregistrement d'émission radiophonique, est un document particulièrement difficile à indexer, car l'extraction de l'information élémentaire se heurte à l'extrême diversité des sources acoustiques. Les segments acoustiques sont de nature très diverses de par leur production et leur enregistrement : l'environnement peut être propre ou plus ou moins bruité, la qualité de l'enregistrement peut être plus ou moins soignée et liée à des éléments extérieurs (canal téléphonique), la musique peut être traditionnelle ou synthétique, la présence de parole peut être observée en monologue ou en dialogue...

Si aucune connaissance a priori n'est donnée et pour tenir compte de cette extrême variabilité, le signal acoustique doit subir un certain nombre de pré-traitements avant de pouvoir espérer extraire une quelconque information pertinente.

Il peut être intéressant de rechercher des « bruits » ou des sons sémantiquement significatifs tels que les applaudissements, les rires ou les effets spéciaux (pistolets, explosions...), de repérer les passages musicaux pour les segmenter et les identifier, de détecter les locuteurs équivalents à des tours de parole dans un dialogue. Enfin la transcription du discours ou la recherche de mots clés (mots isolés, groupes de mots...) fournissent une information importante sur le contenu du message verbal, et permettent l'accès à la recherche d'information telle qu'elle est pratiquée dans des documents textuels.

Si l'on se réfère à la norme MPEG7, indexer un document sonore signifie rechercher aussi bien des composantes de bas niveau dites primaires comme la parole, la musique, les sons clés (jingles, mots-clés...) que des descripteurs de plus haut niveau tels les locuteurs ou les thèmes.

3 Problématique

Les problématiques d'extraction d'information de documents audiovisuels (cf. figure 1) sont au cœur des préoccupations de notre équipe de recherche **SAMoVA** (**S**tructuration, **A**nalyse et **Mo**délisation de documents **V**idéo et **A**udio). Les travaux s'appuient sur l'expérience acquise en traitement automatique de la parole et plus particulièrement en reconnaissance automatique de la parole et de la langue, où la reconnaissance des formes par approche statistique est privilégiée.

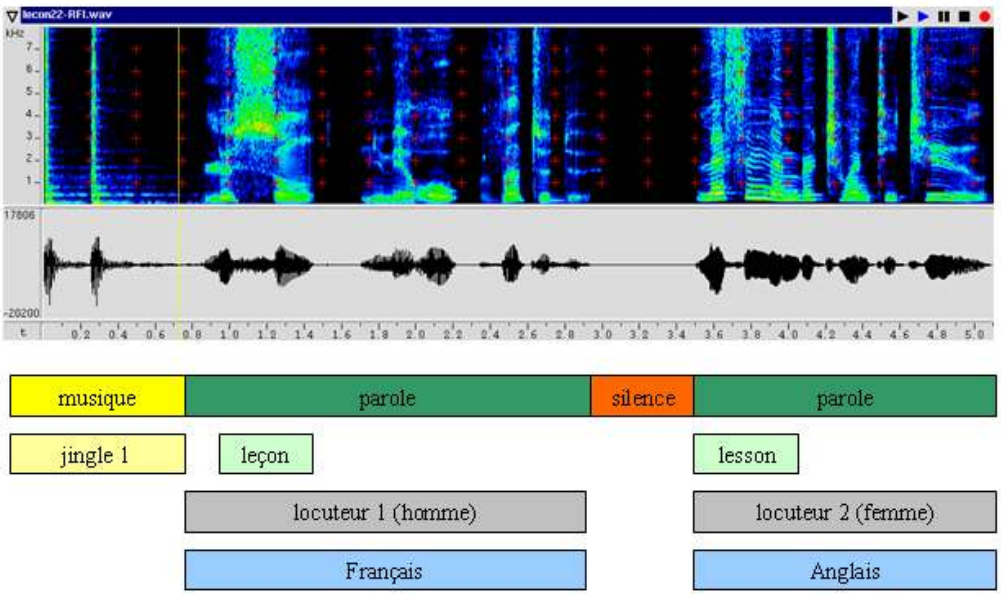

FIG. 1 – Exemple d'indexation sonore sur un document radiophonique de RFI (Radio France Internationale).

Mais dès lors que l'on s'intéresse au problème d'indexation du flux sonore, il est nécessaire de comprendre en quoi une tâche d'indexation diffère d'une tâche de reconnaissance de formes ou de classification automatique ; est-ce un « simple » problème de robustesse des outils classiques compte tenu de la très grande variabilité rencontrée dans une collection de documents outils ? Est-ce un problème de mise en œuvre des algorithmes ou une remise en cause ?

C'est dans cet état d'esprit que nous nous sommes placés afin de relever les **deux défis** suivants. Il s'agit tout d'abord d'une **contribution à l'analyse automatique sonore** grâce à des outils de recherche de composantes primaires. Ces composantes primaires sont des briques de base qui doivent être robustes, utiles et utilisables pour nous permettre d'effectuer ensuite une **structuration automatique de documents audiovisuels**.

Nous avons cherché à développer un système d'indexation. Comme nombre de ces systèmes, nous avons étudié une segmentation et une identification des différentes composantes du flux sonore, avant d'effectuer sur chaque composante des traitements plus spécialisés. À chaque étape, nous avons cherché à définir l'outil le plus adapté au problème de l'indexation en terme de mise en œuvre : nous pouvons évalué par exemple la robustesse des paramètres et la rapidité de mise en œuvre (exécution, complexité).

Cette construction progressive de l'indexation d'un document nous a également amené de manière duale à réfléchir sur la structuration automatique du document.

De manière plus précise, les **objectifs** du travail réalisé tout au long de cette thèse sont les suivants :

- concevoir un système capable de définir la notion de décomposition Parole/Musique/Bruit (PMB) par l'intermédiaire de détecteurs performants et robustes,
- proposer une alternative à ce premier partitionnement en détectant des sons-clés représentant le début et/ou la fin d'une émission afin de segmenter ou de structurer le flux audiovisuel,
- adapter les outils de reconnaissance de locuteur et de parole à la recherche de personnages cibles ou de mots cibles, index naturels d'un flux de parole,
- réfléchir à une fusion des détecteurs audio et vidéo.

Ce travail de recherche s'est développé dans le cadre du projet **RNRT**[1] (**R**éseau **N**ational de **R**echerche en **T**élécommunications) **AGIR**[2] (**A**rchitecture **G**lobale pour l'**I**ndexation et la **R**echerche d'informations par le contenu) et du projet **RIAM**[3] (**R**echerche et **I**nnovation en **A**udiovisuel et **M**ultimédia) **FERIA**[4] (**F**ramework pour l'**E**xpérimentation et la **R**éalisation **I**ndustrielle d'**A**pplications multimédias).

4 Organisation du mémoire

La description de nos différents travaux d'indexation par analyse de la composante sonore se décompose en deux parties principales, l'une est liée à la décomposition en composantes primaires du flux sonore, à savoir la parole, la musique et les autres sources sonores dénommées communément bruit. La seconde traite de l'extraction de constituants relativement courts, ce

[1] http ://www.telecom.gouv.fr/rnrt/
[2] http ://www.ina.fr/recherche/projets/agir
[3] http ://www.riam.org/
[4] http ://www.ina.fr/recherche/projets/encours/feria/

que nous appelons les sons clés ; cela va de la recherche de jingles à celle de mots clés en passant par la recherche d'applaudissements ou par la localisation d'intervenants cibles. Au cours d'une troisième partie, la synthèse des informations extraites permet de proposer quelques axes de structuration de documents audiovisuels.

La **première partie** est consacrée à la classification parole/musique/bruit (PMB).

Après un premier chapitre rappelant définitions et état de l'art en décomposition PMB, nous décrivons au cours du second chapitre un premier système de détection de parole et de musique, dit système de base. Bien que ce système utilise des outils de paramétrisation (spectre, cepstre) et de modélisation (mélanges de lois gaussiennes) très classiques, leur mise en œuvre permet de décorréler les deux détections et d'utiliser ainsi les outils les plus adéquats. L'utilisation des techniques classiques permet de certifier que les performances obtenues soient au niveau de l'état de l'art.

L'inconvénient de ce type de système est sa forte dépendance vis à vis des conditions d'enregistrements. Ce manque de robustesse ne permet pas une utilisation aveugle sur tout type de document. C'est pourquoi nous proposons un nouveau système de classification parole/musique tout à fait original. Sa description et sa mise en œuvre sont détaillées au chapitre trois. Ce système est fondé sur l'extraction de trois paramètres originaux : la modulation de l'entropie, la durée des segments (issue d'une segmentation automatique) et le nombre de ces segments par seconde. Les informations issues de ces trois paramètres sont ensuite fusionnées avec celle issue de la modulation de l'énergie à 4 Hz. De nombreuses expériences sont réalisées afin d'évaluer la robustesse du système.

La **deuxième partie** présente différentes méthodes de détection de sons clés.

Le chapitre quatre propose un système de détection et d'identification de jingles. Fondé sur un simple calcul de distance euclidienne dans le domaine spectral, ce système se révèle très performant et très intéressant dans le cadre d'une structuration élémentaire de flux de type télévisuel.

Lors du cinquième chapitre, deux sons clés caractéristiques d'émissions dites de plateau sont analysés et recherchés : les applaudissements et les rires sont détectés et localisés à l'aide d'une méthode comparable à celle employée dans le système de base PMB. Les performances montrent les limites de ces outils lorsque la modélisation statistique n'est pas adaptée : le rire se montre difficilement maîtrisable !

Avec le problème de détection de mots clés développé au sixième chapitre, nous reprenons un problème classique, vieux et très difficile. Nombre de systèmes d'indexation extraient les mots clés après avoir transcrit automatiquement la totalité de la composante parole. Dans une

approche volontairement dynamique où la liste des mots clés est définie par l'utilisateur, nous reprenons l'approche dite des modèles « poubelles » pour cerner ses limites, mais aussi l'exploiter dans le cadre d'une recherche de thèmes. Bien que cette étude soit préliminaire, les conclusions nous semblent intéressantes en structuration.

Durant la troisième et **dernière partie**, il nous paraît nécessaire d'examiner des scénarios au cours desquels il est indispensable d'enchaîner les différents systèmes développés tout au long de cette thèse. Nous nous attardons sur la structuration d'une émission télévisuelle de plateau, le « Grand Échiquier », et d'un journal télévisé, le « 6 Minutes ». Ce travail nous conduit à une réflexion sur l'apport de la vidéo dans cette analyse essentiellement audio et sur la fusion audiovisuelle en général. De nombreux axes de réflexion sont proposés et permettent de conclure ce document sur de nombreuses et réalistes perspectives.

Afin de valider les différentes approches abordées dans cette thèse, 75 heures de données sont traitées : des journaux d'informations télévisés et radiophoniques, des reportages, des séries, des publicités, des interviews, des émissions de plateau et des chansons.

Première partie

Classification Parole/Musique/Bruit

Chapitre 1

Définitions et état de l'art

Sommaire

1.1 Introduction

Un document sonore est l'enregistrement d'un signal acoustique obtenu à partir de plusieurs sources de production sonore. Il est donc constitué de nombreuses composantes, dont les plus communes sont la parole et la musique. Ces deux composantes sont dites primaires, et il faut leur rajouter les composantes liées aux multiples sources de bruit potentielles. Avant d'aller plus loin, il convient de préciser certaines définitions, notamment sur la parole et la musique afin de lever toute ambiguïté.

1.1.1 Parole

Le signal de parole appartient à la classe des signaux acoustiques produits par des vibrations des couches d'air. Les variations de ce signal reflètent les fluctuations de la pression de l'air [Boi87].

La parole est une suite de sons produits soit par des vibrations des cordes vocales (source quasi périodique de voisement), soit par une turbulence créée par l'air s'écoulant dans le conduit vocal, lors du relâchement d'une occlusion ou d'une forte constriction de ce conduit (sources de bruit non voisées) [Cal89]. La durée d'un son est de l'ordre de 60 à 100 ms (cf. spectrogramme de la figure 1.1).

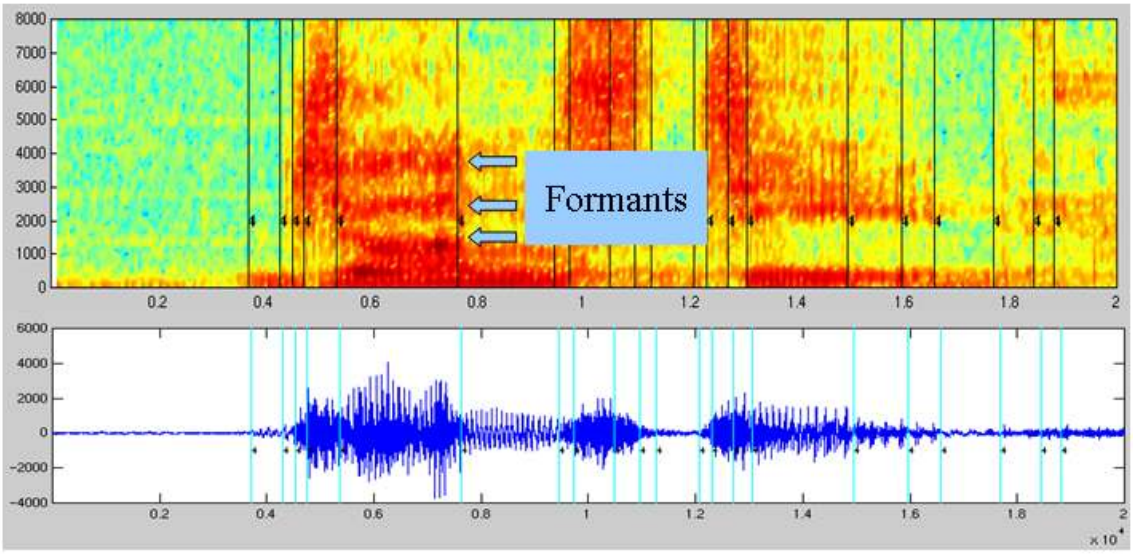

FIG. 1.1 – Spectrogramme de parole : alternance de sons voisés et non voisés sur 1,2 s de signal. Dans le cas voisé, une structure formantique est présente.

La parole est formée de phonèmes et de transitions entre ces phonèmes. Plusieurs types de phonèmes existent : les voyelles, les consonnes fricatives, les consonnes plosives, les nasales et les liquides. Les transitions acoustiques correspondent à des transitions dans l'appareil de production de l'état correspondant au premier phonème à l'état correspondant au suivant [Kor99].

Les voyelles, sons voisés par excellence, sont les « piliers » de la parole ; leur présence est révélée fréquentiellement par les formants qui correspondent aux fréquences de résonance du conduit vocal (cf. figure 1.1). La fréquence d'apparition des voyelles correspond au rythme syllabique.

1.1.2 Musique

Les particularités de la musique, qui la différencient de toutes autres sonorités, ne résident pas seulement dans des différences culturelles, mais dans des propriétés physiologiques très spécifiques du système auditif de l'homme. Ainsi, définir la musique est très difficile car celle-ci peut être produite et perçue de différentes manières.

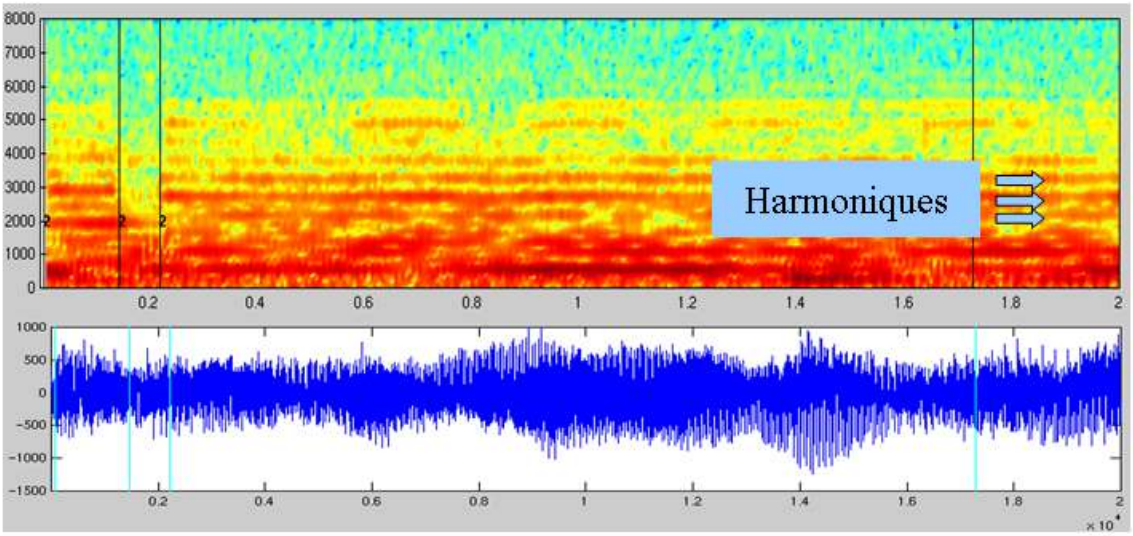

FIG. 1.2 – Spectrogramme de musique traditionnelle sur 1,2 s de signal : une structure harmonique est présente.

C'est pourquoi lorsqu'il s'agit d'extraire cette composante, de nombreux chercheurs se limitent à l'étude de la musique « instrumentale traditionnelle » dans le sens où elle est une composition de sons harmoniques (de notes au sens classique).

Dans un cadre éventuellement polyphonique, le signal acoustique musical se caractérise par l'existence de(s) hauteur(s) ou fréquence(s) fondamentale(s) (cf. figure 1.2).

Remarque : lorsque l'on parle d'un signal harmonique, il s'agit d'un signal composé d'une somme de sinusoïdes disposées sur un peigne harmonique.

1.2 Paramètres (ou caractéristiques)

Pour rechercher des composantes primaires dans un document sonore telles que la parole et la musique décrites auparavant, deux étapes sont indispensables :

- la paramétrisation,
- la décision.

Une telle analyse correspond à une analyse de type reconnaissance des formes. La phase de paramétrisation a pour but l'extraction d'informations pertinentes, dites discriminantes pour la tâche de classification envisagée.

Beaucoup de caractéristiques sont utilisées dans les systèmes actuels, nombre d'entre elles visent à mettre en évidence l'aspect harmonique du signal. Seules les plus fréquemment utilisées sont reprises ici. Elles ont été classées en quatre groupes selon leur mode de calcul :

- les paramètres temporels,
- les paramètres fréquentiels,
- les paramètres mixtes,
- les paramètres issus de modélisation.

1.2.1 Les paramètres temporels

Les deux principaux paramètres temporels sont l'énergie et le ZCR (Zero Crossing Rate). Ils sont en général directement calculés à partir du signal temporel. Utilisés il y a très longtemps en reconnaissance de la parole [Dav52], ils ont prouvé plus récemment leur pouvoir discriminant dans le cadre de ce problème.

1.2.1.1 Le ZCR

Le ZCR est le taux de passage par zéro. Cette caractéristique est fréquemment utilisée pour la classification parole/musique [Sau96], [Sch97] et [Zha98]. Les brusques variations du ZCR sont significatives de l'alternance voisée/non-voisée donc de présence de parole.

La trame acoustique est une suite d'échantillons représentant 20 à 40 ms de signal en général, durant laquelle le signal de parole est supposé quasi stationnaire : des paramètres statistiques peuvent y être calculés.

Le ZCR d'une trame est déduit du nombre de fois où le signal sonore change de signe :

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

avec $x_n(i)$ le nième échantillon de la trame i et N le nombre d'échantillons dans la trame i.

Pour la parole, le ZCR est faible pour les zones voisées et très élevé pour les zones non voisées alors que pour la musique, les variations du ZCR sont très faibles (cf. figure 1.3).

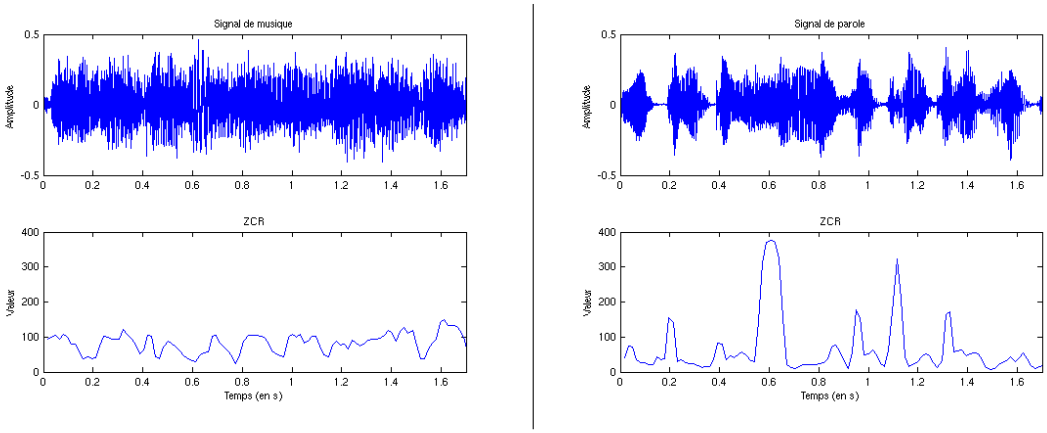

FIG. 1.3 – ZCR d'un signal pour des extraits de musique et de parole sur 1,7 s.

[Lu01] a montré que la variation du ZCR est plus significative que sa valeur exacte. Il en a déduit la mesure « haute proportion du ZCR » (HZCRR) qui correspond au nombre de trames où le ZCR est supérieur à 1,5 fois la valeur moyenne du ZCR, par unité de temps.

$$HZCRR = \frac{1}{2I}\left(\sum_{i=1}^{I}(\text{sign}(ZCR(i) - 1.5 * avZCR) + 1)\right) \tag{1.2}$$

avec I est le nombre total de trames et :

$$avZCR = \frac{1}{I}\sum_{i=1}^{I} ZCR(i) \tag{1.3}$$

1.2.1.2 L'énergie

L'énergie est un paramètre couramment utilisé en traitement du signal. L'énergie d'un signal échantillonné $(x_n(i))_{n=1,\cdots,N}$ à support fini est définie par :

$$E(i) = \sum_{n=1}^{N} x_n^2(i) \tag{1.4}$$

Étant donnée sa dynamique et pour respecter l'échelle perceptive, elle est généralement exprimée en décibels :

$$E_{db}(i) = 10 \times \log_{10}\left(\sum_{n=1}^{N} x_n^2(i)\right) \tag{1.5}$$

Pour un signal échantillonné de longueur infinie, on calcule l'énergie à court terme en prenant des portions de signal relatives à une fenêtre glissante. Cette fenêtre est étroite, de l'ordre de 10 ms, et correspond en général à une trame acoustique.

Pour éliminer la variabilité de ce paramètre, due en partie à des conditions d'enregistrements différentes (une simple variation de la distance entre la source et le microphone suffit pour être élément de perturbation de l'énergie), l'énergie peut être normalisée par rapport au maximum observé sur le signal global.

L'énergie à court terme varie beaucoup pour la parole alors qu'elle est plutôt stable pour la musique. Elle peut permettre, comme le ZCR, de discriminer la parole de la musique (cf. figure 1.4) mais aussi de détecter les silences.

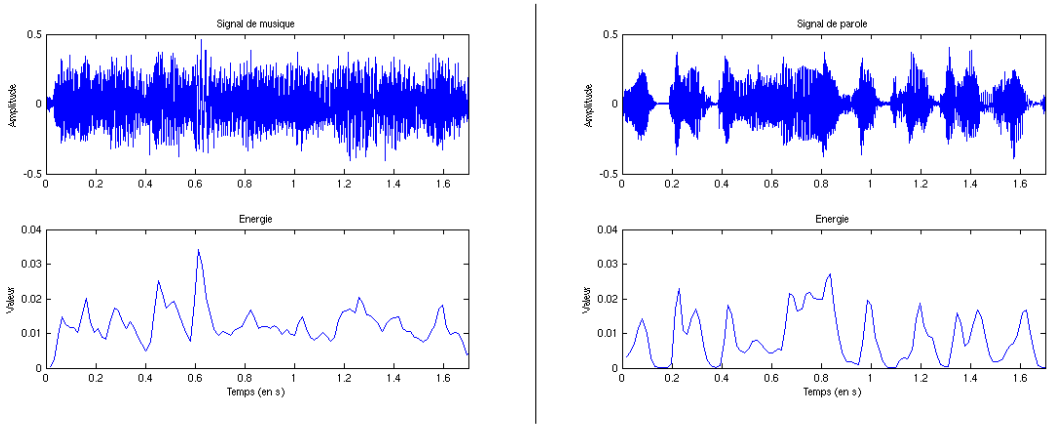

FIG. 1.4 – Évolution de l'énergie pour la musique et pour la parole sur 1,7 s.

Le caractère complémentaire de l'énergie et du ZCR implique que leur couplage permet une bonne discrimination des sons voisés et non voisés : un taux de passage par zéro faible et une énergie forte sont synonymes d'un son voisé alors qu'un taux de passage par zéro élevé et une énergie plus faible caractérisent une zone non voisée.

1.2.2 Les paramètres fréquentiels

Ces paramètres sont issus de la **DSP** (Densité Spectrale de Puissance). La DSP d'un signal est la transformée de Fourier de la fonction d'autocorrélation. Les trois principaux paramètres, au sens de la discrimination parole/musique, sont le centroïde spectral, le flux spectral et le « spectral rolloff point » [Sau96] et [Sch97].

1.2.2.1 Le centroïde spectral

Le centroïde spectral est le centre de gravité fréquentiel d'une DSP.

$$C(i) = \frac{\sum_{n=1}^{N} w_n . S_i(w_n)}{\sum_{n=1}^{N} S_i(w_n)} \tag{1.6}$$

où $S_i(w_n)$ correspond à la composante spectrale de la trame i à la fréquence w_n et N est le nombre d'échantillons dans la trame i.

Il est plus élevé pour la musique car les hauteurs des sons sont réparties fréquentiellement sur une zone plus importante pour la musique que celle pour la parole : en général 6 octaves sont nécessaires pour décrire la musique et 3 suffisent pour la parole (cf. figure 1.5). La variation du centroïde spectral est aussi significative : une variation importante caractérise l'alternance voisée/non-voisée.

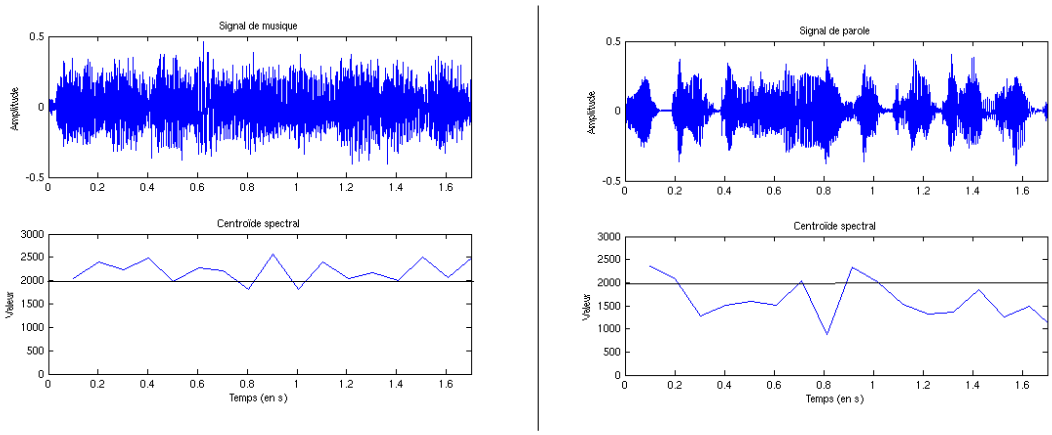

FIG. 1.5 – Centroïde spectral pour la musique et la parole sur 1,7 s.

1.2.2.2 Le flux spectral

Le flux spectral mesure la variation de spectre de deux analyses consécutives sachant que l'intervalle de temps entre ces deux analyses est de l'ordre de 20 ms.

$$FS = \sum_{n=1}^{N} \left(\frac{S_i(w_n)}{||S_i||} - \frac{S_{i-1}(w_n)}{||S_{i-1}||} \right)^2 \tag{1.7}$$

Pour la parole, les variations sont importantes et la valeur du flux spectral peut être faible alors que pour la musique le profil est moins agité et le niveau est élevé [Lu01]. Un flux spectral élevé caractérise un contenu musical et une variance élevée reflète la présence de parole (cf. figure 1.6).

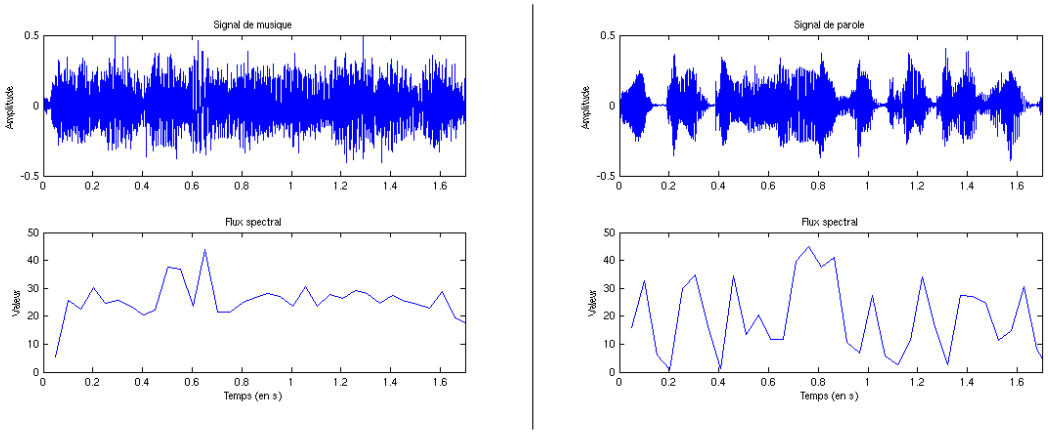

FIG. 1.6 – Flux spectral pour la musique et la parole sur une 1,7 s.

1.2.2.3 Le spectral rolloff point

Le « spectral rolloff point » est la fréquence de coupure en dessous de laquelle 95% de la puissance de la DSP est concentrée (cf. figure 1.7).

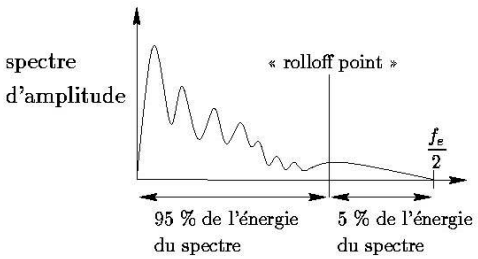

FIG. 1.7 – Définition du « Spectral Rolloff Point ».

Le « spectral rolloff point » est plus élevé pour un son non voisé (son riche en hautes fréquences) que pour un son voisé (énergie concentrée dans des fréquences plus faibles). Cette mesure permet donc de caractériser les alternances voisées/non-voisées de la parole. Pour la musique, cette mesure est sensiblement toujours la même (cf. figure 1.8).

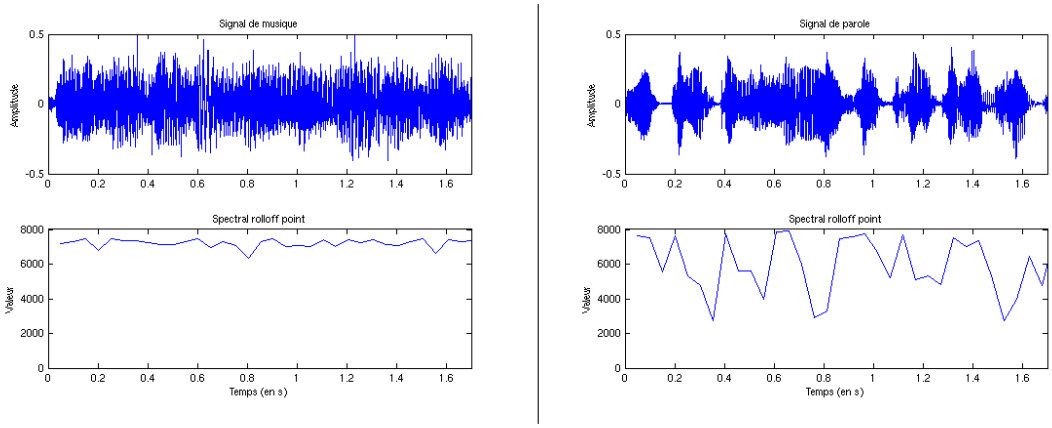

FIG. 1.8 – « Spectral Rolloff Point » correspondant à 1,7 s de musique et de parole.

1.2.3 Les paramètres mixtes

La **modulation de l'énergie à 4 Hertz** est un exemple de paramètre mixte, c'est-à-dire issu à la fois d'analyses fréquentielle et temporelle [Sch97]. Il est intéressant d'observer le comportement de l'énergie et sa modulation autour de 4 Hertz.

Pour la parole, les changements de syllabe se situent aux alentours de cette fréquence, sous l'hypothèse qu'une syllabe soit la combinaison d'une zone de faible énergie (consonne) et d'une zone de forte énergie (voyelle).

La musique possède une variation de l'énergie à 4 Hertz beaucoup plus faible que la parole (cf. figure 1.9).

Remarque : le calcul de ce paramètre est explicité dans le paragraphe 3.2.2, car il sera largement utilisé dans nos travaux.

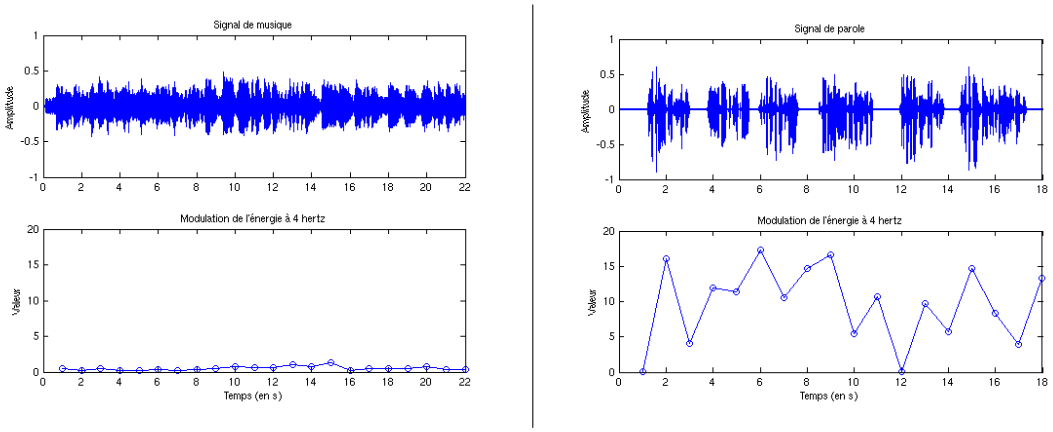

FIG. 1.9 – Modulation de l'énergie à 4 Hertz pour la musique (extrait de Mozart de 22 s) et la parole (6 phrases de parole lue de 18 s).

1.2.4 Les paramètres issus de modélisation : les MFCC

Les MFCC sont utilisés en reconnaissance de parole et en identification du locuteur ou de la langue car ces paramètres sont bien adaptés au signal de parole [Mar02]. Ils sont issus de l'hypothèse suivante, à savoir que le signal de parole est le résultat de la convolution entre un filtre (conduit vocal) et une excitation (cordes vocales) :

$$x_n = g_n * b_n \tag{1.8}$$

avec x_n le signal, g_n l'entrée et b_n le filtre caractérisant le conduit.

Une transformation homomorphique permet de transformer ce produit en une somme qui est ensuite filtrée pour obtenir les **MFCC** : « Mel Frequency Cepstral Coefficient ». Ces MFCC permettent une déconvolution entre la source des sons produits (caractéristiques du locuteur) et le conduit oral (couplé ou non au conduit nasal) :

$$\widetilde{x_n} = \widetilde{g_n} + \widetilde{b_n} \tag{1.9}$$

La transformation homomorphique se décompose en trois étapes principales [Cal89] (cf. figure 1.10) :

- un passage dans le domaine spectral par calcul du module de la transformée de Fourier rapide,
- une application du logarithme,
- un retour au domaine temporel par calcul de la transformée de Fourier rapide inverse.

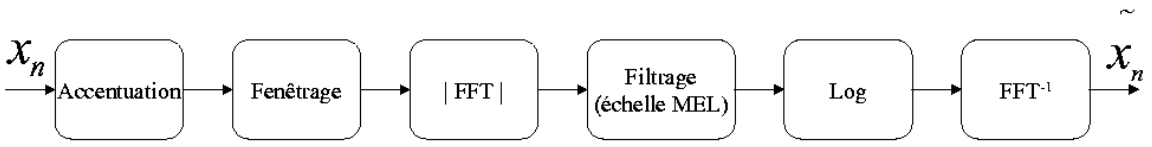

FIG. 1.10 – Processus de création des coefficients cepstraux.

où
FFT : Transformée de Fourier Rapide (passage dans le domaine spectral).
FFT^{-1} : Transformée de Fourier Rapide Inverse (retour dans le domaine temporel).
Le calcul de la FFT se fait sur des fenêtres glissantes.

Une *accentuation des aigus* est présente car les composantes fréquentielles aiguës sont toujours plus faibles que les graves. Un filtrage de type passe-haut est réalisé avec la fonction de transfert :

$$H(z) = 1 - 0.98 * z^{-1} \tag{1.10}$$

L'utilisation d'une *fenêtre de Hamming* (sur une trame acoustique de 256 ou 512 points en général) avec recouvrement sur la moitié (128 ou 256 points) permet d'éviter la formation d'artefacts liés aux effets de bord durant la transformation du domaine temporel au domaine fréquentiel :

$$W_{Hamming}(n) = \left\{ \begin{array}{c|c} 0.54 - 0.46 * cos(\frac{2\pi n}{N}) & pour\ 0 \leq n \leq N-1 \\ 0 & ailleurs \end{array} \right\} \tag{1.11}$$

avec N la taille de la fenêtre.

L'échelle non linéaire Mel est connue pour rendre compte de la perception humaine. Les coefficients sont appelés MFCC car dans le domaine spectral, ce changement d'échelle (utilisation de l'échelle perceptive Mel) est effectué. Ils ont la propriété d'être fortement décorrélés. Dans les systèmes de reconnaissance de la parole, le premier coefficient est souvent utilisé pour définir l'énergie.

Généralement une soustraction cepstrale se fait sur les MFCC pour déconvoluer le signal du bruit du canal (de la source d'enregistrement : micro, canal téléphonique...) et obtenir un signal paramétré débruité [Mok95]. Cette opération résulte du fait que les coefficients cepstraux de la parole ont une moyenne nulle ; pour ôter le bruit causé par le canal, il suffit alors de soustraire à chaque coefficient cepstral du signal bruité leur moyenne, représentative de la moyenne des coefficients cepstraux relatifs au bruit seul.

L'inconvénient majeur de la représentation cepstrale réside dans son manque de lisibilité : il ne s'agit pas d'une représentation directement liée aux informations qu'un expert peut extraire de la lecture d'un sonagramme, ce qui complexifie l'interprétation des paramètres.

Remarque : cette méthode est la plus utilisée et la plus performante à l'heure actuelle en traitement automatique de la parole, que ce soit pour faire de l'identification automatique des langues, de la reconnaissance du locuteur ou pour faire de la classification Parole/Musique/Bruit ([Gau99]).

1.2.5 Mais encore...

Il existe d'autres paramètres, notamment :

- le **pulse métric** ou détection de pulsation [Sch97] caractérise le rythme très présent et régulier de la musique (paramètre temporel).

- la **fréquence fondamentale**, notée F_0, correspond à la fréquence de vibration des cordes vocales ou la hauteur de la note jouée.
  Les algorithmes d'extraction de F_0 utilisent généralement une représentation temporelle ou spectrale du signal. Les méthodes temporelles utilisent la similarité du signal d'une période à l'autre pour identifier la période fondamentale. Il est parfois possible de repérer manuellement la période fondamentale T_0 (telle que $F_0 = 1/T_0$) directement sur le signal. Dans le domaine fréquentiel, les algorithmes cherchent les harmoniques de la fréquence fondamentale.

  Comme le dit [Zha98], la fréquence fondamentale renseigne sur le type de son étudié. Elle permet seulement une discrimination par comparaison à une référence issue du même processus.

– l'**harmonicité** [Wol99] permet d'accorder une confiance plus ou moins grande à la fréquence fondamentale ou de classer les sons.

– la **largeur de bande** [Wol99] présente l'étalement en fréquences d'un spectre.

– le **timbre** permet de faire la différence entre deux sons de même hauteur, puissance et durée [Wol99] et [Zha98] (paramètres fréquentiels).

– le nombre de segments de basse énergie [Sch97] (paramètres mixtes).

Ces critères sont plus longuement explicités dans [Car99] et [Car00b].

1.2.6 Conclusion

Il existe une grande variété de paramétrisations possibles afin de discriminer la parole de la musique. Malheureusement cette première étape, commune à tout système d'extraction de la parole, se limite trop souvent aux calculs des coefficients cepstraux (MFCC). Le pré-traitement mérite de s'y attarder un peu plus car il influe notablement sur les résultats obtenus alors qu'actuellement les efforts se portent plus sur les méthodes de classification (modélisation).

1.3 Méthodes de Classification

Les caractéristiques extraites permettent de définir un vecteur d'observation pour chaque trame de signal analysée. Ces observations sont ensuite comparées à des références ou évaluées à l'aide de modèles représentant les classes pour permettre d'associer une classe à chacune d'elles.

Il est commun de distinguer deux types de stratégies de classification, selon la connaissance a priori dont on dispose sur les observations et les classes elles-mêmes :

- l'approche statistique est privilégiée dès lors que les distributions de ses vecteurs d'observations, supposées être les réalisations de vecteurs aléatoires, sont connues ou estimables, ainsi que les probabilités a priori des classes.
- l'approche discriminante au travers de la recherche de frontières de régions est la méthode alternative. Les fonctions à seuil et les SVM (Support Vector Machines) appartiennent à ce type de stratégie.

Quelle que soit la méthode, ces méthodes nécessitent un apprentissage « supervisé » pour soit définir les références, soit estimer les modèles probabilistes, soit apprendre les paramètres caractérisant les zones de classification.

1.3.1 Approche statistique

L'approche statistique suppose de probabiliser l'espace des observations et l'espace des classes [Dud01].

Chaque classe C_i (la classe musique, la classe parole par exemple) est supposée être caractérisée par la loi a priori $P(C_i)$ et par la distribution des observations conditionnellement à la classe $P(y|C_i)$, où y est un vecteur d'observations ou une suite de vecteurs d'observations.

La décision d'appartenance de y à une classe est prise en recherchant la classe de probabilité a posteriori maximum :

$$P(C_i|y) = \max_j P(C_j|y) = \max_j P(y|C_j)P(C_j)/P(y) \tag{1.12}$$

En supposant toutes les classes équiprobables, le critère de décision devient alors le critère du maximum de vraisemblance et la classe identifiée C_i vérifie :

$$P(y|C_i) = \max_j P(y|C_j) \tag{1.13}$$

Cette approche nécessite une phase d'apprentissage pour estimer $P(C_i)$ et $P(y|C_i)$ à partir d'un ensemble d'échantillons.

Dans le cas où l'espace des observations est continu (de type R^d), différentes méthodes sont envisageables pour estimer l'ensemble de ces probabilités selon les connaissances ou hypothèses faites sur les lois :

– les méthodes paramétriques,
– les méthodes non paramétriques.

Dans les deux cas, les probabilités conditionnelles sont supposées admettre une densité de probabilité.

1.3.1.1 Méthodes paramétriques

La forme de la densité de probabilité $P(y|C_i)$ est supposée connue a priori et être entièrement décrite par un petit nombre de paramètres (loi normale, loi de Poisson, loi gamma...). L'apprentissage à l'aide des échantillons consiste à estimer ces paramètres.

Exemple 1 : le mélange de lois gaussiennes.

Les lois les plus couramment utilisées dans R^d sont les lois gaussiennes multidimensionnelles [Sau96] et les mélanges de lois gaussiennes multidimensionnelles (**MMG** pour **M**odèle de **M**élange de lois **G**aussiennes) [Sch97] [Wol99].

Si y est un vecteur d'observation de R^d alors :

$$P(y|C_i) = \sum_{l=1}^{p} \nu_l^i N(y, \mu_l^i, \Sigma_l^i) \tag{1.14}$$

où
$N(y, \mu_l^i, \Sigma_l^i)$ représente la densité d'une loi multidimensionnelle,
μ_l^i représente le vecteur moyenne,
Σ_l^i représente la matrice de covariance,

ν_l^i représente une probabilité a priori,
et p désigne le nombre de lois gaussiennes multidimensionnelles dans le mélange,

avec pour les densités correspondantes à des lois normales multidimensionnelles :

$$N(y, \mu_i, \Sigma_i) = \frac{1}{(2\pi)^{d/2}|\Sigma_i|^{1/2}} e^{\left[-\frac{1}{2}(y-\mu_i)^t \Sigma_i^{-1}(y-\mu_i)\right]} \tag{1.15}$$

À titre d'illustration, la figure 1.11 présente un exemple en deux dimensions dans lequel sont représentés mille points issus d'un modèle de mélange comportant trois composantes.

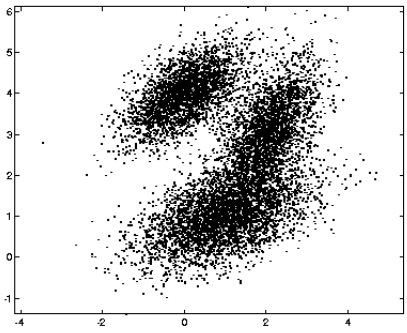

FIG. 1.11 – Exemple d'un mélange de trois densités gaussiennes bidimensionnelles.

On distingue les trois formes ellipsoïdales ; leur orientation différente rend compte des matrices de covariances différentes pour chaque composante du mélange. Il est clair que selon la dispersion liée à chaque loi, l'individualisation des composantes sera plus ou moins bien marquée.

En apprentissage, il s'agit d'estimer les paramètres des lois. Parmi les approches les plus courantes, citons [Dud01] :

- la méthode du **maximum de vraisemblance** : les paramètres de la loi recherchée sont considérés comme « fixes » et le choix optimal est celui qui maximise la vraisemblance des échantillons fournis pour définir la classe,
- la méthode d'**estimation Bayésienne** : les paramètres cherchés sont considérés comme les réalisations d'une variable aléatoire ayant une distribution a priori connue. La stratégie est fondée sur le maximum a posteriori des observations.

Exemple 2 : les MMC (Modèles de Markov Cachés). Ils seront décrits dans la section 6.2.2.

1.3.1.2 Méthodes non paramétriques

L'usage des méthodes dites paramétriques nécessite de faire des hypothèses sur la nature de la loi. De plus, selon la complexité de la loi retenue, l'apprentissage des paramètres peut s'avérer difficile. C'est alors que l'on a recours aux méthodes non paramétriques. De telles méthodes prennent en compte les échantillons issus d'un ensemble d'apprentissage et surtout leur répartition spatiale dans l'espace des observations pour obtenir une estimation de la densité de probabilité de chaque classe.

Les décisions, comme la méthode des k plus proches voisins ou celle des histogrammes, sont fondées sur ces statistiques locales ou globales.

Méthode de décision des k plus proches voisins (kppv) :

La méthode nécessite de disposer en permanence d'un ensemble d'apprentissage pour chaque classe. Ces ensembles permettent, pour toute nouvelle observation y, d'estimer localement $f(y|C)$ la densité conditionnelle relative à chaque classe. Cette estimation est réalisée en calculant le nombre relatif d'éléments d'apprentissage pour la classe donnée dans une boucle centrée sur y. Ce calcul nécessite de choisir une distance dans l'espace des observations ; une pondération de chaque élément est possible (noyau de Parzen).

La méthode de décision dite des « kppv » consiste à définir la boule d'estimation autour de y comme étant la plus petite boule contenant exactement k échantillons, toutes classes confondues. La classe affectée à y est alors la classe majoritairement présente dans cette boule (cf. figure 1.12).

Les distances les plus couramment utilisées sont :

– la **distance Euclidienne** ou L^2 :

$$d(x,y) = \sqrt{\sum_{j=1}^{p}(x^j - y^j)^2} \tag{1.16}$$

– la **distance « city block »** ou L^1 :

$$d(x,y) = \sum_{j=1}^{p}|x^j - y^j| \tag{1.17}$$

– la **distance de Chebychev** ou $L^{\inf}$:

$$d(x,y) = \max_j |x^j - y^j| \tag{1.18}$$

– la **distance de Mahalanobis** :

$$d(x,y) = \sqrt{(x-y)^t \Sigma^{-1} (x-y)} \tag{1.19}$$

avec x et y les éléments de R^d et Σ^{-1} une matrice de covariance inverse.

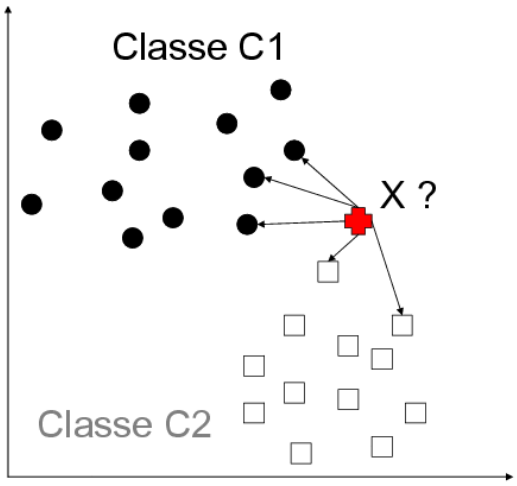

FIG. 1.12 – Exemple de décision selon les kppv. Pour $k = 5$ voisins, x est affecté à la classe $C1$ alors que par la méthode du plus proche voisin il aurait été affecté en $C2$.

Méthode des histogrammes :

Lorsque l'observation à classer est une suite de vecteurs de R^d et est considérée comme une suite de réalisations d'un vecteur aléatoire, il est alors possible de construire l'histogramme de ces réalisations. Cet histogramme, après normalisation, peut être considéré comme une estimation de la densité de probabilité des dites observations (cf. figure 1.13).

En phase d'apprentissage, il est nécessaire de construire l'histogramme des observations relatives à chaque classe.

La classification est obtenue en comparant des histogrammes entre eux à l'aide de mesures. [Foo97] a utilisé plusieurs mesures dont la corrélation, la distance euclidienne et le cosinus.

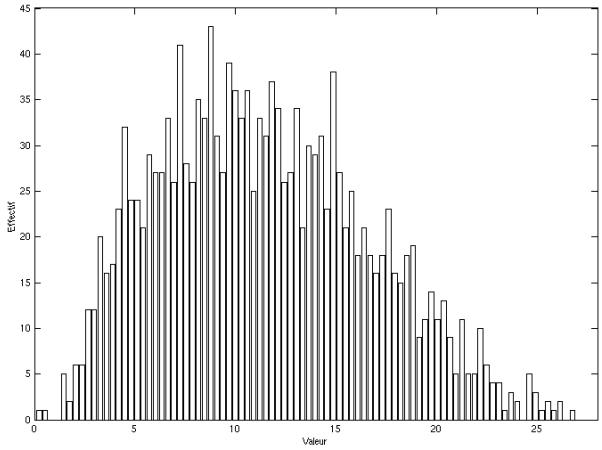

FIG. 1.13 – Exemple d'histogramme.

1.3.2 Méthode de décision fondée sur les réseaux de neurones

Il est classique de distinguer parmi les différents modèles neuronaux existants deux grandes catégories : celle qui nécessite un apprentissage supervisé et celle qui détermine leurs paramètres et leurs classes par apprentissage non supervisé.

Dans la première catégorie, le modèle le plus connu et le plus utilisé est le Perceptron Multicouches. Il permet de traiter les problèmes de régression non linéaire [Cri00].

Dans la seconde catégorie, on trouve le modèle des cartes topologiques dont la première version est due à Kohonen [Koh88]. Utiliser les cartes topologiques permet d'aborder les problèmes classiquement connus en statistique sous le nom de classification automatique non supervisée.

Dans la mesure où nous restreignons notre problème de classification au cas supervisé, nous nous limitons ci-dessous à la description du perceptron, classiquement utilisé en parole et musique.

Un réseau de neurones est un ensemble d'éléments simples, appelés neurones formels et reliés les uns avec les autres, qui se transmettent l'information par l'intermédiaire de ces liens ou connexions (cf. figure 1.14).

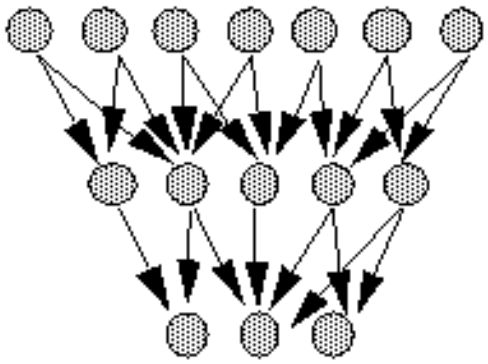

FIG. 1.14 – Exemple de réseau de neurones.

Chaque neurone formel réalise une somme pondérée des valeurs de ses entrées. Sa sortie est une modulation de cette somme (cf. figure 1.15).

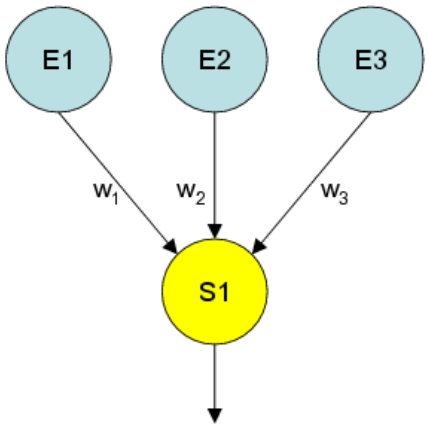

FIG. 1.15 – Schéma d'un neurone formel.

Pour n entrées, la sortie est donnée par la formule générale suivante :

$$S_1 = f_{act}(I) \tag{1.20}$$

avec :

$$I = \sum_{i=1}^{n} w_i E_i \tag{1.21}$$

La fonction f_{act} est appelée fonction d'activation, elle régularise les valeurs de sortie et prend en général ses valeurs dans l'intervalle $[0 - 1]$. Elle interdit les évolutions « catastrophiques », qui peuvent être dues à des effets de boucle où les valeurs deviennent de plus en plus grandes.

Il existe plusieurs fonctions d'activation qui vont des fonctions les plus « naïves » dites fonctions à seuils plus ou moins biseautés aux fonctions sigmoïdes (cf. figure 1.16) :

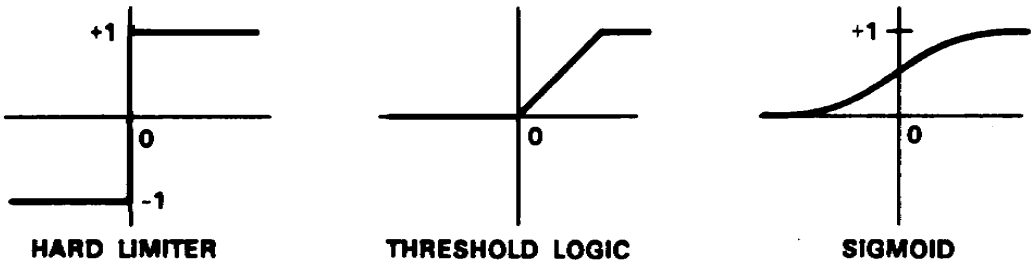

FIG. 1.16 – Quelques fonctions de seuils.

– La fonction « Hard Limiter » est une fonction seuil :

$$f_{act}(I) = \left\{ \begin{array}{lll} 1, & \text{si } I \geq 0 & \text{(neurone activé)} \\ -1, & \text{sinon} & \text{(neurone inhibé)} \end{array} \right\} \tag{1.22}$$

– La fonction « Threshold Logic » est une fonction linéaire :

$$f_{act}(I) = \left\{ \begin{array}{lll} 1, & \text{si } I \geq 1 & \text{(neurone activé)} \\ 0, & \text{si } I \leq 0 & \text{(neurone inhibé)} \\ I, & \text{si } 0 < I < 1 & \end{array} \right\} \tag{1.23}$$

– La fonction « Sigmoïde » est une fonction non linéaire :

$$f_{act}(I) = \frac{1}{1 + e^{-I}},\ 0 \leq f_{act}(I) \leq 1 \tag{1.24}$$

Les perceptrons sont des réseaux sans contre réaction. Ils sont définis à partir d'une répartition des neurones formels en couche : les sorties des neurones de la couche i forment les entrées de la couche $i + 1$.

Généralement, en entrée du perceptron sont données les coordonnées de l'observation à classer ; à chaque neurone de sortie correspond une classe. La classe reconnue correspond au neurone de sortie de plus forte activation.

La figure 1.17 montre la structure d'un perceptron à trois couches dont une couche cachée, appliqué à un problème de reconnaissance globale de mots. Ce modèle est issu des travaux de F. Rosenblatt [Ros62] sur le perceptron monocouche.

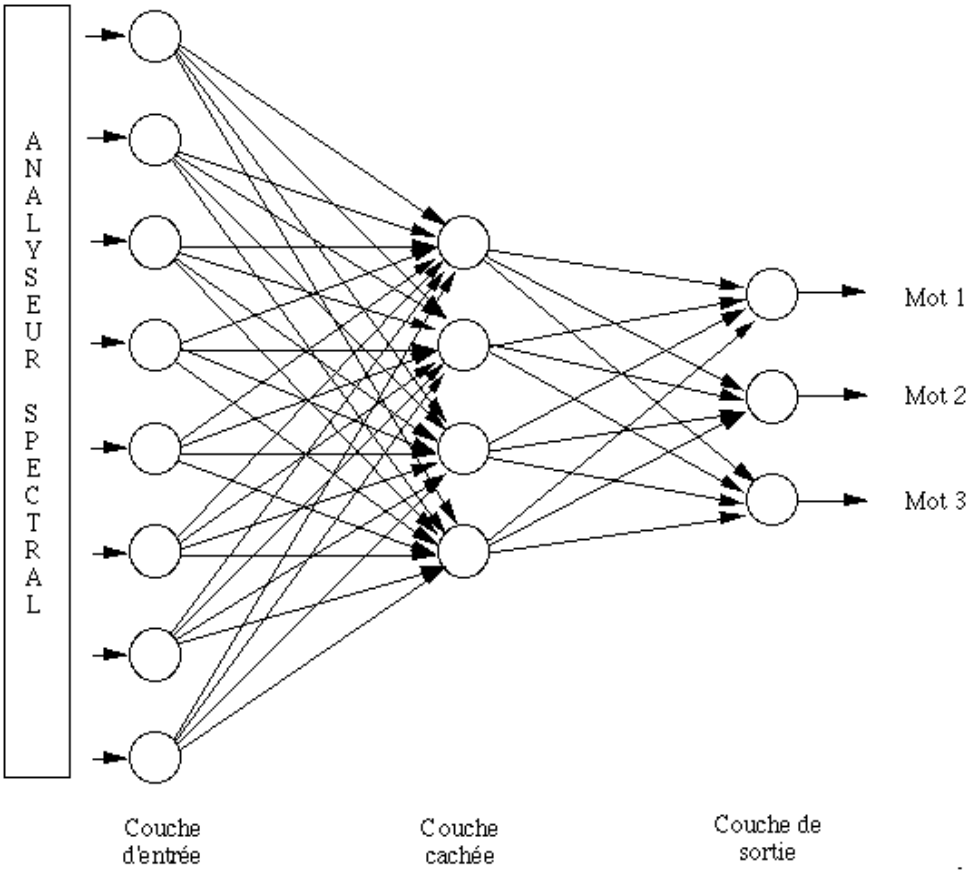

FIG. 1.17 – Structure d'un perceptron à trois couches.

Ce type de classification est actuellement très employé pour la fusion d'informations ou la fusion de scores issus de plusieurs sous-systèmes.

1.3.3 Support Vector Machines : SVM

La méthode des machines à vecteurs de support est une alternative récente pour la classification mais prometteuse [Cha02]. Cette méthode repose sur l'existence d'un hyperplan séparateur dans un espace approprié. Les travaux de [Bur98], [Cor95] et [Vap99] présentent les principes de base des SVM.

Cette méthode est fondée sur la minimisation du risque structurel et est meilleure que celle fondée sur la minimisation du risque empirique traditionnellement utilisée par les réseaux de neurones adaptatifs [Gun97].

La méthode SVM est une méthode discriminante, avec apprentissage supervisé. Il s'agit de rechercher une surface de décision optimale, déterminée par certains points de l'ensemble d'apprentissage appelés « vecteurs supports » en projetant les données d'entrée non-linéairement séparables dans un espace de plus grande dimension appelé « espace des caractéristiques ».

Soit un ensemble d'apprentissage : $(x_1, y_1), \cdots, (x_n, y_n)$ où y_i est le label de classe du vecteur x_i et $y_i = \pm 1$.

1.3.3.1 Cas linéairement séparable

Dans le cas de données linéairement séparables, tout hyperplan séparateur est paramétré par w et b (cf. figure 1.18) :

$$< x, w > +b = 0 \tag{1.25}$$

Soient les deux hyperplans parallèles à l'hyperplan séparateur, définis par :

$$< x, w > +b = \pm m \tag{1.26}$$

tels que :

$$y_i(< x_i, w > +b) \geq m, \ i = 1, \cdots, n \tag{1.27}$$

Le problème consiste à trouver l'hyperplan canonique qui maximise la marge, intervalle entre les deux hyperplans parallèles associés, à savoir (cf. figure 1.18) : $\frac{2m}{||w||}$.

En utilisant les contraintes (cf. équation 1.27), la solution est obtenue par optimisation en utilisant la théorie des multiplicateurs de Lagrange.

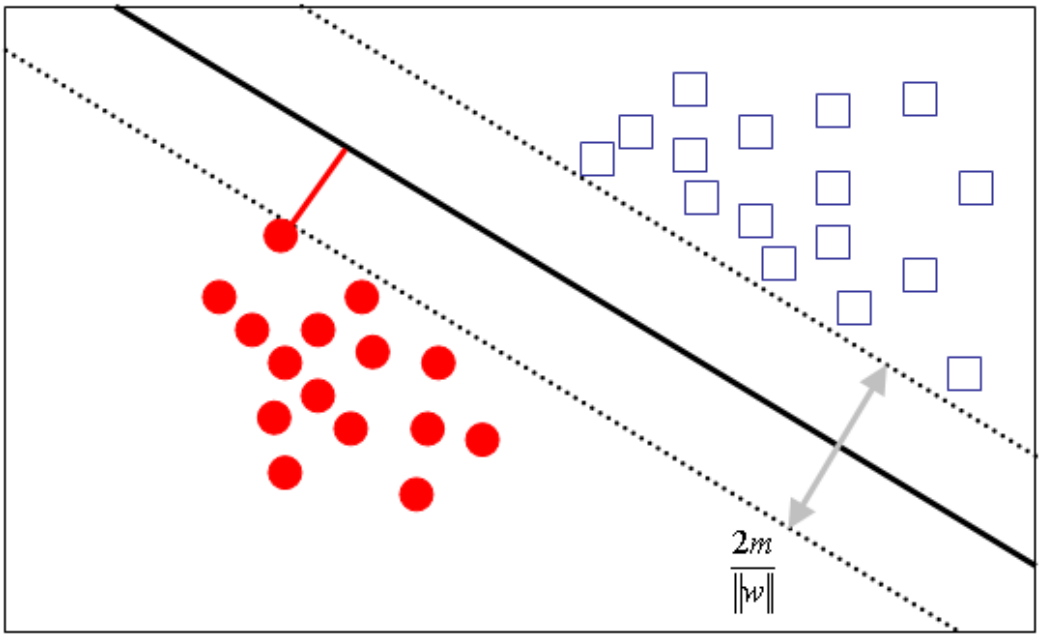

FIG. 1.18 – Hyperplan séparateur de deux classes maximisant la marge dans un cas linéairement séparable.

Le Lagrangien est donné par :

$$L(w, b, \alpha) = \frac{1}{2}||w||^2 - \sum_{i=1}^{n} \alpha_i (y_i(< w, x_i > +b) - 1) \tag{1.28}$$

où $\alpha_i > 0$ et $i = 1, \cdots, n$ multiplicateurs de Lagrange.

La fonction de décision devient :

$$f(x) = \text{sign}\left(\sum_{i=1}^{n} y_i \alpha_i < x, x_i > +b\right) \tag{1.29}$$

L'intérêt de cette méthode est que les α_i sont nuls pour presque tous les vecteurs d'apprentissage. Les autres sont appelés « vecteurs support » et sont en quelque sorte les paramètres du modèle.

1.3.3.2 Cas non linéairement séparable

Dans le cas non linéairement séparable (cf. figure 1.19), le classifieur de marge maximale ne peut pas être utilisé et le problème d'optimisation n'a pas de solution.

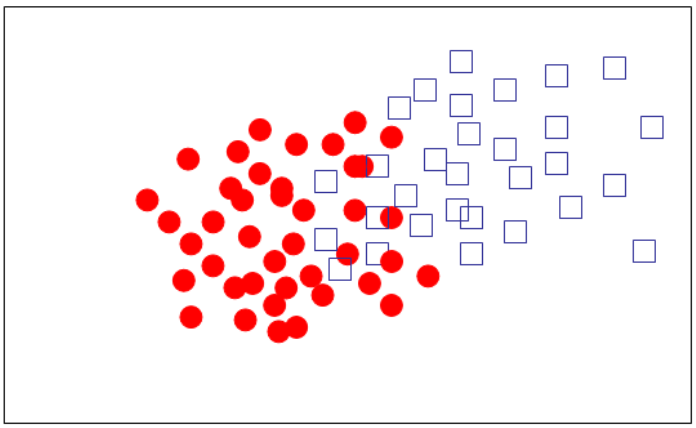

FIG. 1.19 – Exemple d'un cas non linéairement séparable.

L'**idée** consiste à trouver un espace de plus grande dimension dans lequel est plongé l'espace des observations χ. Cet espace, appelé espace des caractéristiques, est un « espace de Hilbert » F muni d'un produit scalaire $< ., . >$ tel que :

$$\begin{aligned} Si\ \Phi : \quad & \chi \mapsto F \\ & x \rightarrow \times = \Phi(x), \end{aligned} \tag{1.30}$$

$$k(x_1, x_2) =< \Phi(x_1), \Phi(x_2) > \text{ pour tout } (x_1, x_2) \in \chi^2.$$

k est la fonction noyau « Kernel trick » associée.

Dans F, l'hyperplan séparateur est paramétré par :

$$w = \sum_{i=1}^{n} \alpha_i \times_i \tag{1.31}$$

et la fonction de décision dans χ devient :

$$f(x) = \text{sign}\left(\sum_{i=1}^{n} y_i \alpha_i k(x, x_i) + b\right) \tag{1.32}$$

Remarque : cas des éléments anormaux

Afin de tolérer le bruit et prendre en compte les données d'apprentissage « hors frontières », il faut utiliser des « marges douces ».
Ceci revient à minimiser :

$$\frac{1}{2}||w||^2 + \frac{1}{n}\sum_{i=1}^{n}\xi_i \tag{1.33}$$

avec $y_i < w, x_i > \geq b - \xi_i$ pour tout $i = 1, \cdots, n$.

Les ξ_i sont des variables de relâchement.

La fonction de décision est inchangée (équation 1.32).

1.4 Les systèmes

Plusieurs méthodes de discrimination parole/musique ont été décrites dans la littérature. Elles peuvent se classer en deux groupes.

– Dans la **communauté des spécialistes en musique**, l'accent porte sur des paramètres permettant de séparer au mieux la musique du reste (non-musique). Par exemple, le taux de passage par zéro (Zero Crossing rate) et le centroïde spectral sont utilisés pour séparer le bruit des parties voisées (donc harmoniques) [Sau96], [Zha98] tandis que la variation de la magnitude spectrale (le flux spectral) permet de détecter les continuités harmoniques [Sch97].

– Dans la **communauté du traitement automatique de la parole**, les paramètres cepstraux sont privilégiés pour extraire les zones de parole [Gau99] et [Foo00]. *Pour plus d'informations voir la campagne d'évaluation NIST [NIS00].*

Quatre approches sont communément utilisées pour la classification : les modèles de mélanges de lois gaussiennes (MMG, [Sch97] et [Wol99]), les k plus proches voisins (kppv, [Car99]), les réseaux de neurones (RN, [Ros00]) et les MMC ([Zha98] et [Kim96]).

Les approches utilisées sont très souvent liées à la culture du laboratoire, comme l'illustrent les deux systèmes présentés ci-dessous. L'un est issu de l'IRCAM [Ros00], laboratoire où l'étude de la musique est privilégiée et l'autre est issu du LIMSI [Gau99] où l'objet d'étude est la parole.

1.4.1 IRCAM

Le système de segmentation en sources étudié par l'IRCAM se décompose en deux modules classiques :

– la *paramétrisation*. Toutes les 20 ms sont extraits le flux spectral, le centroïde spectral, le taux de passages par zéro et le flux spectral modifié en utilisant un lissage du spectre, par calcul du cepstre. Les moyennes et les logarithmes des variances de ces paramètres sont calculés sur une fenêtre d'une seconde et donnent les vecteurs d'observations.

– la *décision*. Trois méthodes de classification ont été testées, les lois gaussiennes multidimensionnelles, les k plus proches voisins ($k = 7$) et les réseaux de neurones (une

couche cachée de 6 à 8 neurones selon le nombre de fonctions d'observation qui varie de 2 à 6). Les expériences ont été réalisées sur des signaux réels, enregistrés à la radio. Elles montrent que le taux de segments mal classés peut descendre en dessous de 10 %, les RN et les kppv semblant donner les meilleurs résultats.

1.4.2 LIMSI

Le premier module du système de transcription automatique d'émissions de radio, développé au LIMSI, permet d'extraire du continuum sonore les segments de parole, en précisant s'il s'agit d'enregistrements bande large ou non (canal téléphonique), et s'il s'agit d'un locuteur masculin ou féminin.

Les segments de musique et les bruits sont éliminés. Pour chaque trame élémentaire de signal, 38 coefficients dérivés des MFCC, sont extraits et constituent le vecteur d'observation. La décision se décompose en trois procédures :

– quatre MMG, correspondant aux classes « parole », « parole bruitée », « musique » et « silence », permettent d'extraire les zones de parole,

– une phase de clusterisation permet de segmenter les segments de parole, de telle façon que les nouveaux segments obtenus correspondent chacun à un locuteur enregistré dans un environnement homogène. L'identification de l'environnement est obtenue en utilisant deux nouveaux MMG correspondant aux environnements canal téléphonique/large bande.

– une segmentation est obtenue de la même manière que précédemment pour préciser le genre du locuteur.

Lors de la campagne d'évaluation DARPA Hub-4E 1996, sur les systèmes de reconnaissance de la parole, le LIMSI a obtenu des résultats de 3 à 8 % d'erreurs sur la tâche de détection de parole.

Remarque : les zones de musique ne sont pas isolées.

1.5 Conclusion

Nous avons présenté dans ce chapitre, divers paramètres et modélisations pouvant caractériser ou discriminer la parole de la musique.

Quelle que soit la méthode de décision proposée, un ensemble d'apprentissage est nécessaire, afin d'apprendre les seuils, les caractéristiques des classes ou les paramètres des modèles statistiques (histogrammes, moyennes, variances, réseaux de neurones ou vecteurs supports). Les performances des classifieurs sont, comme dans de nombreux problèmes de reconnaissance des formes, extrêmement dépendantes de la qualité de cet apprentissage.

En classification parole/musique, il est difficile d'évaluer son propre système : est-ce que le système du LIMSI est plus performant que celui de l'IRCAM ?

- Les corpora utilisés ne sont jamais identiques d'un laboratoire à l'autre : base de données personnelles, corpus radiophonique...
- Le but n'est pas forcément le même : extraction de parole (respectivement musique) pour la communauté du traitement automatique de la parole (respectivement communauté des spécialistes en musique).

De plus les choix sont généralement binaires : y a-t-il de la parole ou de la musique ?
Ces choix ne correspondent pas à la réalité où nous voudrions savoir s'il y a de la parole, de la musique, les deux ou aucun.

Nous allons définir un « système de base » représentatif des systèmes dits « état de l'art », tant sur le plan de la modélisation que sur celui des performances, afin de ne plus se cantonner à une approche binaire d'une part et pour évaluer de manière plus objective les limites de cette tâche d'autre part.

Chapitre 2

Le système PMB de base

Sommaire

2.1 Introduction

Dans une approche classique de discrimination parole/musique, il est question d'un choix binaire : parole ou musique. Dès lors qu'il s'agit d'indexation, le but est de trouver les composantes parole et musique du document de façon indépendante puisque, celles-ci peuvent apparaître simultanément.

Il en résulte que le document doit être annoté de deux manières différentes et indépendantes : l'une parole/non-parole et l'autre musique/non-musique.

De plus, comme nous l'avons vu dans le chapitre précédent, la parole se caractérise par une structure formantique (section 1.1.1), tandis que la musique se caractérise par une structure harmonique (section 1.1.2). Par conséquent, les deux annotations peuvent dériver d'une stratégie de décision différente, que ce soit au niveau de la paramétrisation ou au niveau de la modélisation. C'est ce que nous appelons ensuite la **modélisation différenciée**.

Modélisation différenciée

Traditionnellement en classification, les classes à discriminer partagent le même espace de représentation (paramétrisation) et le même type de modélisation. Or elles peuvent présenter des différences qui motivent des espaces de représentation distincts et des modèles statistiques distincts.

Le but n'est plus de trouver des paramètres qui permettent de séparer au mieux ces classes, mais plutôt de trouver conjointement des ensembles de représentation et des modèles qui caractérisent au mieux chaque classe.

Dans cet esprit, chaque ensemble est alors défini par son espace de représentation, son modèle et son « anti-modèle » :

$$e = \{Espace\ de\ représentation, Modèle\ classe, Modèle\ non\text{-}classe\} \tag{2.1}$$

Un problème de classification à n classes peut se scinder en n problèmes de classification : n sous-systèmes. L'intérêt de ce type d'approche est que la décision n'est plus exclusive. Un même signal peut conduire simultanément à la détection de m classes (avec $m \leq n$).

Dans le cas présent de l'indexation parole/musique, nous proposons de distinguer les deux classes parole et musique en définissant :

$$\begin{aligned} Classe\ parole &= \{domaine\ cepstral, modèle\ parole, modèle\ non\text{-}parole\} \\ Classe\ musique &= \{domaine\ spectral, modèle\ musique, modèle\ non\text{-}musique\} \end{aligned} \tag{2.2}$$

Le choix des espaces de représentation est issu d'une étude effectuée à l'IRIT précédemment [Fon00]. Nous avions remarqué que les résultats étaient légèrement meilleurs en utilisant des coefficients spectraux pour les modèles musique et non-musique qu'en utilisant des coefficients cepstraux, classiquement utilisés pour la parole.

Le système de base

Comme en témoignent les systèmes de la campagne d'évaluation française en cours : ESTER[5] [Raz04], [Lam04] et [Fre04], nombre d'entre eux s'appuient sur une paramétrisation de type analyse cepstrale et une approche bayésienne fondée sur une modélisation par des modèles de mélanges de lois gaussiennes (MMG). *Pour plus d'informations sur cette campagne d'évaluation voir [Gra04].* Les travaux du LIMSI en indexation sonore, que ce soit pour la détection de parole [Gau99], la reconnaissance de locuteurs [Lam97], le suivi de locuteurs [Lam00] ou la transcription de parole dans n'importe quelle langue (l'allemand actuellement [MT03]) sont fondés sur ces mêmes éléments de base (analyse cepstrale, MMG).

Nous avons repris ces outils dans le cadre de la modélisation différenciée [Pin02], afin de définir un **système de base** représentatif des systèmes dits « état de l'art », tant sur le plan de la modélisation que sur celui des performances.

Ce chapitre est divisé en trois sections. Tout d'abord, le système est décrit à travers les paramétrisations (analyses cepstrale et spectrale) et la phase de décision. Puis, l'apprentissage des MMG est développé, notamment l'initialisation, l'optimisation et la ré-adaptation des modèles. Enfin, des expériences sur un corpus télévisuel permettent de valider notre système de base.

[5] http ://www.afcp-parole.org/ester/

2.2 Description du système

La modélisation différenciée impose à notre système de base [Pin02] une décomposition en deux sous-systèmes (cf. figure 2.1).

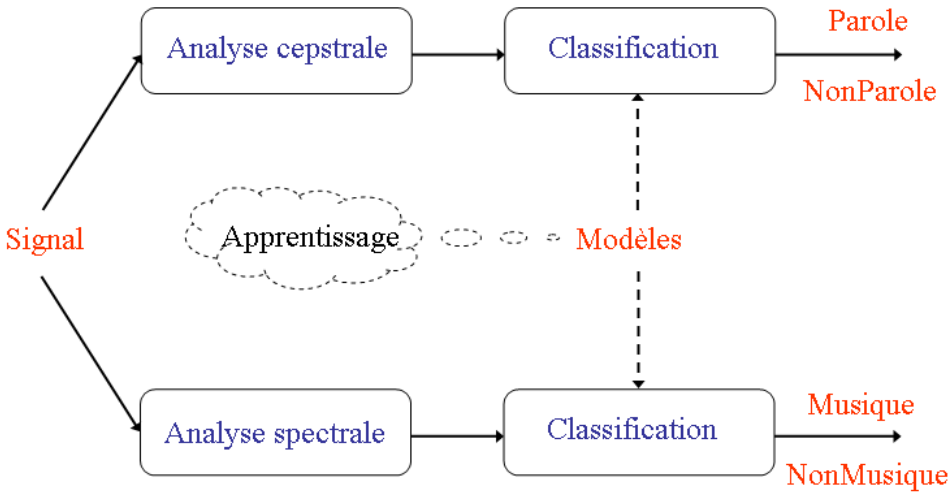

FIG. 2.1 – Système de classification de base.

Chacun de ces sous-systèmes possède la même structure : un pré-traitement acoustique afin d'extraire les paramètres pertinents du signal suivi d'une classification (ou reconnaissance) qui entraîne la prise de décision. Ils diffèrent au niveau du pré-traitement : le premier (parole/non-parole) est fondé sur une analyse cepstrale et le second (musique/non-musique) utilise une analyse spectrale.

2.2.1 Pré-traitement acoustique

Cette étape est essentielle puisqu'il s'agit d'extraire d'un signal fortement redondant, l'information pertinente relative à la tâche de classification donnée.

Le signal acoustique que nous traitons est échantillonné à 16 kHz. Les vecteurs d'observation peuvent être issus :

- d'un **traitement centiseconde**. Toutes les 10 ms, un vecteur de paramètres est extrait de l'analyse d'une centiseconde de signal (appelée aussi « trame »),
- d'un **traitement segmental**. Après une recherche de segments stationnaires de longueur variable, les paramètres sont extraits de chaque segment [AO88].

Une comparaison de l'approche segmentale avec l'approche centiseconde pour la reconnaissance de parole en présence de bruit a été réalisée par [Del96]. Les résultats n'étant pas tranchés, le choix s'est porté sur l'approche centiseconde, approche la plus classique.

2.2.1.1 Analyse cepstrale

Pour la parole, une analyse cepstrale (cf. figure 2.2) selon une échelle Mel est effectuée. Introduite dans la section 1.2.4, nous précisons ci-dessous les quelques détails pratiques de mise en œuvre.

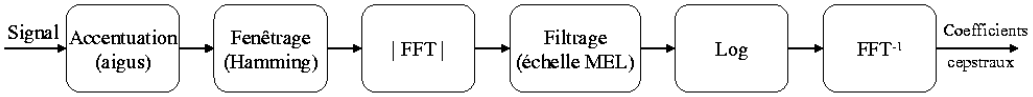

FIG. 2.2 – Analyse cepstrale.

Pour chaque trame une accentuation des aigus est réalisée (car les aigus sont toujours plus faibles en intensité que les graves) ; un fenêtrage est effectué pour atténuer les discontinuités (étalement de l'impulsion de Dirac due à une limitation en temps du signal). La fenêtre de Hamming est traditionnellement employée en parole.

Les énergies dans 24 filtres sont calculées après application du module de la FFT (Transformée de Fourier). Ces canaux sont répartis sur l'échelle de Mel : le nombre de coefficients utilisés pour décrire le signal est réduit tout en respectant une définition suffisante pour les fréquences qu'elles soient basses ou hautes, en respectant la perception humaine.

Les énergies des 24 canaux sont obtenues par l'application de filtres triangulaires se chevauchant et centrés sur les fréquences suivantes : 100, 200, 300, 400, 500, 600, 700, 800, 900, 1000, 1150, 1300, 1500, 1700, 2000, 2350, 2700, 3100, 3550, 4000, 4500, 5050, 5600, 6200 et 6850 Hertz.

L'analyse de chaque trame de signal donne un vecteur d'observation de 18 paramètres :

- l'énergie,
- 8 coefficients cepstraux (MFCC),
- la dérivée de l'énergie,
- 8 dérivées des coefficients cepstraux.

Cette suite de vecteurs subit ensuite une soustraction cepstrale. Chaque coefficient cepstral est diminué de la valeur moyenne des coefficients cepstraux. Ce traitement permet d'assurer une relative indépendance vis à vis du canal de transmission (microphone, ligne téléphonique...).

2.2.1.2 Analyse spectrale

Pour la musique, une simple analyse spectrale (cf. figure 2.3) est effectuée. Ainsi, les paramètres extraits sont les sorties de filtres et l'énergie.

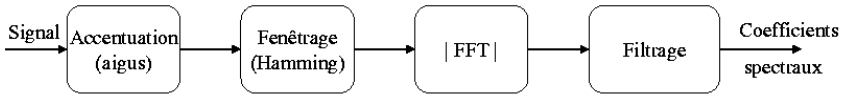

FIG. 2.3 – Analyse spectrale.

Le traitement permettant de trouver les coefficients spectraux est assez analogue au précédent.

Une accentuation des aigus et un calcul du fenêtrage sont effectués (Hamming). Les coefficients spectraux sont alors créés à la suite du calcul des énergies dans 28 filtres après le module de la FFT (Transformée de Fourier) et une pondération triangulaire (filtrage).

Les filtres se chevauchent et sont centrés sur des fréquences différentes de celles utilisées lors de l'analyse cepstrale. La répartition est linéaire par morceaux et les centres sont : 100, 200, 300, 400, 500, 600, 700, 800, 900, 1000, 1200, 1400, 1600, 1800, 2000, 2300, 2600, 2900, 3200, 3600, 4000, 4400, 4900, 5400, 5900, 6400, 6900 et 7400 Hertz.

Contrairement à l'analyse cepstrale, les dérivées ne sont pas calculées, l'énergie est conservée.

À chaque trame de signal, sont associés 29 paramètres :

– l'énergie,
– 28 coefficients spectraux.

2.2.2 Reconnaissance

Comme nous l'avons annoncé, nous avons opté pour une approche statistique. Il faut définir pour chaque sous-système, les distributions probabilistes de formes attendues (cf. figure 2.1). Nous avons choisi les modèles par mélanges de lois gaussiennes (MMG, cf. section 1.3.1.1) pour chacune d'elles, afin de se rapprocher de l'état de l'art.

La classification par MMG se fait par calcul du maximum de vraisemblance. Il en résulte pour chaque trame de signal, une décision parole/non-parole et une décision musique/non-musique.

À la suite de cette phase de classification, une phase d'assemblage permet de concaténer sous forme de segments les trames adjacentes ayant obtenu le même index lors de la classification (cf. figure 2.4.a, un exemple pour la reconnaissance de la parole).

(a) Exemple d'étiquetage d'un fichier pour une classification parole/non-parole après la phase de regroupement.

Temps (en ms)	300	1000	10	800	300	20	600	200	310
Etiquette	non-parole	parole	non-parole	parole	non-parole	parole	non-parole	parole	non-parole

(b) L'exemple (a) subit le premier lissage à 20 ms

Temps (en ms)	300	1810	920	200	310
Etiquette	non-parole	parole	non-parole	parole	non-parole

(c) L'exemple (b) subit le deuxième lissage à 500 ms

Temps (en ms)	2110	1430
Etiquette	parole	non-parole

FIG. 2.4 – Déroulement du lissage.

Une **fonction de lissage** est incorporée de manière à supprimer les segments non significatifs issus de l'assemblage. Généralement, dans une tâche d'indexation, l'important est de détecter un événement significatif, quitte à perdre un peu en précision. Pour une détection de parole, dans

le cas d'un pré-traitement d'une transcription automatique, le lissage ne serait pas si important afin de ne pas perdre le début des phrases et ainsi permettre au modèle de langage de fonctionner correctement.

Le lissage s'effectue en deux étapes. La première est un lissage à 20 ms (cf. figure 2.4.b) pour éliminer les segments trop petits. La seconde a lieu entre 500 ms et 2 s (cf. figure 2.4.c) pour ne garder que les segments représentatifs du son considéré. La valeur de lissage est étudiée expérimentalement.

Dans le cas où plusieurs étiquettes qui se suivent sont lissées, alors l'affectation se fait à la classe majoritairement représentée durant l'unité de lissage choisie.

2.3 Apprentissage des MMG

Les modèles de mélanges de lois gaussiennes sont des outils probabilistes très puissants, à condition de soigner leur apprentissage !

Cet apprentissage dit supervisé, nécessite une base d'apprentissage soigneusement étiquetée (cf. figure 2.5).

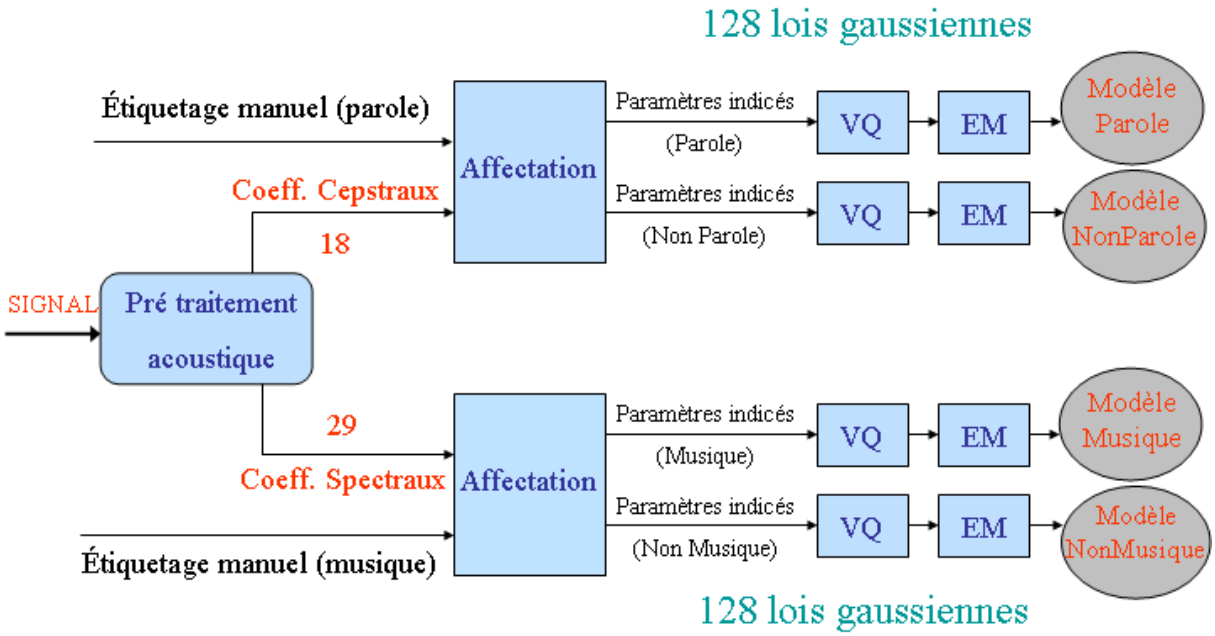

FIG. 2.5 – Déroulement de la phase d'apprentissage du système de base.

Si C désigne un MMG et y représente un vecteur d'observations :

$$P(y|C) = \sum_{k=1}^{K} \nu_k N(y, \mu_k, \Sigma_k) \tag{2.3}$$

où $\theta = (\nu_k, \mu_k, \Sigma_k,\ 1 \leq k \leq K)$ désigne les paramètres du MMG.

L'apprentissage de ces paramètres est classiquement réalisé par un algorithme de type EM (Expectation-Maximisation). Il se déroule en deux étapes. La première est une initialisation du modèle par Quantification Vectorielle (algorithme VQ) fondée sur l'algorithme de Lloyd [Llo57] et [Llo82]. La seconde étape est une optimisation des paramètres du mélange par l'algorithme classique Expectation Maximisation (algorithme EM) [Dem77].

2.3.1 Etiquetage manuel

L'étiquetage manuel s'effectue par rapport aux deux composantes : parole et musique. Deux étiquetages sont faits indépendamment :

- parole/non-parole,
- musique/non-musique.

Ils sont réalisés à l'aide du logiciel Transcriber (cf. annexe A). La figure 2.6 présente un exemple d'étiquetage parole/non-parole avec ce logiciel. Cette transcription manuelle fournit une liste de segments caractérisés par une étiquette et les temps de début et de fin ; ces temps sont exprimés en millisecondes.

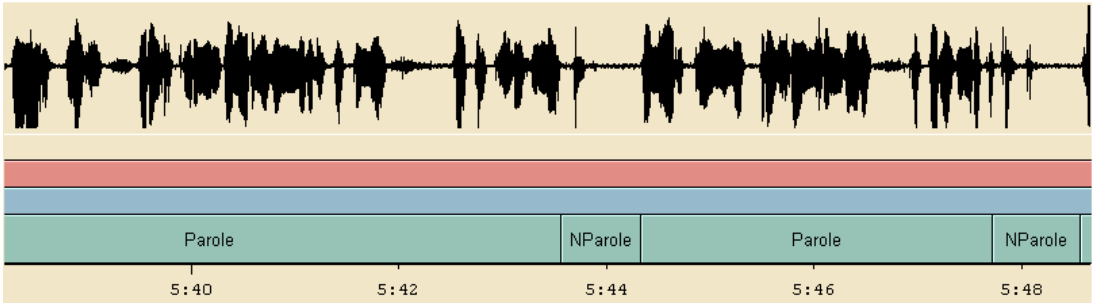

FIG. 2.6 – Exemple d'étiquetage parole/non-parole sur environ 10 secondes de signal à l'aide du logiciel Transcriber.

Chaque trame d'analyse (centiseconde) de l'ensemble d'apprentissage est positionnée par rapport à ces segments. Si la trame est interne à un segment, son étiquette est celle du segment ; si elle chevauche la frontière de deux segments, elle est ignorée et ne servira pas pour l'apprentissage, ceci afin que l'apprentissage soit le meilleur possible.

2.3.2 Initialisation des modèles

L'emploi de l'algorithme EM nécessite de disposer d'un modèle initial de mélanges de lois gaussiennes. Plus ce modèle est soigné, meilleure sera la convergence de cet algorithme.

L'initialisation des modèles est faite par Quantification Vectorielle (VQ). Elle s'appuie sur l'algorithme général des nuées dynamiques, dont la forme la plus utilisée en parole est l'algorithme LBG (Linde, Buzo, Gray) [Lin80]. L'algorithme repose uniquement sur un critère de calcul de distance (cf. annexe C).

La quantification vectorielle consiste à extraire un dictionnaire de prototypes (les « codes ») d'un grand ensemble représentatif de données. Le terme de dictionnaire désigne un ensemble fini de K vecteurs de référence ainsi que la partition de l'espace qui lui est associée au moyen d'une mesure de distance ou de distorsion.

Dans R^d, trouver les représentants du dictionnaire revient à chercher le dictionnaire D vérifiant la minimisation suivante :

$$min_d \left[\sum_{i=1}^{N} d(y_i, d_{\hat{i}}) \right] \tag{2.4}$$

avec :
$\hat{i} = argmin_k \; d(y_i, d_k)$,
$y = \{y_1, \cdots, y_N\}$ l'ensemble des observations,
$d = \{d_1, \cdots, d_K\}$ le dictionnaire.

Pour initialiser un modèle de type MMG d'ordre K (cf. équation 2.3), l'algorithme LBG est appliqué sur l'ensemble d'apprentissage correspondant et fournit un dictionnaire D de K éléments et une partition de l'ensemble d'apprentissage en K classes.

Les éléments de D sont considérés comme les moyennes des lois gaussiennes du mélange, la dispersion de chaque classe donne la matrice de covariance de la loi associée, et l'effectif de la classe sa pondération.

2.3.3 Optimisation des paramètres

L'algorithme EM décrit en annexe D, est fondé sur une approche par maximum de vraisemblance et suppose que les données d'apprentissage sont indépendantes et identiquement distribuées. La maximisation se fait en introduisant une fonction intermédiaire sur laquelle porte l'optimisation.

L'algorithme converge en un temps fini vers un extremum local mais il n'existe pas de critère absolu permettant la convergence vers un maximum global. De plus, l'algorithme EM n'effectue qu'une optimisation des composantes gaussiennes à nombre de lois gaussiennes, K constant.

En résumé, à chaque itération l'algorithme se décompose en deux phases :

- la phase d'**estimation** : la probabilité que la donnée i soit générée par la loi gaussienne k est calculée, pour toutes les observations de l'ensemble d'apprentissage i, $1 \leq i \leq N$ et pour toutes les lois gaussiennes k, $1 \leq k \leq N$, ceci à partir du modèle connu. K représente le nombre de lois gaussiennes et N le nombre d'observations,

- la phase de **maximisation** : une réévaluation des paramètres du modèle est effectuée à partir des observations d'apprentissage, pondérés par les probabilités calculées durant la phase d'estimation.

2.3.4 Adaptation des modèles : critère MAP

Précédemment les paramètres θ des modèles étaient estimés sur la base du critère du « Maximum de Vraisemblance ».

Lorsque le critère **MAP** « **M**aximum **A** **P**osteriori » est utilisé, les paramètres θ sont supposés être les réalisations d'une variable aléatoire de distribution $p(\theta)$. Les paramètres θ sont estimés à partir des données d'apprentissage y_{app} en résolvant [The99] :

$$\hat{\theta}_{MAP} = \arg\max_{\theta} p(\theta) P(y_{app}|\theta) \tag{2.5}$$

L'application classique de cette estimation est l'adaptation d'un modèle appris préalablement à de nouvelles données dites données d'adaptation y_{adp} :

$$\hat{\theta}_{MAP} = \arg\max_{\theta} p(\theta) P(y_{adp}|\theta) \tag{2.6}$$

La procédure d'adaptation de nos systèmes, modélisés par des MMG, est fondée sur cette méthode MAP [Gau94]. Comme [Mei01], nous adaptons uniquement les moyennes des lois gaussiennes. Les probabilités a priori et les matrices de covariance du modèle appris préalablement restent inchangées.

Pour chaque gaussienne g_l de dimension d, la moyenne $\mu^j_{l,adp}$ du modèle à p composantes est une combinaison linéaire de la moyenne estimée $\hat{\mu}^j_l$ et de la moyenne correspondante μ^j_l du modèle appris préalablement :

$$\mu^j_{l,adp} = \alpha\mu^j_l + (1-\alpha)\hat{\mu}^j_l \tag{2.7}$$

où :
$\alpha > 0$,
$j \in \{1, \cdots, d\}$,
et $k \in \{1, \cdots, n\}$.

Nous emploirons ce critère plusieurs fois au cours de ce travail, dès lors que les conditions d'apprentissage varieront notoirement. L'intérêt est qu'un volume de données d'adaptation très faible est nécessaire à la réestimation des modèles.

Comme l'a précisé [Dem77], l'algorithme EM permet d'appliquer le processus d'estimation MAP.

2.4 Expériences et évaluation

2.4.1 Corpus

Le corpus expérimental est formé à partir de documents audiovisuels appartenant à la base d'archivage de l'**INA**[6] (**I**nstitut **N**ational de l'**A**udiovisuel). Ce corpus a fait l'objet d'un accord de confidentialité lors du projet **RNRT AGIR**.

Il est composé d'une série télévisuelle « Chapeau melon et bottes de cuir » (corpus AIM), de journaux télévisés (le « 20 heures » de France2 et « SOIR3 »), de journaux sportifs (« Stade2 » et « Sport dimanche ») et d'un championnat du monde de patinage artistique.

Le signal est échantillonné à 16 kHz et a l'avantage de présenter de longues périodes de parole comme de musique dites « pures » ainsi que des zones dites « mixtes » contenant de la parole et de la musique et/ou du bruit.

La parole est présente sous diverses conditions, de pure à très bruitée (enregistrements en extérieur, foule, poursuites en voiture, parole téléphonique, deux locuteurs simultanés, cocktail party...).

Pour la musique, des variétés, de la techno, du jazz et de la musique classique sont majoritairement présents : vents (surtout des cuivres), guitare électrique et batterie-percussions. Quelques autres instruments (harpe, piano...) sont également utilisés sporadiquement.

La difficulté la plus importante du corpus correspond à la séquence de patinage artistique : pendant le passage sur la glace de chacun des candidats, le présentateur commente sur le fond musical. La superposition de la parole et de la musique dans ce cas là est problématique, il s'y ajoute le problème de deux canaux de transmission différents.

Le corpus est divisé arbitrairement en deux parties. Pour l'apprentissage des MMG de ce système de base, nous avons utilisé environ 12 heures du corpus (1ère partie) et pour l'ensemble des tests, environ 6 heures (2ème partie).

L'étiquetage manuel des données contenues dans ce corpus suit le protocole décrit dans la section 2.3.1.

[6] http ://www.ina.fr/

2.4.2 Élaboration des modèles

Après expérimentation, le nombre de lois gaussiennes dans le mélange est fixé à 128 pour chacun des modèles : parole, non-parole, musique et non-musique. Les matrices utilisées lors de l'apprentissage sont diagonales, ce qui équivaut à supposer les coordonnées des vecteurs d'observations indépendantes.

Le lissage est fixé à 500 millisecondes pour la détection de parole et à 2 secondes pour la détection de musique. Une phrase de parole dure en général 1 à 2 secondes : choisir un lissage d'une demi seconde n'est pas aberrant dans notre tâche d'indexation. Un extrait de musique, même court comme un jingle, dure au moins 2 secondes : le lissage n'est pas dramatique là non plus.

De manière classique :

- le nombre d'itérations des algorithmes VQ (cf. section 2.3.2) et EM (cf. section 2.3.3) est fixé à 10,
- le seuil (critère d'arrêt) pour la variation de la vraisemblance est fixé à 10^{-5}.

La plupart du temps, l'arrêt des algorithmes VQ et EM se fait par le nombre d'itérations. Mais ceci ne menace pas la convergence des modèles car la variation de la vraisemblance est alors très faible.

2.4.3 Évaluation

2.4.3.1 L'accuracy

L'évaluation de l'indexation automatique est effectuée en comparaison directe avec l'étiquetage manuel. Après avoir aligné les deux segmentations, plusieurs mesures sont possibles. L'une d'elles, l'**accuracy**, en l'absence de substitution s'écrit :

$$\frac{D_{Corpus_{test}} - D_{Insertions} - D_{Omissions}}{D_{Corpus_{test}}} \tag{2.8}$$

avec D représentant la durée. Les insertions et les omissions sont très liées dans le cas de deux classes : une insertion de l'une correspond à une omission de l'autre.

L'accuracy n'est pas une mesure intéressante en indexation car elle n'intègre pas le décalage que nous introduisons par notre lissage : ce décalage étant jugé acceptable dans notre tâche

d'indexation ; la précision de notre système est très liée à la valeur du lissage.

De ce fait, nous utilisons une mesure développée par le **NIST**[7] (**N**ational **I**nstitute of **S**tandards and **T**echnology). Il s'agit d'un outil de mesure de performance dédié aux segmentations et fondé sur l'accuracy dans lequel le décalage autorisé est paramétrable.

Nous avons choisi la moitié de la valeur du lissage comme décalage, soit 250 millisecondes pour la détection de parole et 1 seconde pour la détection de musique. En faisant un lissage de n millisecondes, la précision ne peut être que de $n/2$ millisecondes : le lissage se fait sur la classe majoritaire (au moins $n/2 + 1$ millisecondes).

Par la suite, la valeur renvoyée par cet outil sera notée : **Taux de Classification Correcte** (TCC).

2.4.3.2 Résultats

Les résultats obtenus, que ce soit pour la parole ou pour la musique, sont bons compte tenu de la nature du corpus. En effet, les taux de classification correcte sont de 93,9 % pour la détection de parole et de 91 % pour la détection de musique (cf. tableau 2.1).

TAB. 2.1 – *Taux de classification correcte pour les classifications parole/non-parole et musique/non-musique.*

Classification	Performances : TCC (Taux de Classification Correcte)
parole/non-parole	93,9 %
musique/non-musique	91 %

Les principales erreurs sont dues au fond musical, pouvant être très faible ou très important, présent sur environ quarante minutes d'une émission comme le patinage. Ce fond musical n'est pas reconnu lorsqu'il possède une intensité très faible d'où les erreurs en détection de musique et lorsqu'il est important la détection de parole échoue.

Les autres erreurs sont plus classiques. En général, il s'agit du problème de la voix chantée qui est soit reconnue comme de la parole, soit comme de la musique ou soit comme les deux. La variété des données d'apprentissage est assez faible, elle ne représente pas tous les cas possibles pour les données de tests. Malheureusement pour améliorer notre apprentissage, un étiquetage manuel coûteux et très pénible est nécessaire.

[7] http ://www.nist.gov/speech/tools/

Le tableau 2.2 présente les effets de la variation de la valeur du décalage sur les détections de parole et de musique. Lorsque la valeur du décalage permis augmente, le taux de classification correcte s'améliore également jusqu'à atteindre plus de 97 %. Ceci signifie que le système est robuste car les erreurs sont dues en majeure partie à des écarts de positionnement des frontières entre la détection manuelle et automatique et très peu à des erreurs de détection.

TAB. 2.2 – *Taux de classification correcte pour les classifications parole/non-parole et musique/non-musique en fonction de la valeur de décalage.*

Classification	TCC	TCC	TCC	TCC	TCC
Décalage parole/non-parole	5 ms (trame) 80,8 %	100 ms 88,9 %	250 ms (référence) *93,9 %*	500 ms 96,6 %	1 s 97,4 %
Décalage musique/non-musique	5 ms (trame) 61,8 %	250 ms 74,5 %	1 s (référence) *91 %*	2 s 94,5 %	4 s 97,6 %

La phase d'adaptation des modèles selon le critère MAP (adaptation des moyennes uniquement) donne des résultats légèrement moins bons (cf. tableau 2.3) que les précédents mais corrects. L'apprentissage est effectué ici uniquement sur 10 heures de la première partie du corpus et l'adaptation sur les 2 heures restantes. Le corpus de test est bien sûr identique !

Ce phénomène s'explique simplement par la forte hétérogénéité du corpus. Une adaptation n'a de sens que si elle rapproche les conditions d'apprentissage de celles de test, ce qui n'est pas assuré ici.

TAB. 2.3 – *Résultats pour les classifications parole/non-parole et musique/non-musique après adaptation des modèles (critère MAP).*

Classification	TCC (référence)	TCC (critère MAP)
parole/non-parole	**93,9 %**	92,8 %
musique/non-musique	**91 %**	90,1 %

Le système peut, bien sûr, incorporer un nouveau module faisant la fusion de la reconnaissance de la parole et de la musique. Un exemple de regroupement est présenté sur la figure 2.7 et deux nouvelles étiquettes sont alors créées :

– « PM » : pour « parole » et « musique »,
– « - » : pour ni « parole » (non-parole) ni « musique » (non-musique).

Classification parole/non-parole	P	NP	P	NP	
Classification musique/non-musique	NM	M			NM
Regroupement	P	M	PM	M	-

FIG. 2.7 – Exemple de fusion des indexations parole/non-parole et musique/non-musique.

La figure 2.8 est un exemple de résultats obtenus par les classifications parole/non-parole et musique/non-musique.

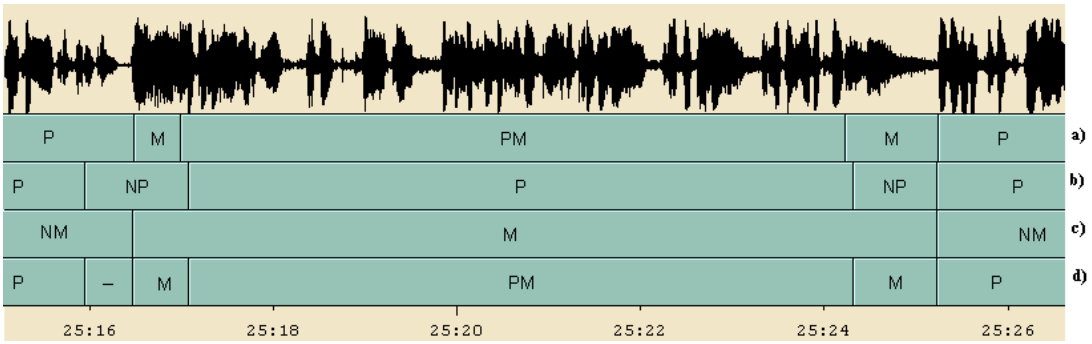

FIG. 2.8 – Exemple de résultats de notre système de base pour des classifications parole/non-parole b) et musique/non-musique c). La fusion est représentée sur la ligne d) et l'étiquetage manuel correspondant sur la ligne a). « P » désigne la parole, « M » la musique, « PM » parole et musique et « - » le bruit.

Remarque : Le fichier de sortie créé est au format XML d'une part pour être portable et réutilisable mais aussi pour permettre de visualiser les résultats et de les comparer à une indexation manuelle qui aurait été faite sur le logiciel Transcriber (cf. annexe A) par exemple.

2.5 Conclusion

Le système décrit dans ce chapitre constitue la base d'un système d'indexation automatique classique. Il permet une séparation parole/non-parole d'une part et musique/non-musique d'autre part.

Fondé sur une modélisation différenciée, il permet de redéfinir la notion de décomposition PMB (Parole/Musique/Bruit). L'idée de la méthode est d'associer à chacune des classes son propre espace de représentation et sa propre modélisation. Cette approche est mise en œuvre à partir de MMG sur la base d'une analyse cepstrale pour la parole et d'une analyse spectrale pour la musique.

Les résultats d'indexation obtenus pour la parole et pour la musique sont corrects : 93,9 % pour la classification parole/non-parole et 91 % pour la classification musique/non-musique. Ils sont similaires aux systèmes correspondant à l'état de l'art [Gau99] (jusqu'à 7,9 % d'erreurs en détection de parole sur l'évaluation DARPA). Ce système nous servira de référence en classification parole/musique et nous le nommerons par la suite « système de référence ».

Malheureusement, les modèles MMG nécessitent un apprentissage soigné, à partir d'un étiquetage manuel, long et rébarbatif. Compte tenu de la grande variété des documents rencontrés en indexation et structuration de bases de données sonores, ceci nous a amené à faire de la réestimation des modèles par le critère MAP en adaptant les moyennes. Mais, cette adaptation est assez lourde : un étiquetage, bien que très faible, est encore nécessaire et les résultats sont revus à la baisse.

Il devient utile de proposer une nouvelle alternative en reconsidérant de nouveaux paramètres et modèles.

Chapitre 3

Le système de classification Parole/Musique/Bruit

Sommaire

3.1 Introduction

Lors du chapitre 1, nous avons présenté deux approches en discrimination parole/musique à travers les communautés des spécialistes en musique et en traitement automatique de la parole. Deux systèmes ont été décrits : celui de l'IRCAM et celui du LIMSI.

S'inspirant de ces travaux, et notamment de ceux du LIMSI, nous avons développé un système d'indexation fondé sur une modélisation différenciée, où il n'était plus question de chercher à discriminer la parole de la musique, mais à les caractériser au mieux de façon indépendante afin de faire une séparation de type classe/non-classe : c'est-à-dire parole/non-parole et musique/non-musique (cf. chapitre 2).

L'approche est mise en œuvre à partir de MMG en utilisant des paramètres spectraux (fréquentiels) pour la classification musique/non-musique et cepstraux (MFCC) pour la classification parole/non-parole. Cette approche est concluante en terme de résultats (environ 90 %) mais nécessite un changement (ou une réestimation) de nos modèles (parole, non-parole, musique et non-musique) lorsque nous utilisons un corpus de test différent de celui de l'apprentissage.

C'est pourquoi nous proposons une méthode alternative de classification. Celle-ci consiste à toujours détecter de manière disjointe les segments de parole et de musique puisque, de par la nature même des signaux de parole et de musique, leur extraction ne peut résulter de l'utilisation d'outils communs.

L'**originalité** de cette nouvelle approche est l'utilisation de paramètres inhabituels : la modulation de l'entropie, la durée des segments (issue d'une segmentation automatique) et le nombre de ces segments par seconde.

Le choix de ces paramètres est fait de manière à détecter les deux composantes aussi bien dans des zones simples (zones contenant seulement de la parole ou de la musique) que dans des zones critiques (zones contenant à la fois de la parole, de la musique et/ou du bruit). À ces paramètres est adjointe la classique modulation de l'énergie à 4 Hertz. La stratégie de décision repose sur de simples fonctions à seuil. Différentes fusions de décisions sont proposées.

L'**intérêt** de cette méthode est sa robustesse vis à vis des conditions d'enregistrements, et des types de documents. Une fois les quelques seuils appris, aucun nouvel apprentissage n'est nécessaire pour traiter un document d'un nouveau type et enregistré dans de nouvelles conditions : l'estimation de nos paramètres de fonctionnement est faite une fois pour toutes.

Ce chapitre est divisé en quatre parties. Une première section est dédiée à la description du système global de classification et des paramètres. Une première expérience, effectuée sur un corpus de parole lue et de toutes sortes de musique, montre la pertinence du choix des paramètres par leur distribution et fait l'objet de la deuxième section. Au cours de la troisième partie, un deuxième corpus (radiophonique) est employé afin de vérifier la robustesse attendue des paramètres et du système de fusion proposé dans des conditions très diverses (reportages, informations, chansons, interviews...) : une comparaison au système de référence est effectuée. Lors de la dernière section, différentes méthodes de fusion (théorie des probabilités et théorie de l'évidence) sont utilisées et comparées.

3.2 Le système global et ses paramètres

Le système de fusion d'informations proposé [Pin03b] est fondé sur l'extraction de quatre paramètres (cf. figure 3.1) :

- la modulation de l'énergie à 4 Hertz,
- la modulation de l'entropie,
- le nombre de segments par seconde,
- la durée de ces segments.

Le système se décompose en deux systèmes de classification correspondant aux deux détections disjointes de la parole et de la musique. Nous les appellerons respectivement système parole/non-parole et système musique/non-musique. Ainsi, les passages contenant de la parole, de la musique mais aussi simultanément de la parole et de la musique sont détectés. La décision est prise en comparant les scores (vraisemblances) issus de la modélisation de chacun des paramètres considérés.

3.2.1 Le système global

Un pré-traitement commun aux deux sous-systèmes consiste à détecter le silence afin de ne traiter par la suite que les zones d'activité acoustique. Cette détection se fait sur la base de calculs d'énergie par rapport à un seuil. Cette méthode est classiquement utilisée dans la littérature [Zha98]. Dans notre système, les résultats sont donnés pour chaque seconde : une classification « silence » signifiera que le silence est majoritairement représenté durant la seconde de test, c'est-à-dire au moins durant 0,5 s.

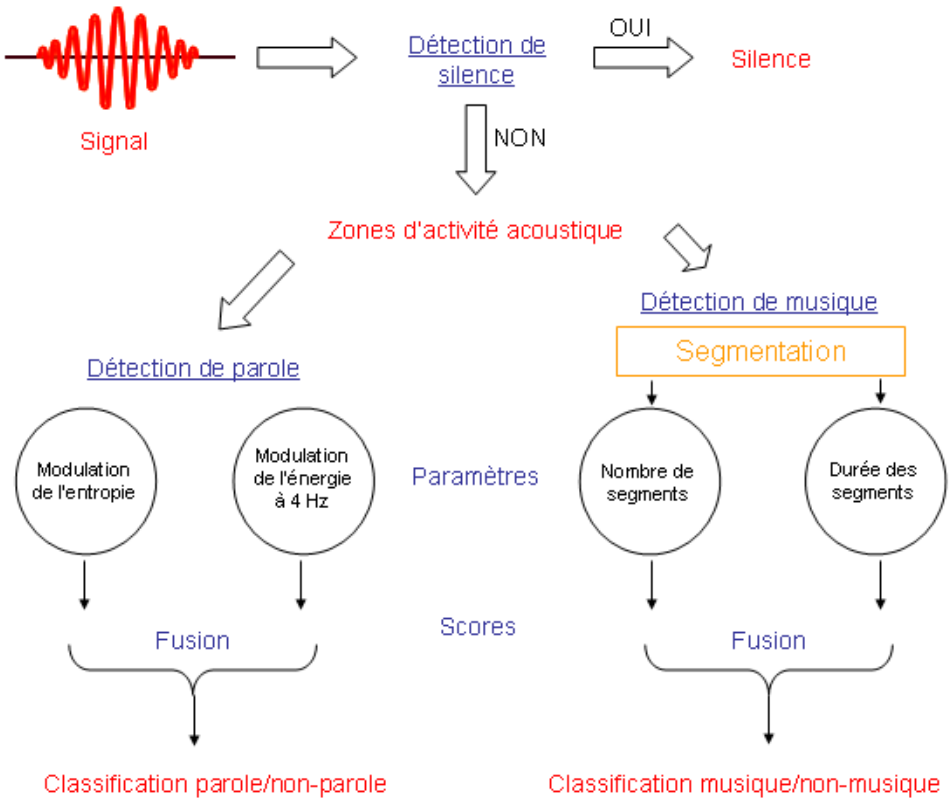

FIG. 3.1 – Le système global de fusion de paramètres

Dans le système parole/non-parole, l'information primaire traitée est issue de l'analyse d'une trame de 16 ms, alors que dans le système musique/non-musique, l'information est obtenue à partir du traitement de segments de taille variable : la notion de trame n'existe pas. Malgré tout, dans chaque cas, une décision est prise sur l'analyse complète d'une fenêtre glissante d'une seconde.

Le système parole/non-parole repose sur l'extraction de la modulation de l'énergie à 4 Hertz et la modulation de l'entropie tandis que le système musique/non-musique utilise les deux paramètres issus d'une segmentation automatique.

Chaque système correspond en fait lui même à deux « classifieurs » statistiques liés à chacun des deux paramètres en cause. 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 un certain indice de confiance (score de vraisemblance) et il faut fusionner les scores obtenus pour obtenir la décision de chaque système.

La classification hiérarchique proposée se décline de la manière suivante :

– dans le premier système, consacré à la détection de parole, nous avons fusionné les deux paramètres de modulation de l'énergie à 4 Hertz et de modulation de l'entropie par maximisation des scores de vraisemblance. Le score de vraisemblance le plus important détermine le choix (parole ou non-parole).

– dans le second, consacré à la détection de musique, les vraisemblances des deux paramètres de segmentation (le nombre de segments par seconde et la durée moyenne de ces segments par seconde) ont été fusionnées. La méthode de fusion est la même que précédemment, fondée sur la maximisation des scores.

3.2.2 Modulation de l'énergie à 4 Hertz

Le signal de parole possède un pic caractéristique de modulation en énergie autour de la fréquence syllabique 4 Hertz [Hou85]. En effet, ces modulations correspondent au rythme syllabique. Pour extraire cette information, la procédure suivante est appliquée.

1. Le signal est découpé en trames de 16 ms sans recouvrement afin de garder les transitions intéressantes (événements courts).
2. Pour chaque trame, après un fenêtrage de Hamming, 40 coefficients spectraux sont extraits suivant l'échelle Mel et correspondent à l'énergie des 40 bandes de fréquence en accord avec les propriétés de l'oreille humaine [Hou85]. Ces bandes sont également appelées canaux. L'énergie dans chacun des canaux fait apparaître des syllabes (cf. figure 3.2).

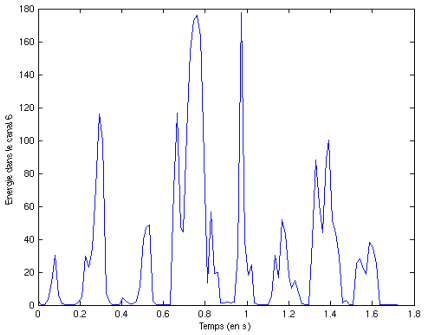

FIG. 3.2 – Évolution de l'énergie dans un canal sur 1,7 secondes de parole : les syllabes apparaissent (canal 6 : <500 Hz, 700 Hz>).

3. L'énergie de chaque bande est filtrée grâce à un filtre à Réponse Impulsionnelle Finie (RIF) passe-bande de fréquence centrale 4 Hertz (cf. figure 3.3).

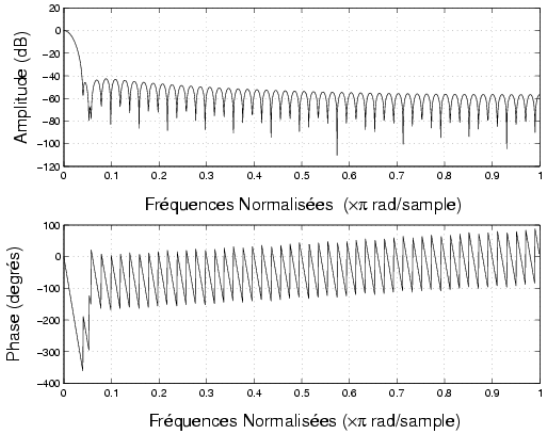

FIG. 3.3 – Réponse en fréquence et en phase du filtre RIF.

4. La somme des énergies filtrées est effectuée sur l'ensemble des canaux et est normalisée par l'énergie moyenne.
5. La modulation est obtenue en calculant la variance de l'énergie filtrée en décibels, sur une seconde de signal.

La parole possède une modulation de l'énergie à 4 Hertz plus forte que la musique (cf. exemples de la figure 3.4).

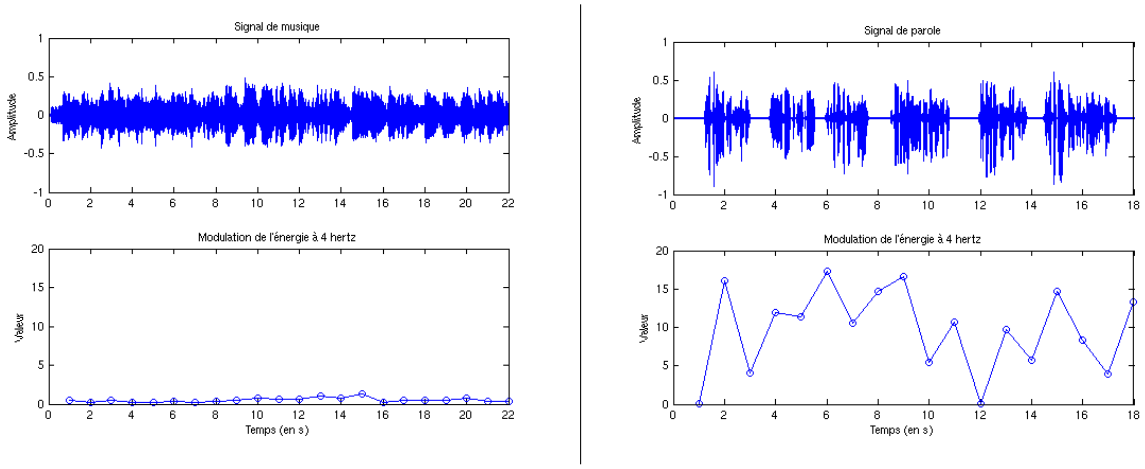

FIG. 3.4 – Modulation de l'énergie à 4 Hertz pour la musique (extrait de Mozart) et la parole (6 phrases de parole lue).

3.2.3 Modulation de l'entropie

Des observations menées sur le signal ainsi que sur le spectrogramme font apparaître une structure plus « ordonnée » du signal de musique que de parole. Pour mesurer ce « désordre », nous avons calculé un paramètre fondé sur l'entropie du signal [Mod89] :

$$H = \sum_{i=1}^{k} -p_i log_2 p_i \tag{3.1}$$

avec p_i = probabilité de l'événement i et k = nombre d'évènements.

Une procédure similaire à celle employée pour le paramètre de modulation de l'énergie à 4 Hertz est appliquée.

1. Le signal est découpé en trames de 16 ms sans recouvrement.
2. L'entropie est estimée pour chaque trame grâce à un estimateur non biaisé. Le calcul se fait en deux étapes. Tout d'abord, un histogramme est calculé afin de préciser la notion d'évènements, puis l'entropie est ensuite estimée.

   Soit N le nombre d'échantillons contenus dans la fenêtre considérée.

   – **Calcul de l'histogramme**

   La borne minimale de l'histogramme est définie par :

   $$min_h = min(x) - \frac{\Delta}{2} \tag{3.2}$$

   La borne maximale est définie par :

   $$\max_h = \max(x) + \frac{\Delta}{2} \tag{3.3}$$

   avec $\Delta = \frac{\max(x) - min(x)}{N-1}$.

   Le pas de quantification de l'histogramme est défini de telle sorte que le nombre N_k de classes soit l'arrondi supérieur de la racine carré du nombre d'échantillons :

   $$N_h \approx \sqrt{N} \tag{3.4}$$

– **Éstimation de l'entropie**

Une fois l'histogramme obtenu, on possède les probabilités d'apparition des différentes valeurs de l'amplitude (on notera h_i l'effectif de la classe i, $i = 1, \cdots, N_h$).
En considérant que les échantillons sont indépendants ($\widehat{H} = \sum_{n=1}^{N} \widehat{H_n}$), on effectue le calcul de l'estimateur biaisé [Mod89] :

$$\widehat{H}_{biased} = \frac{\sum_i(-h_i log(h_i))}{N} + log(N) + log(\frac{\max_h - min_h}{N_h}); \tag{3.5}$$

dont le biais est :

$$nbias = -\frac{N_h - 1}{2N} \tag{3.6}$$

L'estimateur non biaisé utilisé est donné par :

$$\widehat{H}_{unbiased} = \widehat{H}_{biased} - nbias \tag{3.7}$$

3. La modulation est obtenue en calculant la variance de l'entropie sur une seconde de signal. Puisque nous avons choisi de calculer l'entropie sur des trames de 16 ms, nous obtenons 62 valeurs de l'entropie $\widehat{H}$ par seconde.
On pose $\Psi = \{\widehat{H_1}, \cdots, \widehat{H_{62}}\}$, alors la modulation d'entropie est définie par :

$$modulation_H = var(\Psi) \tag{3.8}$$

Compte tenu du « désordre » annoncé de la parole par rapport à la musique, la modulation de l'entropie doit être plus élevée pour la parole que pour la musique (cf. figure 3.5).

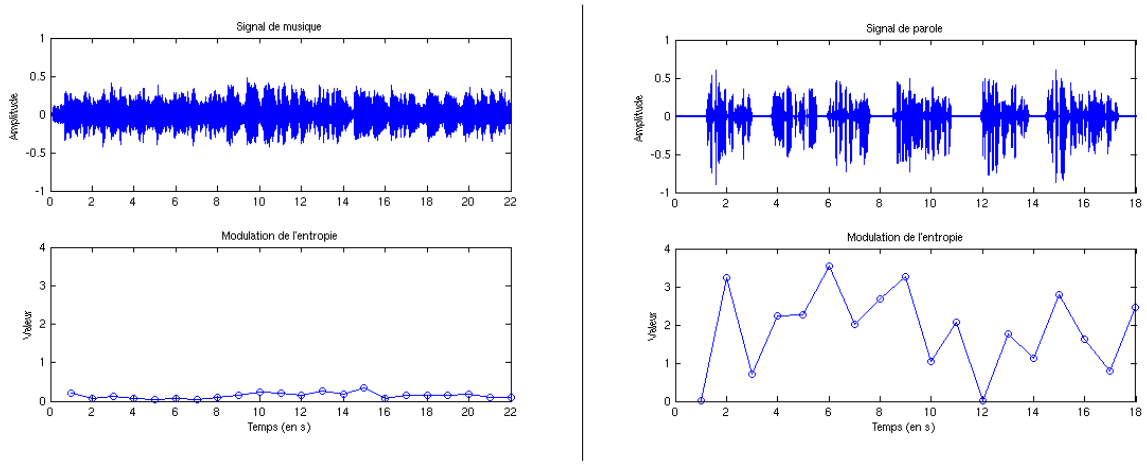

FIG. 3.5 – Modulation de l'entropie pour la musique et la parole.

Les courbes représentées sur la figure 3.5 paraissent similaires à celles obtenues avec le critère de modulation de l'énergie à 4 Hertz (cf. figure 3.4). L'échelle n'est certes pas la même, mais quelques variations permettent de supposer que la modulation de l'entropie apporte une information complémentaire. Cette hypothèse devra être confirmée lors de l'étape de fusion.

3.2.4 Paramètres de segmentation

Comme nous l'avons vu au cours du chapitre 1, la longueur des segments quasi stationnaires est différente pour la parole et la musique. En utilisant une segmentation du signal en zones quasi stationnaires, nous cherchons à mettre en évidence cette information.

3.2.4.1 Segmentation automatique

La segmentation est issue de l'algorithme de « Divergence Forward-Backward » (DFB) [AO88] qui est fondé sur une étude statistique du signal dans le domaine temporel. En faisant l'hypothèse que le signal est décrit par une suite de zones quasi stationnaires, chacune est caractérisée par un modèle statistique, le modèle autorégressif gaussien :

$$\begin{cases} x_n = \sum a_i x_{n-i} + e_n \\ \\ var(e_n) = \sigma_n^2 \end{cases} \tag{3.9}$$

où (x_n) est le signal et (e_n) est un bruit blanc gaussien.

La méthode consiste à détecter les changements de modèles autorégressifs au travers des erreurs de prédiction calculées sur deux fenêtres d'analyse (cf. figure 3.6). La distance entre ces deux modèles est obtenue à partir de l'entropie mutuelle des deux lois conditionnelles correspondantes.

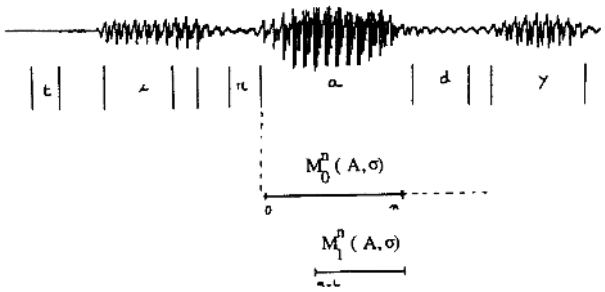

FIG. 3.6 – Localisation des fenêtres d'estimation des modèles M_0^n et M_1^n à l'instant n ; l'instant 0 correspond à la dernière frontière validée. La phrase prononcée est : « il se garan**tira du**... »

La distance entre ces deux modèles est dérivée de l'entropie mutuelle des deux lois conditionnelles correspondantes.

Définissons pour chaque instant k :
– le « passé » de l'échantillon x_k :

$$X_{k-1}^T = (x_1, \cdots, x_{k-1}) \tag{3.10}$$

– les deux lois conditionnelles associées respectivement aux modèles M_0^k and M_1^k :

$$g_0^k(x|X_{k-1}) \; and \; g_1^k(x|X_{k-1}) \tag{3.11}$$

L'entropie mutuelle est donnée par :

$$w_k = \int g_0^k(x|X_{k-1}) log \frac{g_1^k(x|X_{k-1})}{g_0^k(x|X_{k-1})} dy - log \frac{g_1^k(x_k|X_{k-1})}{g_0^k(x_k|X_{k-1})} \tag{3.12}$$

ce qui s'écrit dans le cas gaussien :

$$w_k = \frac{1}{2} \left\{ 2 \frac{e_k^0 e_k^1}{\sigma_1^2} - \left[1 + \frac{\sigma_0^2}{\sigma_1^2} \right] \frac{e_k^0}{\sigma_0^2} + \left[1 - \frac{\sigma_0^2}{\sigma_1^2} \right] \right\} \tag{3.13}$$

avec l'erreur de prédiction à l'instant k pour chacun des modèles M_i^k :

$$e_k^i = x_k - \sum_{j=1}^{p} a_j^i x_{k-j}, i = 0, 1. \tag{3.14}$$

Finalement, la statistique à l'instant n est définie comme une somme cumulée des entropies mutuelles calculées à tout instant k, $1 \leq k \leq n$:

$$W_n = \sum_{k=1}^{n} w_k \tag{3.15}$$

Le comportement de la somme cumulée est décrit par la courbe de la figure 3.7 : le problème consiste à détecter un changement de moyenne à l'instant de rupture r.

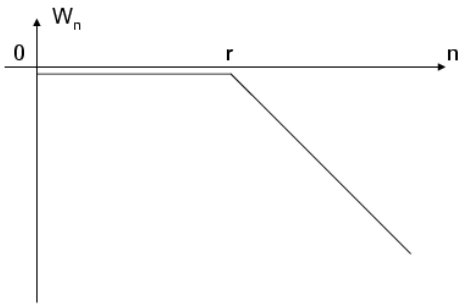

FIG. 3.7 – Variations de la somme cumulée W_n.

Cette méthode est comparée à de nombreuses autres méthodes de segmentation [AO93]. Nous pouvons citer, en particulier, le système Arial II [Cae79] où la segmentation est fondée sur les variations d'indices spectraux et les systèmes fondés sur la décomposition temporelle [Ata83] et [Bim88].

Elle a déjà fourni des résultats intéressants pour la reconnaissance automatique de la parole : des expériences ont montré que la durée des segments est porteuse d'une information pertinente [AO97].

Elle permet d'atteindre, pour la parole, une segmentation subphonétique où 3 types de segments se distinguent :

- les segments quasi stationnaires qui correspondent à la partie stable des phonèmes lorsqu'elle existe,
- les segments transitoires,
- les segments courts (environ 20 ms), révélateurs de gestes articulatoires rapides.

Leur longueur varie entre 20 et 100 ms pour la parole (cf. figure 3.8).

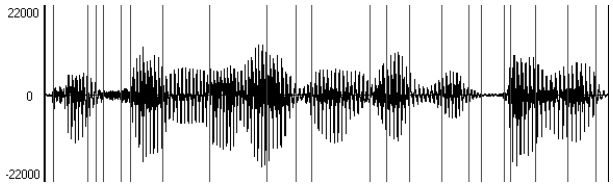

FIG. 3.8 – Résultat de la segmentation sur environ 1 seconde de parole. La phrase prononcée est : « Confirmez le rendez-vous par écrit ».

Pour la musique, un segment correspond à la tenue d'une note ; il peut être beaucoup plus long (cf. figure 3.9).

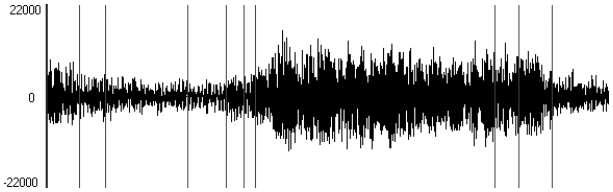

FIG. 3.9 – Résultat de la segmentation sur environ 1 seconde de musique d'un extrait de Mozart.

3.2.4.2 Paramètres

Les deux paramètres sont calculés après application de l'algorithme DFB.

– **Nombre de segments**

Le nombre de segments présents durant chaque seconde de signal est calculé. Les signaux de parole présentent une alternance de périodes de transition (voisées/non-voisées) et de périodes de relative stabilité (les voyelles en général) [Cal89]. Au niveau de la segmentation, cela se traduit par de nombreux changements. La musique, étant plus tonale (ou harmonique), ne présente pas de telles variations.

Le nombre de segments par unité de temps (ici la seconde) est donc plus important pour la parole (23 segments dans notre exemple, cf. figure 3.8) que pour la musique (10 segments dans notre exemple, cf. figure 3.9).

– **Durée des segments**

La durée des segments, obtenue après segmentation automatique (DFB), est fortement corrélée au nombre de segments par seconde. Afin de limiter la corrélation de ces deux paramètres de segmentation, la durée moyenne des segments sur une seconde est calculée sur les 7 segments les plus longs de la seconde. Le nombre de segments caractéristiques est fixé expérimentalement. Les segments sont généralement plus longs pour la musique (180 ms dans notre exemple, cf. figure 3.9) que pour la parole (80 ms dans notre exemple, cf. figure 3.8).

3.2.5 Récapitulatif des échelles de temps du système

Le fonctionnement de notre système global et la fonctionnalité de chaque paramètre reposent sur plusieurs échelles de temps que nous rappelons ici afin de faciliter la compréhension du module de décision.

– La **trame** est l'unité correspondant au découpage du signal (ici, toutes les 16 ms) en vue de calculer les paramètres de modulation de l'énergie à 4 Hertz et de modulation de l'entropie.

– Le **segment** est l'unité représentative de la segmentation par l'algorithme DFB. Il est de taille variable et les paramètres du sous-système de classification musique/non-musique l'utilisent.

– La **fenêtre** de décision est quant à elle représentative de la prise de décision. Dans notre système, celle-ci est de longueur fixe (1 seconde) quel que soit le paramètre utilisé.

La figure 3.10 représente ces trois échelles de temps sur un même signal.

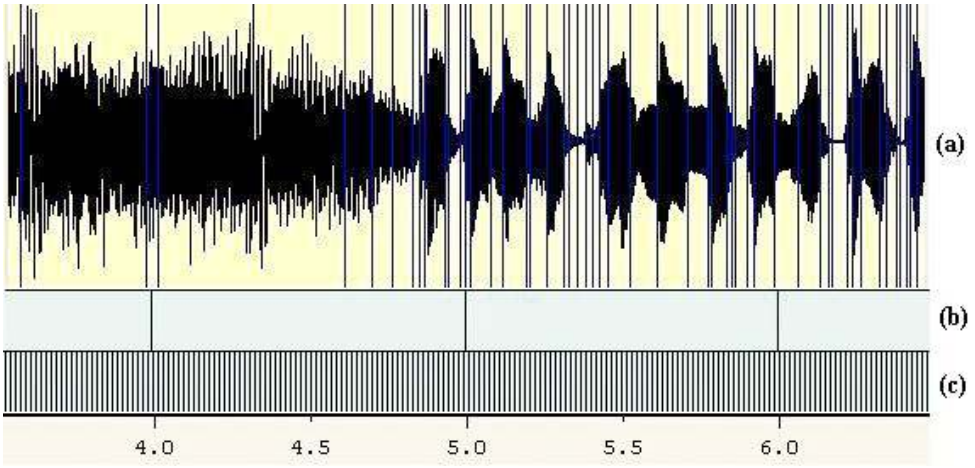

FIG. 3.10 – Représentation des échelles de temps de notre système : les segments de taille variable en (a), les fenêtres de décision (une par seconde) en (b) et les trames d'analyse (une toutes les 16 ms) en (c).

3.3 Étude des distributions des paramètres

Pour évaluer la pertinence des paramètres, nous avons étudié leur distribution sur un corpus de parole lue contenant 5 langues européennes, le corpus MULTEXT [Cam98]. Ce corpus est lu par 10 locuteurs par langue (5 hommes et 5 femmes). Les enregistrements sont de bonne qualité, le taux d'échantillonnage est de 20 kHz. Nous avons également créé un corpus d'extraits musicaux contenant différentes sortes de musique, allant du rock au classique avec un taux d'échantillonnage de 16 kHz, pour compléter cette pré-étude.

La durée totale pour chaque corpus (parole et musique) est d'environ 35 minutes (soit plus de 2000 segments d'une seconde).

3.3.1 Modulation de l'énergie à 4 Hertz

L'histogramme obtenu pour le paramètre de modulation de l'énergie à 4 Hertz pour la parole et la non-parole est représenté sur la figure 3.11.

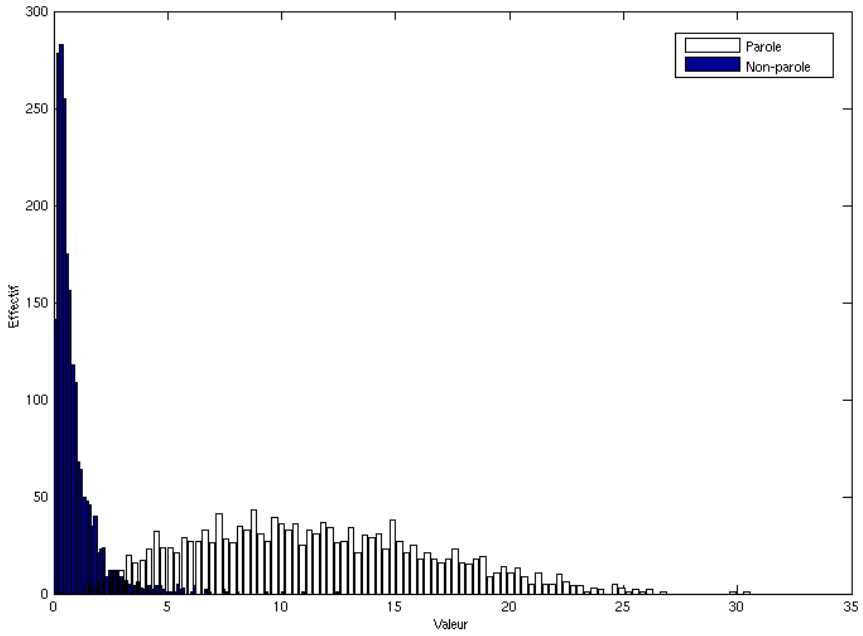

FIG. 3.11 – Distribution de la modulation de l'énergie à 4 Hertz par seconde pour la parole et pour la non-parole.

La parole et la non-parole sont clairement dissociées. L'intersection des deux histogrammes (modulation de l'énergie = 2,5) peut être utilisée comme seuil, dans le cadre d'une approche bayésienne fondée sur le maximum de vraisemblance.

En faisant l'hypothèse que le volume et la diversité des données sont suffisamment significatifs et que les distributions sont des lois gaussiennes, nous pouvons estimer les probabilités d'erreur :

$$\begin{cases} Pr(non\text{-}parole|parole) = Pr(non\text{-}parole > seuil) = 6,4\%. \\ Pr(parole|non\text{-}parole) = Pr(parole < seuil) = 3,2\%. \end{cases} \tag{3.16}$$

3.3.2 Modulation de l'entropie

La même expérience est reconduite avec le paramètre de modulation de l'entropie. Les résultats sont montrés sur la figure 3.12.

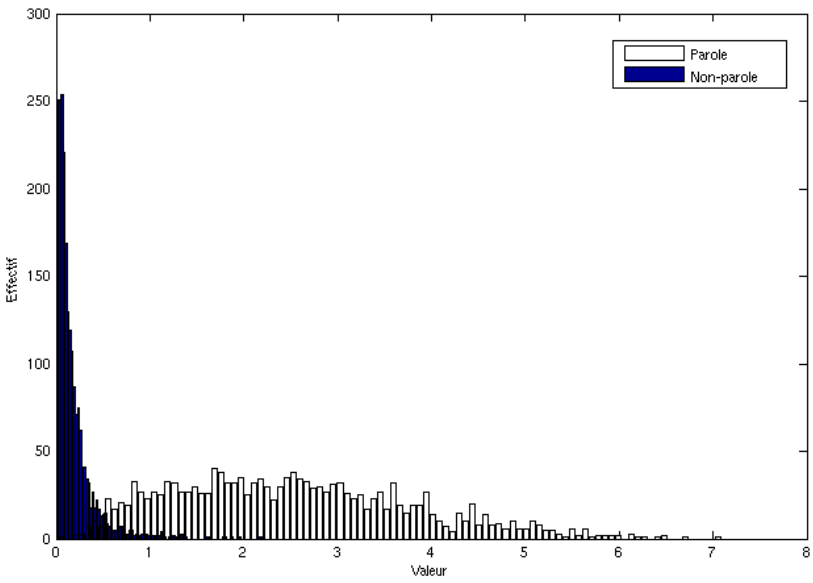

FIG. 3.12 – Distribution de la modulation de l'entropie par seconde pour la parole et pour la non-parole.

Ce paramètre est également pertinent dans la tâche de séparation parole/non-parole. Chaque histogramme est clairement séparé, et nous pouvons également déterminer un seuil expérimental (modulation de l'entropie = 0,5).

En faisant la même hypothèse que précédemment sur le volume, la diversité des données et les distributions probabilistes, nous pouvons également estimer les probabilités d'erreur :

$$\begin{cases} Pr(non\text{-}parole|parole) = Pr(non\text{-}parole > seuil) = 7,2\%. \\ Pr(parole|non\text{-}parole) = Pr(parole < seuil) = 3,4\%. \end{cases} \tag{3.17}$$

3.3.3 Paramètres de segmentation

3.3.3.1 Nombre de segments

La répartition du nombre de segments obtenus automatiquement est représentée sur la figure 3.13.

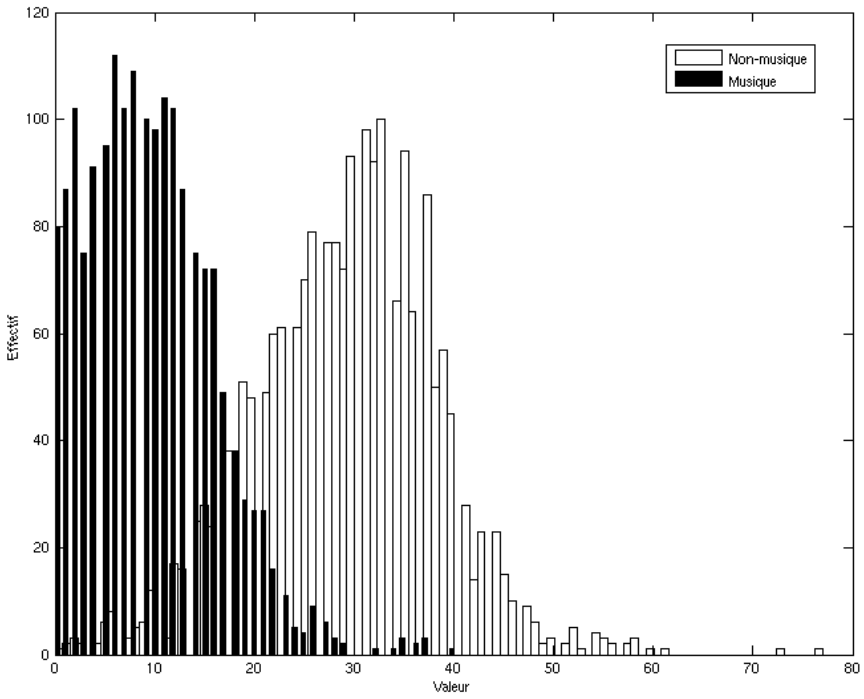

FIG. 3.13 – Distribution du nombre de segments par seconde pour la non-musique et pour la musique.

Les deux histogrammes montrent que ce paramètre est pertinent ; la non-musique et la musique peuvent être discriminées au moyen d'un simple seuillage (seuil = 17 segments par seconde).

Les probabilités d'erreur, sous les mêmes hypothèses que précédemment, sont :

$$\begin{cases} Pr(musique|non\text{-}musique) = Pr(musique > seuil) = 11,6\%. \\ Pr(non\text{-}musique|musique) = Pr(non\text{-}musique < seuil) = 3,6\%. \end{cases} \tag{3.18}$$

3.3.3.2 Durée des segments

Comme le montre la figure 3.14, l'hypothèse gaussienne pour modéliser ce paramètre est impossible : les deux histogrammes sont difficilement séparables.

Une étude statistique [Sua94] montre que la loi gaussienne inverse (loi de Wald) est une loi de probabilité qui modélise mieux la durée des sons que les lois gaussiennes ou gammas. C'est pourquoi la durée des segments est modélisée à l'aide d'une loi de Wald paramétrée par μ et λ.

Par définition, une variable aléatoire g suit une distribution inverse gaussienne si elle présente une fonction de densité de probabilité (pdf) de la forme [Joh70] :

$$\begin{cases} p(g) = \sqrt{\frac{\lambda}{2\pi g^3}} * e^{\frac{-\lambda(g-\mu)^2}{2\mu^2 g}}, & \text{si } g \geq 0 \\ p(g) = 0, & \text{sinon} \end{cases} \tag{3.19}$$

avec μ = valeur moyenne de g et $\frac{\mu^3}{\lambda}$ = variance de g.

La figure 3.14 décrit la répartition des durées des segments pour la non-musique et pour la musique. Les lois de Wald, correspondant aux distributions, sont tracées avec les paramètres estimés pour la non-musique et la musique.

Les paramètres λ et μ du modèle gaussien inverse ont été estimés :

$Parole$	$Musique$
$\lambda = 15,2753$	$\lambda = 50,6069$
$\mu = 30,1865$	$\mu = 74,9350$

Dans le cadre de la stratégie bayésienne, la décision est prise par maximum de vraisemblance en utilisant les deux lois ainsi estimées.

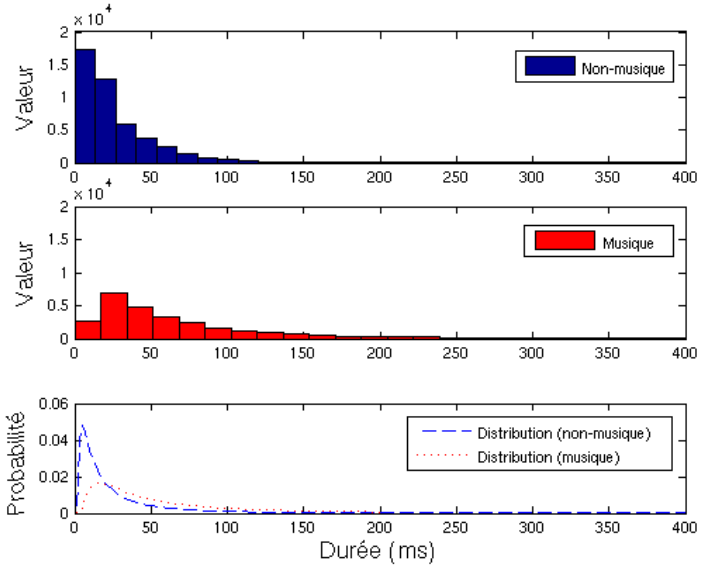

FIG. 3.14 – Répartition des durées des segments pour la non-musique et la musique ainsi que les lois de Wald correspondantes.

3.4 Expériences et évaluation

3.4.1 Corpus

Les expériences sont effectuées sur un corpus totalement différent de celui utilisé pour étudier la distribution des valeurs des paramètres et déterminer les seuils, afin d'évaluer la robustesse de nos paramètres et des seuils.

Le corpus expérimental correspond à une base de données qui est réalisée à partir d'enregistrements de RFI[8] (Radio France Internationale) et la fréquence d'échantillonnage est de 16 kHz. Cette base de données contient de longues périodes de parole, de musique, ainsi que des zones de chevauchement pouvant contenir de la parole, de la musique et/ou du bruit.

La parole est enregistrée dans différentes conditions (canal téléphonique, enregistrements en extérieur, bruit de foule, cocktail party...). Des locuteurs à fort accent sont présents : des africains francophones. Il y a également des traductions simultanées.

La musique est présente sous diverses formes également : de nombreux instruments sont représentés. Il y a également des parties de voix chantée. Le corpus est multilocuteur et multilingue (français, anglais et espagnol).

[8]Dans le cadre du projet RAIVES (projet « Sciences de l'information » CNRS).

Nous utilisons également ce corpus pour mettre en œuvre notre système de référence adapté (cf. chapitre 2). Pour l'adaptation des MMG, nous avons utilisé environ 2 heures de ce corpus. Les tests se font sur plus de 6 heures (différentes de l'apprentissage bien sûr !).

3.4.2 Étiquetage manuel

Les étiquetages manuels sont effectués au préalable comme dans la section 2.3.1, puis ramenés à la seconde, temps où sont prises les décisions. Cet étiquetage est fondé sur la classe majoritairement représentée sur chaque seconde.

La figure 3.15 présente un exemple d'alignement à la seconde obtenue pour une classification manuelle parole/non-parole (p/-).

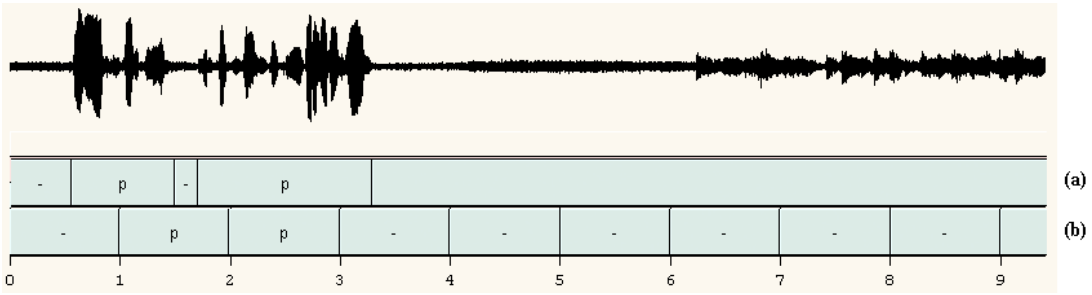

FIG. 3.15 – Exemple d'étiquetage manuel parole/non-parole (p/-) adapté au traitement par seconde, sur un extrait contenant de la parole, du bruit puis de la musique. (a) Étiquetage manuel (cf. section 2.3.1). (b) Étiquetage manuel adapté.

Les deux étiquetages à la seconde (parole et musique) ainsi obtenus serviront à la comparaison et l'évaluation des performances de chacun de nos paramètres et de notre système global.

3.4.3 Évaluation

Afin d'évaluer et comparer les performances de notre système, nous utilisons le « taux de classification correcte » décrit dans la section 2.4.3.1.

Évaluation isolée des paramètres

Les deux seuillages définis sur la modulation de l'énergie à 4 Hertz et la modulation de l'entropie montrent que ces paramètres sont très discriminants ; les taux de classification correcte sont similaires, de l'ordre de 87 % pour la classification parole/non-parole (cf. tableau 3.1).

TAB. 3.1 – *Classification parole/non-parole.*

Paramètres	Performances (Taux de classification correcte)
Modulation de l'énergie à 4 Hertz	87,3 %
Modulation de l'entropie	87,5 %
Fusion (détection de parole)	**90,5 %**

Le seuillage défini sur le nombre de segments issus de l'algorithme de divergence donne un taux supérieur à 86 % pour la détection de musique (cf. tableau 3.2). Le paramètre de durée des segments associé à la loi gaussienne inverse fournit un taux de classification correcte légèrement plus bas, 76 % pour de la musique/non-musique.

TAB. 3.2 – *Classification musique/non-musique.*

Paramètres	Performances (Taux de classification correcte)
Nombre de segments	86,4 %
Durée des segments	78,1 %
Fusion (détection de musique)	**89 %**

Un examen plus précis des résultats reflète bien le fait que le corpus est assez disparate. En effet, dans le cas des émissions d'informations, le taux de classification correcte atteint jusqu'à 98 % pour la parole alors que pour des reportages où un fond musical est présent durant toute l'émission, avec plus ou moins de variabilité, le taux peut baisser jusqu'à 84 % pour la parole et 79 % pour la musique.

Évaluation de la fusion par le maximum

Les résultats fournis par les différents paramètres ont été ensuite fusionnés :

- le premier sous-système, consacré à la détection de parole, résulte de la fusion de la modulation de l'énergie à 4 Hertz et de la modulation de l'entropie (cf. tableau 3.1). En prenant comme décision celle qui correspond au maximum des deux scores, le taux de classification correcte atteint 90,5 % (+ 3 %).

- le second sous-système, consacré à la détection de musique, est issu de la fusion des deux paramètres de segmentation (cf. tableau 3.2). Avec la même stratégie (maximisation des scores), le taux de classification correcte atteint ici 89 % (gain de 2,5 %).

Il est à souligner que, bien que dans chaque sous-système les deux paramètres soient fortement corrélés, l'utilisation conjointe des deux augmente la performance de décision.

Remarque : l'annexe B présente des résultats plus détaillés sur ce corpus et sur d'autres base de données.

Les résultats des deux sous-systèmes sont ensuite alignés et fusionnés (cf. figure 3.16).
L'étiquetage résultant possède quatre symboles :
– « P » : correspondant à de la parole et de la non-musique,
– « M » : pour de la musique et de la non-parole,
– « PM » : pour de la parole et de la musique,
– « - » : pour tout le reste (bruit et silence).

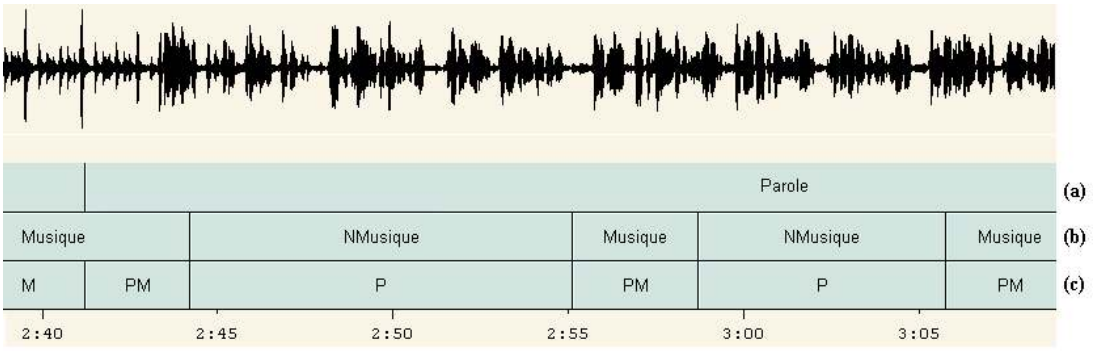

FIG. 3.16 – Exemple de résultats de notre système global pour une classification parole/non-parole (a) et musique/non-musique (b). La fusion est représentée sur la ligne (c) avec « P » pour parole, « M » pour musique et « PM » pour parole et musique.

3.4.4 Comparaison avec le système référence

Le système de référence (chapitre 2), dans sa configuration d'origine, est appris sur le corpus AGIR, composé d'émissions télévisuelles (des séries et du sport notamment). Nous avons vu que ses performances sur ce corpus sont de 93 % d'identification correcte pour la parole/non-parole et 91 % pour la musique/non-musique pour une prise de décision toutes les secondes.

Dans la mesure où notre système global et notre système de référence sont tous deux divisés en deux sous-systèmes parole/non-parole et musique/non-musique, nous pouvons comparer les sous-systèmes un à un. Nous avons utilisé le système de référence de deux manières : sans

adaptation et avec adaptation sur le corpus RFI. L'ensemble des résultats sont rassemblés dans les tableaux 3.3 et 3.4.

– Pour la détection de parole, le résultat du système global (90,5 %) est équivalent à celui obtenu par le système de référence (90,9 %) avec un apprentissage (ou une adaptation) sur le corpus RFI. Cela permet de montrer la pertinence du choix de nos paramètres. Dans les même conditions, c'est-à-dire avec un apprentissage sur des données différentes (système de référence avec sa configuration d'origine), notre système est plus performant que le système de référence (86,1 %). La modélisation statistique de nos paramètres est validée.

TAB. 3.3 – *Comparaison de notre système global de fusion avec le système de référence pour la détection de parole.*

Détection de parole	Performances (Taux de classification correcte)
Système global (fusion)	**90,5 %**
Système de référence	86,1 %
Système de référence avec adaptation sur RFI	90,9 %

– Pour la détection de musique, le résultat du système global (89 %) est supérieur au système de référence qu'il soit réestimé (87 %) ou non (79,7 %) sur le corpus RFI. Cela nous confirme dans le choix de nos paramètres et dans la fusion utilisée par notre système global.

TAB. 3.4 – *Comparaison de notre système global de fusion avec le système de référence pour la détection de musique.*

Détection de musique	Performances (Taux de classification correcte)
Système global (fusion)	**89 %**
Système de référence	79,7 %
Système de référence avec adaptation sur RFI	87 %

Nous avons étudié la complémentarité qui pouvait exister entre notre système global et le système de référence adapté [Pin03a]. Pour cela, nous avons décidé de fusionner les deux systèmes (toujours par maximisation des scores de vraisemblance). Ainsi, pour la détection de parole, les coefficients cepstraux, la modulation de l'entropie et la modulation de l'énergie à 4 Hertz sont utilisés et pour la détection de musique, les coefficients spectraux et les deux paramètres issus de la segmentation automatique sont combinés (cf. figure 3.17).

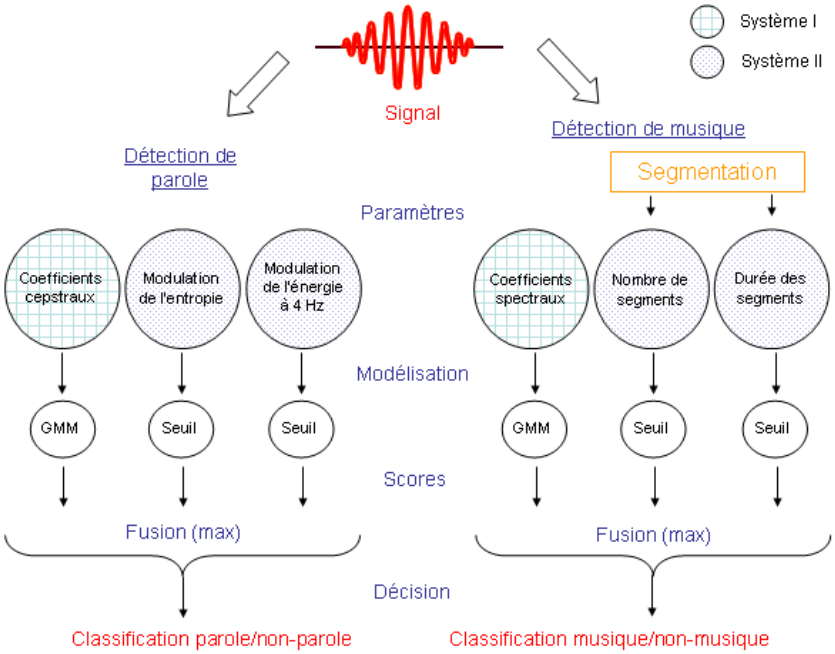

FIG. 3.17 – Schéma de fusion des deux systèmes de classification parole/musique.

Chacun des sous-systèmes (parole et musique) voit son taux de classification correcte augmenter d'environ 3 points (cf. tableau 3.5). Ainsi, la détection de parole atteint 93,9 % et la détection de musique 89,8 %.

TAB. 3.5 – *Fusion des deux systèmes de classification parole/musique.*

Détection	Performances (Taux de classification correcte)
Parole	
Système global	90,5 %
Système de référence adapté	90,9 %
Fusion	**93,9 %**
Musique	
Système global	89 %
Système de référence adapté	87 %
Fusion	**89,8 %**

Il apparaît donc très intéressant d'améliorer un système classique « lourd » (par son volume d'apprentissage et d'étiquetage), fondé sur une analyse spectrale et des MMG, par des paramètres « simples » et robustes.

3.5 Fusion de données

Pour chacun de nos sous-systèmes, nous avons adopté comme stratégie de fusion, la maximisation des scores associés à chaque paramètre, technique tout à fait primaire. Afin de valider cette méthode, nous avons étudié d'autres approches empruntées à la théorie des probabilités et la théorie de l'évidence[9].

3.5.1 Introduction

La fusion d'informations suscite un vif intérêt dans la communauté scientifique depuis quelques années [Dub94] et elle apparaît dans le domaine du traitement de la parole [Bes98] et [Mor00]. Elle consiste à mettre à profit le maximum d'informations sur les données afin de réduire les faiblesses de certaines à l'aide des avantages des autres. Les techniques de fusion, en particulier celles de la théorie des probabilités et de la théorie de l'évidence, permettent de gérer l'incertitude, la redondance, les conflits et les incohérences entre les sources d'information encore appelées experts.

Plus précisément, nous appellerons par la suite « expert » un système de classification et « classe » une réponse de l'expert.

Dans le cadre de cette étude, nous disposons de quatre experts primaires :
- l'expert « Mod4Hz » associé aux classes parole et non-parole,
- l'expert « Entropie » associé aux classes parole et non-parole,
- l'expert « NbSeg » associé aux classes musique et non-musique,
- l'expert « DureeSeg » associé aux classes musique et non-musique.

Nous étudions la fusion entre les experts « Mod4Hz » et « Entropie » et la fusion entre les experts « NbSeg » et « DureeSeg ».

3.5.2 Théorie des probabilités

Les méthodes les plus utilisées pour la fusion de décision ont tout d'abord été envisagées sous un angle probabiliste. Les informations qui doivent être fusionnées, délivrées par des experts, sont des vraisemblances issues de la modélisation des observations par des distributions de probabilités conditionnelles. L'un des inconvénients majeurs de l'approche probabiliste réside dans l'exigence de la connaissance parfaite des probabilités, et plus particulièrement de la

[9] Ce travail a été réalisé en collaboration avec Julie Mauclair lors de son stage de DEA [Mau03].

probabilité a priori des classes.

Pour pallier à ce problème de l'ignorance, on peut à l'issue d'une étape de développement de l'expert, étudier le risque conditionnel d'une prise de décision en s'appuyant sur des indices de confiance relative à l'expert et la classe [Ler00].

Afin de développer cette technique, notons :

- α : l'indice de confiance de l'expert dans sa décision, de manière générale,
- β : l'indice de confiance de classe qui traduit en quelque sorte l'expérience que l'on a de l'expert.

L'expert sait discriminer l'ensemble des classes avec un taux de confiance α_e :

$$\alpha_e = 1 - taux\ d'erreur \tag{3.20}$$

Dans le cas de nos quatre experts, les deux classes distinguées sont la classe C (parole ou musique) et la non-classe NC (non-parole ou non-musique).

On en déduit que (cf. équation 3.16) :

$$\alpha_e = 1 - (P(NC|C) + P(C|NC)) \tag{3.21}$$

Pour chacun de nos quatre experts, les expériences menées sur un ensemble de développement donnent accès à une estimation de la matrice de confusions $\beta_{e.}$, dont on peut extraire de la diagonale, les indices de confiance des classes :

$$\begin{aligned} \beta_{eC} &= \frac{P(C|C)}{P(C|C)+P(NC|C)} \\ \beta_{eNC} &= \frac{P(NC|NC)}{P(C|NC)+P(NC|NC)} \end{aligned} \tag{3.22}$$

Les indices de classe (β_{eC},β_{eNC}) permettent de définir une fonction de coût l entre décision et classe :

$$\begin{aligned} &l(classe\ attendue, classe\ trouve) = 0, \\ &l(classe\ attendue, non\text{-}classe\ trouve) = 1 - \beta_{eNC}, \\ &l(non\text{-}classe\ attendue, classe\ trouve) = 1 - \beta_{eC}, \\ &l(non\text{-}classe\ attendue, non\text{-}classe\ trouve) = 0. \end{aligned} \tag{3.23}$$

Le calcul du risque conditionnel nous donne la fonction de décision modifiée pour chaque expert :

$$s_e^*(y) = min\left\{\left\{(1-\beta_{NC}) * \frac{P(y|C)}{P(y)}\right\}, \left\{(1-\beta_C) * \frac{P(y|NC)}{P(y)}\right\}\right\} \tag{3.24}$$

où y est l'observation extraite toutes les secondes.

Les probabilités conditionnelles des observations restent issues des histogrammes approximés par des lois gaussiennes.

Finalement, la décision issue de la fusion des experts sera celle prise avec l'expert e qui correspond au maximum de la formule :

$$\alpha_e * (1 - s_e^*(y)) \tag{3.25}$$

En conclusion, l'expert qui a le meilleur taux de confiance évalué comme le produit de son indice d'expert avec le complémentaire du risque minimum qu'il prend, donne la décision finale.

3.5.3 Théorie de l'évidence

La plupart des techniques de fusion traite l'imprécision, mais des notions telles que l'incertitude ou encore la fiabilité sont ignorées. La théorie de l'évidence permet de modéliser et d'utiliser des données incertaines [Jan96] et [Alt03].

Supposons qu'un système global de décision puisse prendre N décisions et qu'il résulte de la fusion de plusieurs experts ; soit θ cet ensemble.

Dans notre cadre expérimental, l'ensemble θ est constitué de $N = 4$ classes :

$$\theta = \{P, M, PM, B\} \tag{3.26}$$

avec P pour parole, M pour musique et B pour bruit (en fait, ni parole, ni musique).

L'ensemble des parties de θ est donné par :

$$2^\theta = \{A|A \subseteq \theta\} = \{\emptyset, \{P\}, \{M\}, \{PM\}, \{B\}, \{P \cup M\}, ..., \theta\} \tag{3.27}$$

Il sert de référentiel de définition pour l'ensemble des grandeurs utilisées par la théorie de l'évidence pour évaluer la véracité d'une proposition.

Toute information provenant de n'importe quelle source (capteur, agent, expert...) traduit une opinion sur l'état d'un système global. Cette information porte non seulement sur les hypothèses singletons, mais aussi sur les conjonctions ou les disjonctions de celles-ci ; elle s'apparente à un des éléments de 2^θ.

L'opinion sur le système est alors caractérisée par des degrés de croyance dans les différentes hypothèses. Ces degrés de croyance peuvent être définis par une fonction notée m_θ. La fonction de croyance m_θ est définie par :

$$m_\theta \ : \ 2^\theta \rightarrow [0,1] \tag{3.28}$$

et vérifie les propriétés suivantes :

$$\begin{array}{ll} 1. & m_\theta(\emptyset) = 0 \\ 2. & \sum_{A \subseteq \theta} m_\theta(A) = 1 \end{array} \tag{3.29}$$

La modélisation issue de cette fonction est appelée jeu de masses. Elle consiste à répartir toute la connaissance disponible sur l'ensemble 2^θ. $m_\theta(A)$ représente la partie du degré de croyance placée exactement sur la proposition A. $m(\theta)$ représente la masse associée à l'ignorance.

Plusieurs jeux de masses peuvent être définis : en général un jeu de masse par expert ; ils sont ensuite fusionnés grâce à la loi de Dempster-Shafer [Jan96] pour construire un jeu de masse final unique et ainsi accéder à une information plus fiable.

Dans notre cas, à partir des quatre experts qui correspondent aux quatre paramètres de notre système de classification, quatre jeux de masses (m_1, m_2, m_3, m_4) sont définis. $m_e(\theta)$ représente alors l'erreur de l'expert. Pour les autres hypothèses, chaque expert fournit un jeu de masses qui provient des scores de vraisemblances.

Les différents jeux de masses sont fusionnés avec la loi de Dempster-Shafer afin d'obtenir un jeu de masses global. On peut ainsi fusionner les experts par deux, l'expert « Mod4Hz » avec l'expert « Entropie » et l'expert « NbSeg » avec l'expert « DureeSeg » :

$$\begin{aligned} m_{Mod4Hz,Entropie} &= m_{Mod4Hz} \oplus m_{Entropie}, \\ m_{NbSeg,DureeSeg} &= m_{NbSeg} \oplus m_{DureeSeg}, \end{aligned} \tag{3.30}$$

Les jeux de masses intermédiaires sont de la forme :

$$m_{Mod4Hz,Entropie}(A) = \sum_{A_i \cup B_j = A} m_{Mod4Hz}(A_i) * m_{Entropie}(B_j) \tag{3.31}$$

où A_i et B_j sont des éléments focaux respectivement du premier et du deuxième expert, définis sur 2^θ.

La décision est effectuée sur le maximum de plausibilité pour tenir compte du poids des disjonctions d'hypothèses qui apparaissent dans le jeu de masses final. La plausibilité se définit comme une somme de degrés de croyance :

$$Pl_\theta(A) = \sum_{B \cap A \neq \emptyset} m_\theta(B) \tag{3.32}$$

La théorie de l'évidence permet de gérer des hypothèses composées et de réduire la part d'a priori au sein de la modélisation du problème.

3.5.4 Expériences

Cadre expérimental

– L'apprentissage des indices de confiance pour la théorie des probabilités et celui des jeux de masse pour la théorie de l'évidence s'effectue sur le corpus ayant servi à l'apprentissage des MMG du système de référence adapté (cf. section 3.4.1). Il s'agit de 2 heures de RFI.

– La reconnaissance s'effectue toujours sur les 6 heures de RFI afin que les résultats entre les différents systèmes soient comparables.

Pour la **théorie des probabilités**, les indices de confiance des experts sont :
- $\alpha_{Mod4Hz} = 1 - 9,6 = 90,4\%$,
- $\alpha_{Entropie} = 1 - 10,6 = 89,4\%$,
- $\alpha_{NbSeg} = 1 - 15,2 = 84,8\%$,
- $\alpha_{DureeSeg} = 1 - 46,2 = 53,8\%$.

Les indices de confiance de classe sont obtenus à partir de la matrice de confusion classe/non-classe. Le tableau 3.6 représente la matrice de confusion parole/non-parole dans le cas de l'expert « Mod4Hz » :

TAB. 3.6 – *Matrice de confusion parole/non-parole pour l'expert « Mod 4Hz ».*

	Parole	Non-parole
Parole	86,8 %	13,2 %
Non-parole	21,8 %	78,2 %

Nous avons ainsi : $\beta_{Mod4Hz,P} = \frac{86,8}{86,8+21,8} = 79,9\%$ et $\beta_{Mod4Hz,NP} = \frac{78,2}{78,2+13,2} = 85,5\%$.

En faisant de même avec les autres experts, nous obtenons :
- $\beta_{Entropie,P} = 79,7\%$ et $\beta_{Entropie,NP} = 88,8\%$,
- $\beta_{NbSeg,M} = 75,2\%$ et $\beta_{NbSeg,NM} = 67,7\%$,
- $\beta_{DureeSeg,M} = 89,3\%$ et $\beta_{DureeSeg,NM} = 65,7\%$.

Pour la **théorie de l'évidence**, les jeux de masse associés à l'ignorance sont :
- $m_{Mod4Hz}(\theta) = 9,6\%$,
- $m_{Entropie}(\theta) = 10,6\%$,
- $m_{NbSeg}(\theta) = 15,2\%$,
- $m_{DureeSeg}(\theta) = 46,2\%$.

Résultats

La fusion par approche probabiliste offre des résultats similaires au système de fusion fondé sur le maximum des scores. En effet, cette théorie est sous-jacente dans ce dernier et la modélisation ne peut que conforter les résultats obtenus.

La pondération des indices de confiance apporte une augmentation pour la discrimination parole/non-parole (tableau 3.7).

TAB. 3.7 – *Comparaison des méthodes de fusion pour la détection de parole.*

Méthode de fusion	Performances (Taux de classification correcte)
Système de fusion (max)	*90,5 %*
Théorie des probabilités	90,7 %
Théorie de l'évidence	**90,9 %**

Les résultats musique/non-musique (tableau 3.8) dégradés peuvent s'expliquer par le fait que la durée des segments n'a pas un taux de confiance d'expert suffisamment élevé ($\alpha_4 = 53\,\%$) et que ce paramètre ne rentre donc pas en compte dans la fusion finale. Celle-ci repose uniquement sur la décision du paramètre « nombre de segments ».

TAB. 3.8 – *Comparaison des méthodes de fusion pour la détection de musique.*

Méthode de fusion	Performances (Taux de classification correcte)
Système de fusion (max)	***89 %***
Théorie des probabilités	84,8 %
Théorie de l'évidence	86,9 %

Pour la théorie de l'évidence, les jeux de masse offrent une amélioration des résultats. Le score obtenu en musique/non-musique s'explique de la même façon que pour la théorie précédente.

3.6 Conclusion

Nous avons présenté dans cette section quatre paramètres fondés sur différentes propriétés du signal. Chaque paramètre est pertinent dans le sens où il permet, en lui-même, de faire une discrimination parole/non-parole ou musique/non-musique correcte. En considérant chaque paramètre individuellement, le taux de classification correcte varie d'environ 78 % pour la durée des segments à plus de 87 % pour la modulation de l'entropie.

La fusion entre ces paramètres par maximisation des scores de vraisemblances permet d'améliorer les résultats et d'obtenir environ 90 % de reconnaissance correcte pour chacune des classifications (parole/non-parole et musique/non-musique).

Les résultats obtenus au cours de cette série d'expériences sont meilleurs que ceux obtenus avec l'approche classique fondée sur des Modèles de Mélanges de lois Gaussiennes (MMG). L'avantage principal de cette nouvelle méthode est qu'elle est utilisable sur tout nouveau corpus sans requérir de nouvel étiquetage : il n'y a plus de phase d'apprentissage et/ou d'adaptation des modèles alors que les MMG doivent être systématiquement adaptés pour donner des résultats comparables.

La théorie des probabilités est un cadre satisfaisant pour fusionner les informations des experts de la détection parole/non-parole. Des résultats corrects sont obtenus pour la détection de parole, ce qui valide la modélisation de ce sous-système. Pour la détection de musique, les scores sont inférieurs à ceux du paramètre « nombre de segments » seul, ce qui semble montrer que la durée des segments n'est pas forcément le meilleur soutien pour celui-ci.

La théorie de l'évidence nous conduit aux meilleurs résultats en détection de parole et les résultats en détection de musique restent stables.

Le corpus RFI est assez difficile car très varié : la parole est présente sous diverses conditions, de pure à très bruitée. La qualité des résultats permet de montrer non seulement la robustesse de nos paramètres, mais aussi l'intérêt de notre approche quant à son utilisation sur n'importe quel type de document sonore.

Deuxième partie

Les sons clés

Chapitre 4

Les jingles

Sommaire

4.1 Introduction

4.1.1 Problématique

Au delà du partitionnement primaire d'un document audiovisuel en parole/musique/autre, il est intéressant de détecter des sons clés ou des jingles représentant le début et/ou la fin d'un segment sonore afin de structurer le flux audio-visuel [Car00a]. Il ne s'agit pas de rechercher des thèmes [Ama01], mais plutôt de proposer une macrosegmentation de l'audio en trouvant sa structure temporelle. Cette approche trouve une application directe dans la recherche d'une segmentation automatique des programmes télévisés ou radiophoniques.

La section « reconnaissance de sons » du document de spécifications de MPEG7 [ANS01] propose une liste d'effets sonores classés en catégories afin de décrire les documents sonores. Les sons clés de référence sont répertoriés dans un tableau dynamique : structurer un document sonore signifie en particulier, détecter et localiser les occurrences de ces sons clés.

Notre étude se situe dans ce cadre scientifique. Un son clé peut être un jingle, c'est-à-dire un extrait sonore caractéristique, servant à introduire, conclure ou distinguer les différentes parties d'une émission : il doit attirer l'attention de l'auditeur (ou spectateur) et lui donner des repères temporels.

La détection d'un jingle ne peut pas se faire en le comparant directement au signal car le coût de calcul serait alors trop important. Nous proposons une approche par analyse du signal de manière à réduire le nombre d'observations à traiter et lisser l'information pour obtenir un pas d'évolution assez important.

Ce chapitre, consacré à la détection de jingles, est divisé en deux sections. La première présente le système global qui permet de détecter et d'identifier un jingle présent dans le flux sonore à condition que celui-ci fasse partie de notre catalogue de sons clés. La seconde section décrit les expériences effectuées sur des documents audiovisuels et radiophoniques.

4.1.2 Le jingle

Un jingle est un extrait sonore qui dure généralement quelques secondes. Il a pour but de présenter le début ou la fin d'une émission (météo, journal, publicité...) ou d'attirer l'attention de l'auditeur. Celui-ci a la particularité de pouvoir aussi bien contenir de la musique que de la parole. Il est, de plus, généralement redondant dans une collection de documents audiovisuels. La figure 4.1 présente un signal correspondant à un jingle composé uniquement de musique ainsi que son spectrogramme associé.

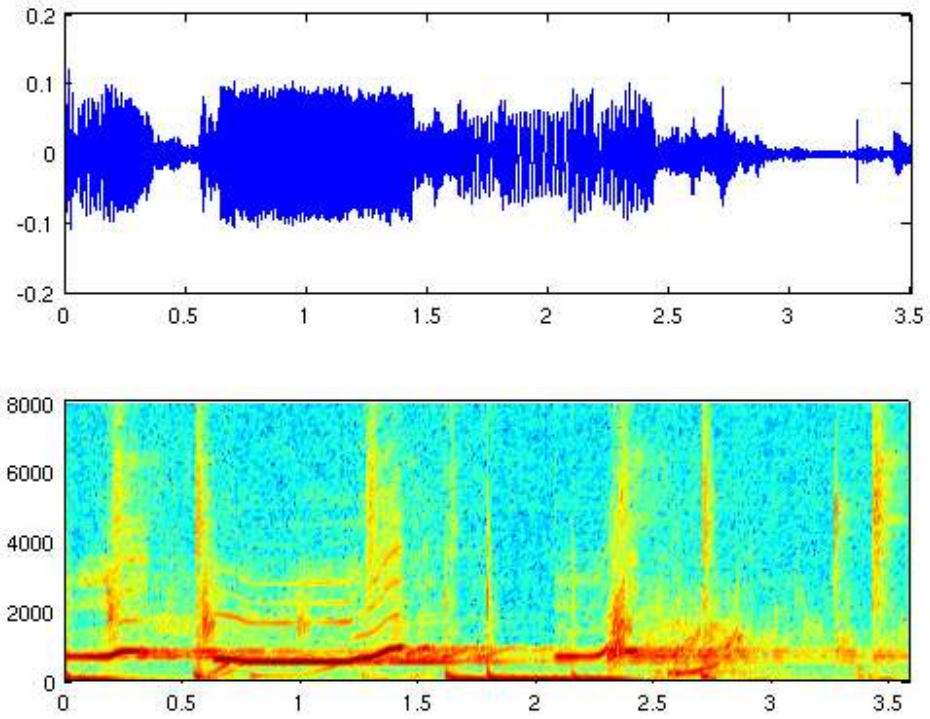

FIG. 4.1 – Signal et spectrogramme d'un jingle d'environ 3,5 secondes comprenant uniquement de la musique.

Les différentes occurences d'un jingle ont des longueurs très peu variables. Cette propriété nous conduit à proposer comme méthode de détection une stratégie de type appariement de motifs (« template matching »), qui consiste à rechercher dans le flux sonore une reproduction, plus ou moins exacte, d'un **jingle de référence**, occurrence sélectionnée préalablement.

4.2 Le système de détection de jingle

Le système de détection d'un jingle [Pin04a] est divisé en trois modules classiquement utilisés dans un problème de reconnaissance de formes (cf. figure 4.2) :

- un pré-traitement acoustique permet de caractériser au mieux le signal par une suite de vecteurs afin de comparer ceux-ci à la **signature** du jingle recherché,
- une phase de détection propose des candidats potentiels issus de la comparaison par mesure de dissimilarité,
- une identification confirme ou annule le choix des candidats grâce à un ensemble d'heuristiques.

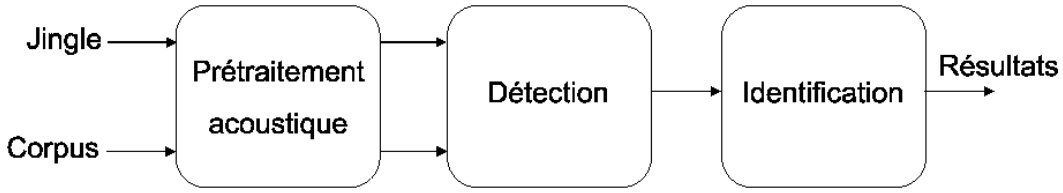

FIG. 4.2 – Le système global de détection et d'identification d'un jingle.

4.2.1 Pré-traitement acoustique

Le pré-traitement acoustique est fondé sur une analyse spectrale (cf. figure 4.3).

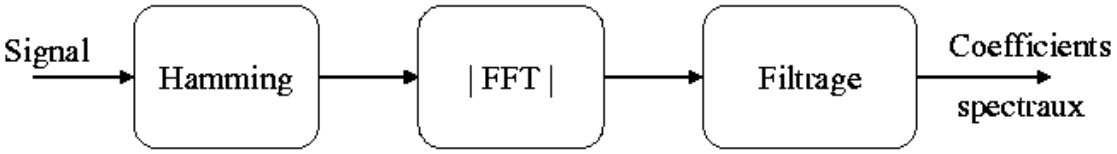

FIG. 4.3 – Extraction des paramètres par analyse spectrale.

Le signal est découpé en trames de 32 ms avec recouvrement sur la moitié de la trame. Le traitement de chaque trame est analogue à celui effectué pour la détection musique/non-musique (cf. section 2.2.1.2).

Rappelons que pour chaque trame d'analyse, une accentuation des aigus et un calcul du fenêtrage sont effectués (Hamming). Après le module de la transformée de Fourier (FFT), sont calculées les énergies dans 28 filtres triangulaires, répartis sur la plage de fréquences [100 Hz - 8000 Hz].

Afin de ne pas tenir compte du facteur bruit/intensité qui peut varier au cours du temps ou des enregistrements, le spectre de l'énergie est normalisé par l'énergie moyenne dans chacun des canaux. 28 coefficients spectraux sont extraits toutes les 16 ms.

4.2.2 Détection

Un jingle de référence est caractérisé par une suite de N vecteurs spectraux que nous appelons « signature » du jingle. Cette valeur N correspond au nombre de fenêtres d'analyse obtenues sur la durée totale du jingle considéré. Ce nombre dépend du jingle.

La détection consiste à trouver cette séquence (suite de vecteurs) dans le flux de données à analyser.

La distance Euclidienne est utilisée afin de comparer la signature du jingle et le signal représenté par une suite de vecteurs spectraux. Cette comparaison s'effectue donc sur une fenêtre glissante de vecteurs que l'on déplace par un pas de S vecteurs (cf. figure 4.4).

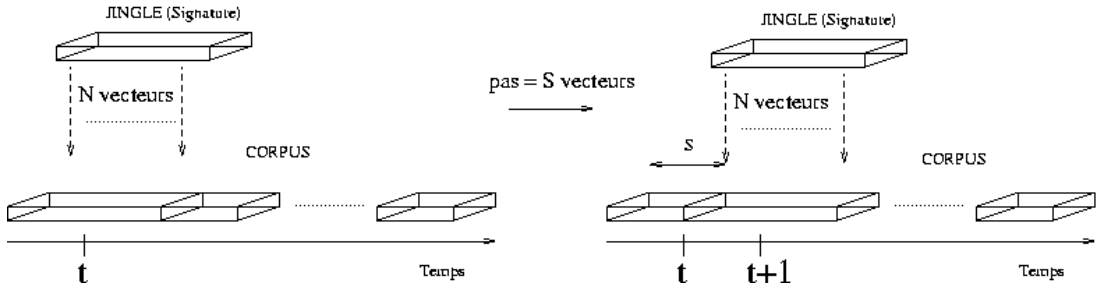

FIG. 4.4 – Comparaison entre le jingle et le corpus par distance Euclidienne.

Exemple : Prenons un corpus échantillonné à 16 kHz et un jingle de durée 3 secondes. Avec des trames de 32 ms de longueur, c'est-à-dire 512 échantillons, et un recouvrement de 16 ms, c'est-à-dire 256 échantillons, nous avons environ 60 vecteurs de coefficients spectraux pour chaque seconde de signal. La signature du jingle est alors composée de $N = 180$ vecteurs spectraux. Si le pas S est de 60 vecteurs, environ 1 s, nous aurons ainsi 180 comparaisons sur 3 minutes de flux sonore.

Les candidats potentiels correspondent à certains minima locaux de la distance signature/flux calculée :

- si la distance courante est inférieure à la moitié de la moyenne de la distance (moyenne calculée sur 120 s), notée M (cf. figure 4.6), cette distance est considérée comme une valeur minimale,
- seuls les minima locaux, correspondant à des valeurs minimales, sont détectés comme des jingles potentiels (cf. figure 4.5).

4.2.3 Identification

Un exemple de calcul de la distance Euclidienne entre la signature d'un jingle de référence et un signal est représenté sur la figure 4.5.

Nous pouvons observer cinq minima détectés (cf. section 4.2.2). Les deux premiers correspondent à un « bon » jingle : il s'agit d'une occurrence du jingle recherché. Les trois autres minima sont des jingles mais ne correspondent pas au jingle recherché. Ils ressemblent fortement aux deux premiers car les sons (notes de musique) qui composent ces jingles sont les mêmes mais ils sont joués dans un ordre différent.

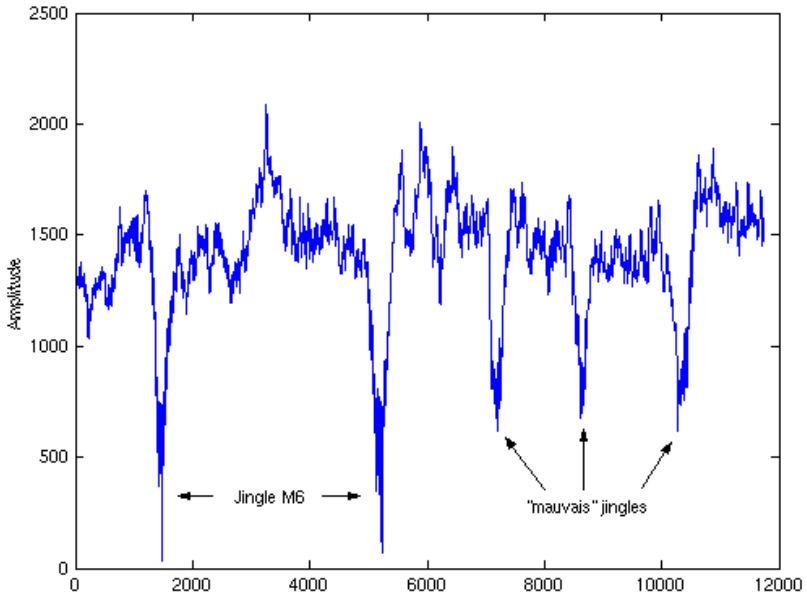

FIG. 4.5 – Distance Euclidienne lors de la détection du « jingle M6 » sur 3 minutes du « corpus M6 ».

Afin de sélectionner ces « bons » jingles, nous proposons le processus suivant. Suite à l'observation, *tous les minima correspondant au jingle de référence, ont une particularité commune ; la distance décroît très rapidement et la courbe présente, sans exception, un pic très fin* (cf. figure 4.5), l'analyse proposée consiste à étudier la largeur des pics de chacun des minima locaux.

Pour cela, nous introduisons les variables suivantes (cf. figure 4.6) :
– h la valeur du minimum local,
– H la hauteur relative à h. $H = \alpha.h$ ($\alpha > 1$),
– L la largeur du pic à la hauteur H.

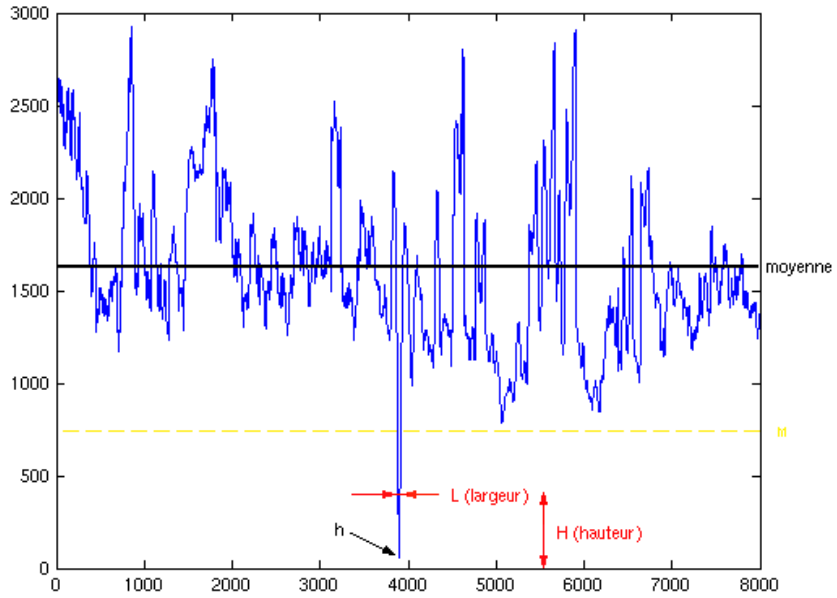

FIG. 4.6 – Identification des « bons » jingles par analyse de chacun des pics correspondant aux minima locaux détectés.

Si $L < \lambda$, le pic est déclaré « fin » et le minimum local est considéré comme un « bon » jingle. Sinon, le candidat est rejeté (« mauvais » jingle).

Les paramètres S, α et λ sont appris expérimentalement (cf. section 4.3.2). À noter que le seuil λ dépend du jingle recherché. Nous l'avons choisi proportionnel à N afin de pouvoir le rendre indépendant de la nature du jingle.

4.3 Expériences

Afin de rendre réaliste cette recherche de jingles, nous avons élaboré un **catalogue de jingles de référence** à rechercher simultanément dans un flux sonore.

L'étape la plus longue de notre système étant la paramétrisation du corpus par le calcul de la FFT, le nombre de jingles dans le catalogue n'est pas pénalisant. En effet, tout nouvel ajout d'un jingle ne nécessite qu'une comparaison de plus entre la paramétrisation du signal, déjà calculée, et la signature du nouveau jingle. Ces calculs peuvent d'ailleurs se faire en parallèle.

4.3.1 Corpus

Notre base de données est composée de six corpora différents (cf. tableau 4.1). La durée totale est d'environ 10 heures. Cette base est échantillonnée à 16 kHz.

TAB. 4.1 – *Description de la base de données.*

Corpus	Durée	Jingles	Occurrences
France 3	15 min	1	4
M6	15 min	1	16
Canal+	30 min	1	6
France Info	60 min	1	12
RFI	360 min	3	60
Publicités	90 min	25	34
Total	**570 min**	**32**	**132**

– Le corpus « France Info » est composé d'émissions radiophoniques : des actualités, des reportages, du sport et des bulletins météorologiques. Quelques chansons et des publicités sont présentes. Ce corpus a une durée d'une heure ; un seul jingle est recherché et il apparaît à douze reprises.

– Le corpus « France 3 » est un court extrait télévisuel avec deux chansons et diverses publicités ; l'ensemble dure quinze minutes et le jingle « France 3 » sélectionné est présent quatre fois.

– Les corpora « Canal+ » et « M6 » correspondent à des journaux d'informations de la télévision française. Le premier possède une durée de trente minutes et six occurrences du jingle « Canal+ ». Le second, composé de l'émission « 6 minutes », est plus court (quinze minutes), mais le jingle « M6 » est très présent (seize fois).

– Le corpus « RFI » est multilingue (français, anglais et espagnol) avec une majorité d'interviews, de reportages et d'informations de Radio France Internationale (RFI). Ce corpus est assez long (six heures environ) : trois jingles de référence différents, soixante occurrences sont à retrouver.

– Le dernier corpus est une compilation de plusieurs publicités télévisuelles et radiophoniques. Il s'agit ici de repérer une publicité à travers son jingle de référence. Nous avons vingt cinq jingles différents qui représentent trente quatre publicités.

La durée d'un jingle varie de une à quatre secondes dans notre base de données. Les jingles de référence sont une sélection des jingles présents dans les corpora de tests. Notre catalogue de sons clés est donc composé de 32 jingles différents représentatifs soit du début (ou de la fin) d'une émission (corpus « France 3 », « M6 », « Canal+ », « France Info » et « RFI »), soit d'une publicité (corpus « Publicités »). Aucune information sur l'appartenance d'un jingle à un corpus n'est conservée.

Plus de 200 jingles apparaissent dans notre base de données. Notre but est de détecter et d'identifier seulement les jingles similaires à ceux de notre catalogue de sons clés. Les jingles recherchés pourront être :

– superposés à de la parole si le présentateur parle durant le jingle,
– plus longs que le jingle de référence (jusqu'à deux fois),
– plus courts que le jingle de référence (jusqu'à la moitié).

Finalement, nous avons 132 jingles à retrouver et reconnaître parmi les 200 présents dans la base de données, sachant que notre catalogue est composé de 32 jingles de référence. Il en reste 68 (200-132) afin de vérifier qu'il n'y a pas de détections de jingles hors catalogue.

4.3.2 Apprentissage

Afin d'implémenter notre méthode d'identification, nous devons tout d'abord fixer les paramètres S, H et λ, présentés dans les sections 4.2.2 et 4.2.3.

Le corpus « France 3 » a servi de base d'apprentissage. Nous avons examiné le comportement de la distance Euclidienne entre les jingles de référence et le corpus « France 3 ».

Nous nous sommes aperçus que S pouvait être très grand sans pour autant que les résultats soient dégradés : nous avons pris $S = 60$ vecteurs. Ce délai important (environ une seconde pour un corpus échantillonné à 16 kHz) permet au système de fonctionner en temps réel.

La constante α est fixée à $log2$. Le seuil de rejet λ est corrélé au rapport entre la durée du jingle de référence N et le pas d'analyse S. Expérimentalement nous choisissons : $\lambda = 5 * N/S$.

4.3.3 Résultats

Nous avons testé **chaque** jingle de référence du catalogue (contenant 32 jingles) sur **tous** les corpora (cf. tableau 4.2).

TAB. 4.2 – *Détection manuelle et automatique des jingles de référence sur chacun des corpus de la base de données.*

Corpus	Détection auto	Détection manuelle
France 3	*4*	*4*
M6	16	16
Canal+	6	6
France Info	11	12
RFI	60	60
Publicités	33	34
Total	**130**	**132**

Les tests sur le corpus « France 3 » sont en italique car ce corpus nous a servi d'apprentissage de nos paramètres ; la reconnaissance parfaite prouve la pertinence du choix des paramètres.

Sur les 132 jingles que nous devions localiser et identifier, nous en avons détecté 130, soit 98,5 % de taux de reconnaissance. Les deux seuls jingles omis (un jingle « France Info » et un jingle publicitaire) sont complètement recouverts de parole (le présentateur parle durant le jingle !) et leur pic est dans ce cas beaucoup trop large (cf. figure 4.7).

La détection est excellente car nous n'avons aucune fausse alarme (insertion) et seulement deux omissions alors que d'autres jingles n'appartenant pas au catalogue de sons clés sont présents dans la base de données.

Bien que la base de données soit très variée, notamment par la différence des enregistrements entre les programmes de télévision et ceux de radio et la nature des émissions, le système

possède un comportement très satisfaisant. En outre, ces expériences prouvent la robustesse de notre système face aux choix des paramètres (S,α,λ). L'apprentissage fut court et aisé.

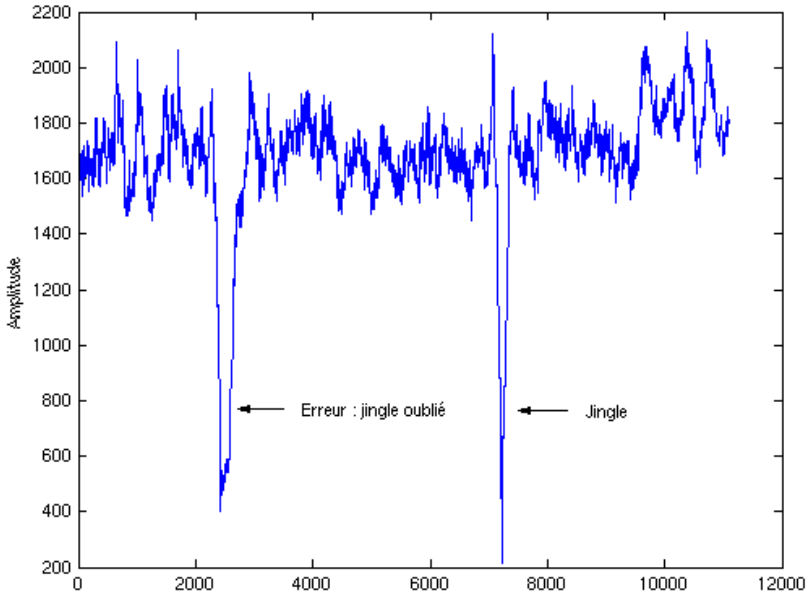

FIG. 4.7 – Exemple d'erreur (omission) du « jingle France Info » sur un extrait du corpus France Info (3 minutes).

Durant la phase d'évaluation, nous avons étudié la précision de la détection. La localisation des jingles est très bonne : la différence entre les localisations manuelle et automatique est très faible, inférieure à 500 ms quel que soit le jingle ; elle correspond au pas S utilisé.

Dans une optique d'indexation et de structuration, où généralement les décisions sont prises pour chaque seconde d'analyse, cette précision est largement suffisante. Le tableau 4.3 nous donne un exemple de la précision de notre système pour les quatre premiers jingles du corpus « France Info ».

TAB. 4.3 – *Précision de la localisation des jingles pour le corpus France 3.*

Corpus France Info	Détection manuelle	Détection automatique
1er jingle	62,4 s	62,69 s
2ème jingle	204,14 s	203,73 s
3ème jingle	408,04 s	408,34 s
4ème jingle	510,31 s	510,58 s

Nous pouvons aussi noter que notre système fonctionne en temps réel. En effet, pour traiter un fichier sonore d'une heure avec notre catalogue de 32 jingles de référence, moins d'une heure est nécessaire en utilisant un processeur AMD cadencé à 1,4 GHz.

La figure 4.8 nous montre un exemple de classification parole/musique sur la première ligne (a) obtenue avec notre système développé au chapitre 3, ainsi que la détection d'un jingle sur la deuxième ligne (b). Ce jingle est détecté aux environs de la quinzième seconde du corpus RFI commun aux expériences de nos deux systèmes.

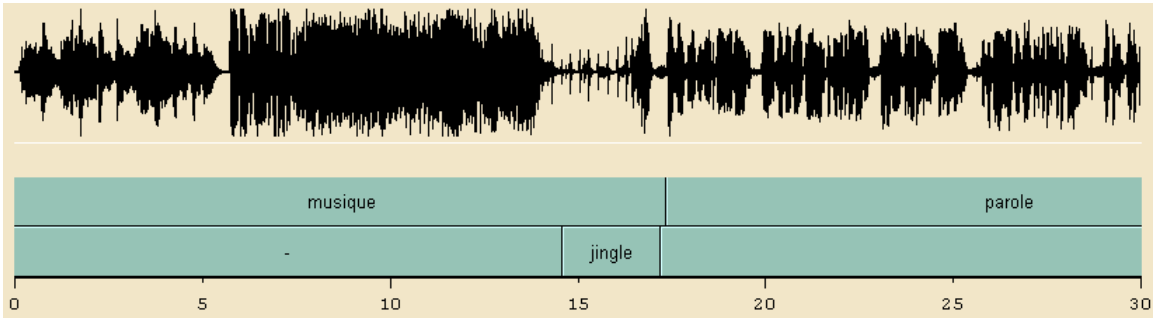

FIG. 4.8 – Exemple de partitionnement bas niveau : (a) classification parole/musique, (b) détection de jingle.

La majorité des jingles présents dans notre base de données est étiquetée en musique par notre système.

4.4 Conclusion

Dans le but de structurer les documents sonores, nous avons présenté un système de détection et d'identification de jingles fondé sur un calcul de distance euclidienne dans le domaine spectral.

Cette méthode est extrêmement simple néanmoins les résultats sont excellents. Sur 570 minutes (environ 10 heures) de flux sonore, nous n'observons aucune fausse alarme et seulement deux omissions dans des conditions extrêmes : la parole est superposée au jingle pendant l'**intégralité** de celui-ci. La localisation est très satisfaisante : nous pouvons déterminer le début d'un jingle avec une marge inférieure à la demi seconde, ce qui est largement suffisant pour une tâche d'indexation.

Notre système peut être considéré comme efficace par sa simplicité (seulement fondé sur une analyse spectrale), sa rapidité (temps réel), sa robustesse (indépendant du corpus) et la qualité de ses résultats. Il peut être utilisé pour une description des documents sonores de plus haut niveau, de manière à aider à la structuration ou au classement des émissions, comme nous le verrons au chapitre 7.

Chapitre 5

Les applaudissements, les rires et le locuteur cible

Sommaire

5.1 Introduction

5.1.1 Problématique

Les contenus sonores les plus variables d'une émission télévisée sont issus de sons provenant du milieu ambiant. Sons d'animaux, rires, voitures, avions, cloches, cris, explosions sont autant d'exemples qui peuvent aider à la détermination du contenu sémantique du document. Or, ils ont été jusqu'ici fort peu étudiés, car ils sont en général difficiles à appréhender.

Afin d'apporter une première réponse à cette attente, nous nous sommes intéressés à deux types de sons : les rires et les applaudissements. Le choix de ces sons clés est motivé par le fait qu'ils sont très présents dans des émissions de télévision dites de « plateau » (divertissement, jeu...), et que leur détection révèle la présence d'un événement caractéristique au sein de l'émission.

Ces émissions se caractérisent également par une alternance régulière du présentateur et des invités ou des reportages. La localisation des interventions du présentateur est une indication précieuse pour la structuration. Lors de sa prise de parole, le sujet ou l'unité traitée de manière adjacente est spécifiée ; la transcription de son intervention sera elle aussi intéressante (cf. chapitre 6).

Lors de ce chapitre, nous allons présenter un système de détection de rires et d'applaudissements et un système de détection d'un locuteur connu. Ces trois systèmes sont traités au cours de ce même chapitre car ils mettent en œuvre les mêmes outils, à savoir des mélanges de lois gaussiennes et une analyse spectrale (ou cepstrale) du continuum sonore.

Les rires et les applaudissements sont des sons inhabituellement étudiés, c'est pourquoi nous commençons par une brève description acoustique. Le système de détection est ensuite présenté à travers les trois modules classiques de reconnaissance des formes : le pré-traitement, la reconnaissance et l'apprentissage. Des expériences sur un corpus télévisuel de type divertissement, « Le Grand Échiquier », évaluent les trois systèmes considérés.

5.1.2 Les applaudissements et les rires

Comme pour la musique instrumentale dite « traditionnelle » (cf. section 1.1.2) et contrairement à la parole (cf. section 1.1.1), le signal correspondant aux **applaudissements** est d'un point de vue statistique stable (cf. figure 5.1). Les applaudissements sont des signaux d'un contenu spectral et d'une durée assez uniformes.

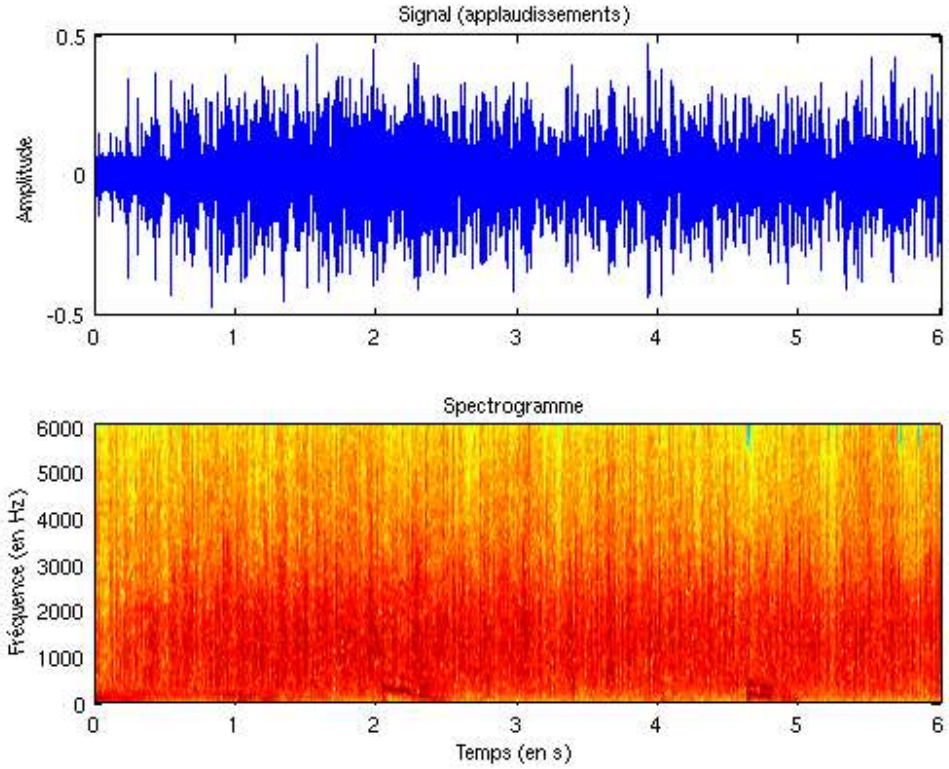

FIG. 5.1 – Signal et spectrogramme d'une séquence d'applaudissements durant six secondes d'une émission télévisuelle.

Par contre, il est difficile visuellement, que ce soit sur le signal ou le spectrogramme (cf. figure 5.2), de reconnaître des **rires** ; le signal est fortement bruité et non stationnaire : la modélisation de ce son clé apparaît d'ores et déjà difficile. Les rires présentent une grande variabilité naturelle car les personnes rient de plusieurs manières différentes alors que la manière d'applaudir semble universelle.

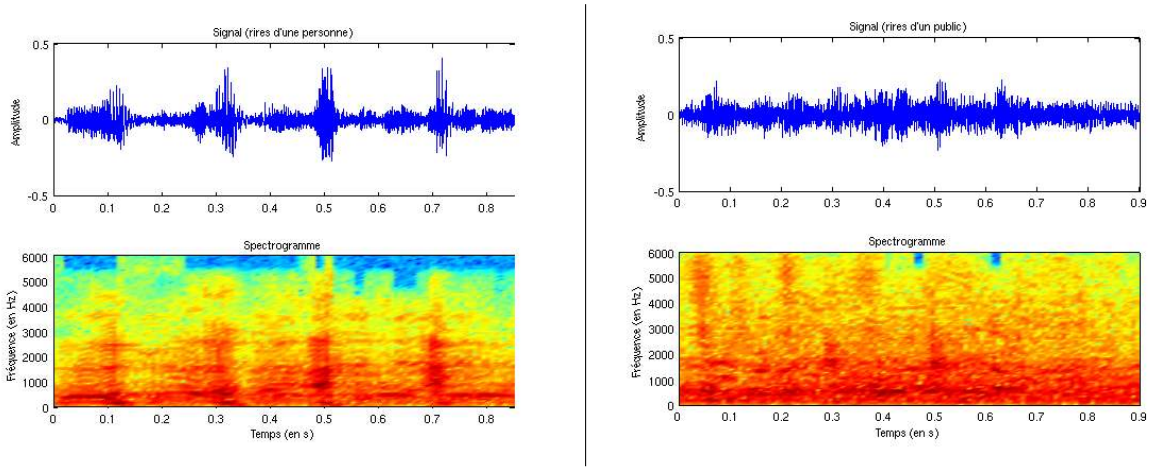

FIG. 5.2 – Variation entre les spectrogrammes correspondant à des rires d'une personne et ceux d'un public. Les extraits durent environ une seconde.

5.1.3 Le locuteur cible

Lorsque l'on parle de reconnaissance de locuteur, il s'agit de prendre une décision en utilisant les caractéristiques du signal de parole, afin de trouver l'identité du locuteur qui prononce une phrase donnée. Plusieurs approches sont possibles :

- l'**identification du locuteur** revient à retrouver un locuteur parmi N,
- la **vérification du locuteur** authentifie un locuteur,
- la **caractérisation du locuteur** cherche l'appartenance d'un locuteur à une catégorie.

Notre objectif est de localiser les interventions d'un présentateur ; nous sommes en présence d'un problème semblable à celui de la vérification du locuteur, si ce n'est que nous devons localiser le locuteur en question dans un flux de parole qui n'est pas uniquement composé de tours de parole d'individus isolés.

Nous aurions pu aborder le problème en le décomposant en deux étapes :

- segmentation du flux de parole en locuteur,
- recherche du locuteur dans les segments.

Pour éviter les problèmes difficiles rencontrés en segmentation de locuteur, nous avons préféré utiliser une stratégie de type détection classe/non-classe comparable à celle utilisée pour la détection de rires et d'applaudissements.

Ajoutons que le modèle « non-classe » correspond alors à une modélisation de la parole, tout environnement et individu confondus. Il est différent du modèle du « monde » employé en vérification du locuteur [Rey00] : celui-ci s'apprend à partir d'énoncés isolés d'individus.

5.2 Le système de base

Les systèmes de détection mis en place [Pin04b] s'inspirent très largement du système de base PMB (cf. chapitre 2). Chacun des systèmes consiste à identifier sur chaque trame de signal, la présence ou l'absence du phénomène considéré (applaudissements, rires ou présentateur) en question. Il se ramène donc à un problème de classification en classe et non-classe :

- applaudissements/non-applaudissements,
- rires/non-rires,
- locuteur/non-locuteur.

Comme dans le chapitre 2, nous utilisons un système composé de deux modules principaux : le pré-traitement et la décision (ou reconnaissance) (cf. figure 5.3). Les classes et les non-classes sont modélisées à l'aide de MMG et la décision se fait par maximum de vraisemblance.

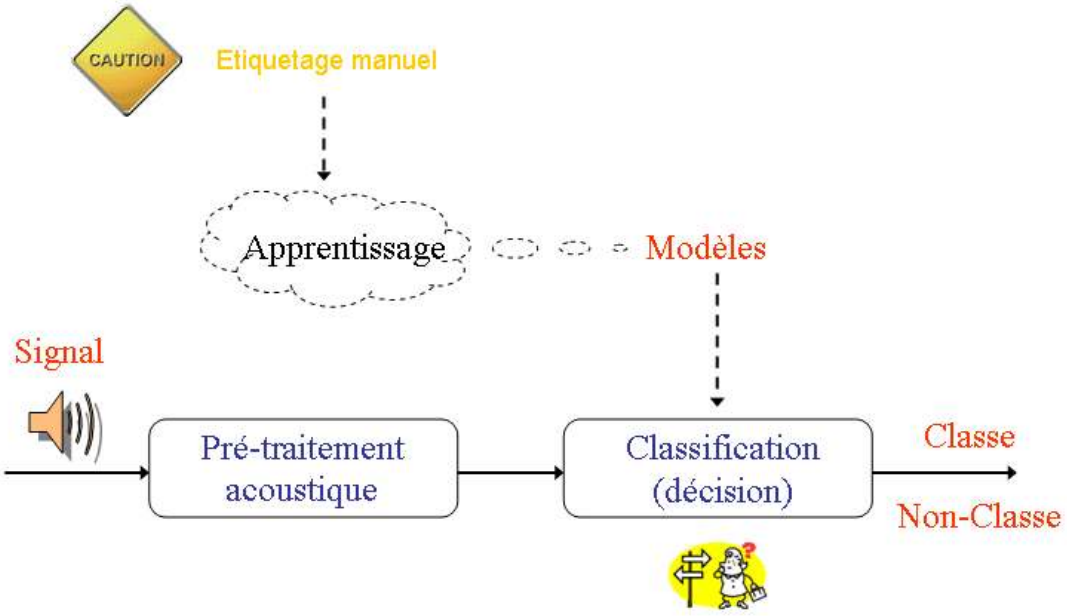

FIG. 5.3 – Schéma général du système de détection de base.

Un apprentissage est nécessaire afin de créer les modèles applaudissements et non-applaudissements, rires et non-rires, présentateur (locuteur) et « monde » (non-locuteur). Là encore, une phase d'étiquetage manuel est indispensable à l'apprentissage de nos modèles.

5.2.1 Pré-traitement

Les analyses cepstrale et spectrale que nous avons présentées précédemment ont été étudiées (cf. section 2.2.1) pour les systèmes « rires » et « applaudissements » :

- **analyse cepstrale** : 18 paramètres sont extraits (l'énergie, 8 coefficients cepstraux et les dérivées correspondantes),
- **analyse spectrale** : 29 paramètres sont extraits (l'énergie, 28 coefficients spectraux).

Afin de trouver la paramétrisation la plus appropriée à la création des modèles rires, non-rires, applaudissements et non-applaudissements et la taille de la trame la plus adaptée, le signal est soumis à divers pré-traitements acoustiques :

- la longueur des trames d'analyse varie de 128 à 1024 échantillons,
- le nombre de coefficients spectraux varie également et ses dérivées sont ajoutées,
- le recouvrement utilisé est de la moitié de la trame.

Pour caractériser le signal dans la recherche du locuteur, nous avons repris l'analyse cepstrale sur 10 ms, analyse couramment rencontrée dans la littérature en vérification du locuteur.

5.2.2 Apprentissage et reconnaissance

La phase de reconnaissance est similaire à celle présentée dans la section 2.2.2 :

- la classification se fait suivant la règle du maximum de vraisemblance entre les modèles classe et non-classe,
- une fonction de lissage permet de ne garder que les segments significatifs, après regroupement des trames correspondant à la même décision.

Pour les sons clés, applaudissements et rires, la phase d'apprentissage est identique à celle du système de base PMB (cf. section 2.3). Le nombre de lois gaussiennes dans le mélange varie de 32 à 256 gaussiennes pour des matrices de covariance diagonales et de 1 à 4 gaussiennes pour des matrices de covariance pleines.

Pour le locuteur, la modélisation est la même, fondée sur des MMG avec un apprentissage en deux temps, initialisation et optimisation des modèles. Cependant, les deux modèles ne sont

pas appris en parallèle, l'obtention du modèle « locuteur » est différente : le modèle locuteur (classe) est une réestimation du modèle « monde » par le critère MAP (cf. section 2.3.4) sur les moyennes. L'algorithme EM permet d'effectuer cette réestimation (cf. figure 5.4).

Le modèle « monde » est appris sur toute les zones de parole, y compris celles où se trouve le locuteur cible : le monde n'est plus un modèle d'anti-locuteur ; cela n'induit pas une difficulté supplémentaire comme le montrent les performances.

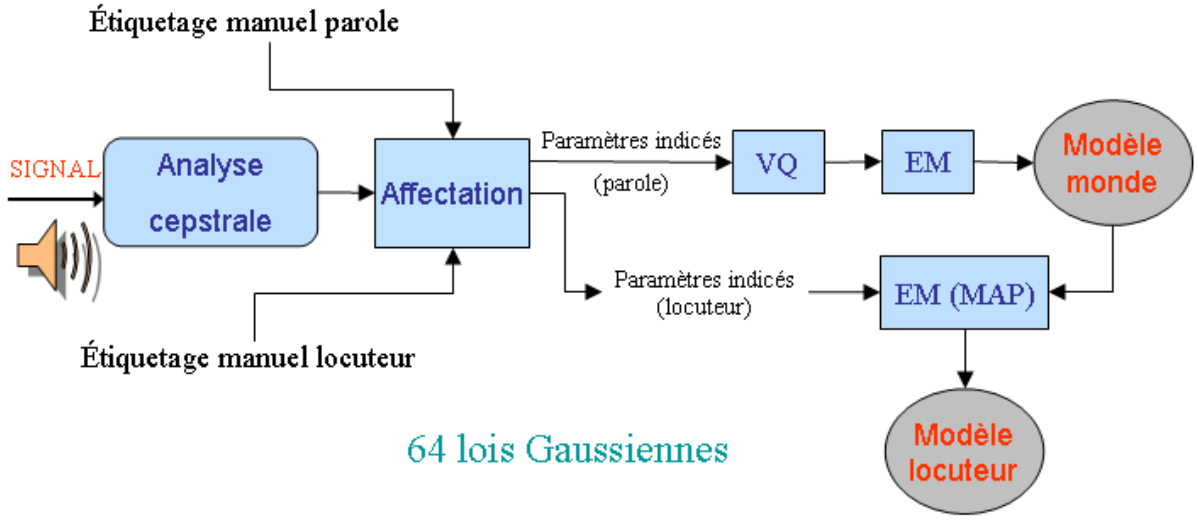

FIG. 5.4 – Procédure d'apprentissage des modèles de mélanges de lois gaussiennes correspondant au locuteur et au monde.

5.3 Expériences et résultats

5.3.1 Corpus

Le corpus est composé de plusieurs documents audiovisuels qui correspondent à cinquante quatre émissions « Le Grand Échiquier », soit plus de 160 heures. Ces émissions, de type plateau, ont la particularité de contenir un grand nombre de nos sons clés (applaudissements et rires), et elles sont présentées par un même intervenant : Jacques Chancel.

Le contenu de ce corpus est assez divers : de la musique (classique, jazz, variété française...), des interviews et des sketches. Chaque émission a une durée d'environ 190 minutes. La première que nous appelons « GE1 » nous sert d'apprentissage et la seconde « GE2 » nous sert de test.

Nous utilisons donc au total seulement 6 heures de notre corpus. L'utilisation d'autres émissions est possible, des tests ont d'ailleurs été effectués sur deux autres émissions. Malheureusement, la tâche d'annotation manuelle nécessaire pour évaluer les résultats « a priori » est très pénible, et nous avons fait cet étiquetage pour deux émissions seulement.

Ce corpus a fait l'objet d'un accord de confidentialité entre l'**INA** et l'**IRIT** dans le cadre du projet **RIAM FERIA**.

Le canal audio est séparé du canal vidéo (MPEG 1, layer II) et mis au format suivant :
- type : wave ou raw (brut, sans les entêtes),
- quantification : 16 bits,
- fréquence d'échantillonnage : 16 kHz,
- canal monophonique.

Les séquences d'applaudissements sont en général claires et bien définies avec une durée comprise entre 5 et 8 secondes. Par contre, les rires sont plus variables : ils peuvent durer de 0,5 à 4 secondes. Les signaux de rires les plus réguliers sont les rires du public, mais ce type de signal est très proche du bruit.

La localisation manuelle du présentateur ne pose pas de problèmes et celle du « monde » correspond à l'étiquetage en « parole ».

5.3.2 Les applaudissements et les rires

5.3.2.1 Critère d'évaluation

Au niveau de l'apprentissage, nous avons effectué des choix quant à la création des modèles que nous allons expliciter. Tout d'abord et de manière classique (cf. section 2.4.2), le nombre d'itérations des algorithmes VQ et EM est fixé à 10 et le seuil d'arrêt à 10^{-5} pour la variation relative de la vraisemblance.

Le critère d'évaluation des résultats se base sur le rapport (exprimé en pourcentage) entre le temps correctement segmenté et sa durée totale. L'outil d'évaluation du NIST est utilisé ici avec un paramètre de délai (cf. section 2.4.3.1).

L'évaluation par cet outil n'est significative que si les classes sont homogènes. Or les zones de rires et d'applaudissements ont de très faibles durées comparées aux zones « non-classes » correspondantes.

L'étiquetage manuel sur l'émission de test de 3 heures environ (11396 secondes) donne :
- durée des applaudissements : 906 secondes,
- durée des non-applaudissements : 10490 secondes,
- durée des rires : 652 secondes,
- durée des non-rires : 10744 secondes.

L'ordre de grandeur est de plus de 1 pour 10 pour nos sons clés. De ce fait, il faut prendre des précautions lors de l'analyse des résultats car un système, par exemple, qui ne détecte aucun rire, nous donne tout de même 94,28 % (11396-652/11396) de taux de classification correcte. Ce taux correspond à ce que nous appelons la « chance ».

Pour les applaudissements, cette chance se situe à 92,05 %. Un système qui donne un taux inférieur à la chance est donc un système qui ne fonctionne pas !

Bien que ce critère d'évaluation soit suffisant pour juger si un système est meilleur qu'un autre, il faudra trouver un moyen de savoir si le système est performant. Pour cela, nous allons regarder quelle proportion de segments nous avons retrouvée par rapport à celle que nous cherchions et la cohérence de ceux-ci.

5.3.2.2 Détection des applaudissements

Différents choix sont possibles afin d'effectuer la paramétrisation et la modélisation :
- le type de paramétrisation : coefficients cepstraux (MFCC), coefficients spectraux (SPL) avec/sans dérivées,
- la taille de la fenêtre : de 128 à 1024 points,
- le nombre de lois gaussiennes : de 32 à 256,
- le lissage : 500 à 2000 ms.

L'ensemble des combinaisons possibles est testé. Dans les tableaux 5.1 et 5.2, quelques extraits de ces nombreux tests sont présentés.

Nous sommes arrivés aux conclusions suivantes :

- l'analyse spectrale (notée « SPL ») est plus performante que l'analyse cepstrale (notée « MFCC ») pour ce son clé.
- l'ajout des dérivées aux coefficients spectraux (notation « SPLΔ ») n'améliore pas les résultats et les pénalise même légèrement,
- pour la taille de la fenêtre d'analyse (la trame), les meilleurs résultats sont obtenus pour des valeurs de 1024 échantillons (ou points), soit 64 millisecondes pour un signal échantillonné à 16 kHz,
- un lissage à 1 s donne de meilleurs résultats.

TAB. 5.1 – *Résultats de tests sur la détection des applaudissements. Les matrices de covariance de lois gaussiennes sont diagonales d'où la notation « D ».*

Paramétrisation	MFCC	MFCC	SPL	SPL	SPL	SPL	**SPL**
Taille fenêtre	256	256	128	256	256	512	**1024**
Nb lois gaussiennes	32D	64D	64D	32D	64D	64D	**64D**
Score (en %)	91,7	85,3	93,7	93,9	94,2	97	**98,58**

TAB. 5.2 – *Résultats de tests sur la détection des applaudissements avec une taille de fenêtre de 1024 points. Les notations « D » et « P » au niveau du nombre de lois gaussiennes correspondent à l'emploi de matrices de covariance respectivement diagonales et pleines. L'ajout des dérivées premières des coefficients spectraux est repéré par la notation « SPLΔ ».*

Paramétrisation	**SPL**	SPLΔ	SPL	SPL
Taille fenêtre	**1024**	1024	1024	1024
Nbe lois gaussiennes	**64D**	64D	1P	4P
Score (en %)	**98,58**	98,49	98,47	98,56

Par rapport à notre volume d'apprentissage (émission GE1 d'environ 3 heures), le nombre optimum de lois pour les MMG est de 64 pour des matrices de covariances diagonales (notées D). Un résultat comparable est obtenu avec des MMG de 4 lois gaussiennes avec matrices de covariance pleines.

Finalement, la configuration la plus efficace est de choisir des trames de **1024 échantillons** avec une **analyse spectrale** (sans dérivées) en utilisant **64 lois gaussiennes**, des matrices de covariances diagonales et un lissage d'**une seconde**. Nous obtenons 98,58 % pour la détection des applaudissements, ce qui semble excellent.

Mais, comme nous l'avons précisé précédemment, ce taux est à relativiser : il nous faut observer les applaudissements retrouvés. Sur les 906 secondes d'applaudissements que nous

avons repérées manuellement, nous avons relevé 144 segments. Seulement 72 de ces segments sont **significatifs** et ceux-ci sont tous bien détectés par notre système. Lorsque nous employons le terme « significatif », il désigne des *segments assez longs*, de durée supérieure à 1 seconde, et *pur* : ce ne sont pas des segments de faibles amplitudes ou superposés à de la parole.

Dans notre tâche visant à structurer les documents audiovisuels, seuls les événements caractéristiques (segments significatifs) sont importants à détecter car les autres ne nous permettent pas d'interprétation ultérieure.

Les insertions sont très faibles (seulement 2 segments d'environ 2 secondes chacun). Ces insertions peuvent être supprimées assez facilement en faisant un lissage plus important (par exemple un lissage sur 3 secondes), les segments significatifs font entre 5 et 8 secondes.

En conclusion, nous avons :

- pour l'étiquetage manuel : 906 secondes (144 segments dont 72 significatifs),
- pour l'étiquetage automatique : 771 secondes d'étiquetage correct (97 segments dont 72 significatifs).

Le taux de reconnaissance des applaudissements, au sens classique, est de 85 % (771/906).

Notre système de détection des applaudissements est très performant. La figure 5.5 présente les résultats d'une classification applaudissements/non-applaudissements sur la ligne 1. La ligne 3 correspond à l'étiquetage manuel correspondant.

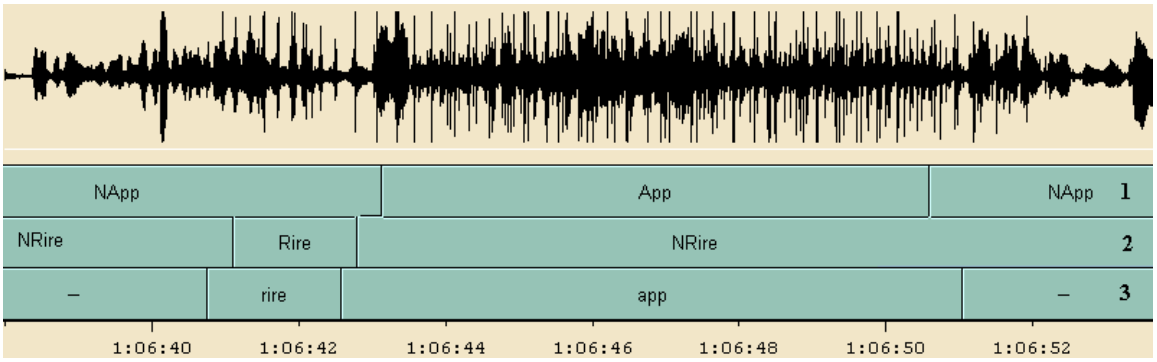

FIG. 5.5 – Exemple de résultats de classification applaudissements/non-applaudissements (ligne 1) et rires/non-rires (ligne 2) par rapport à un étiquetage manuel (ligne 3).

5.3.2.3 Détection des rires

Comme pour la détection des applaudissements, de nombreux choix sont à faire. Dans les tableaux 5.3 et 5.4, quelques tests parmi les nombreux effectués sont présentés.

TAB. 5.3 – *Résultats de tests sur la détection des rires en utilisant des matrices de covariance de lois gaussiennes diagonales.*

Paramétrisation	MFCC	MFCC	SPL	SPL	SPL	SPL	**SPL**
Taille fenêtre	256	256	128	256	256	512	**1024**
Nbe lois gaussiennes	32D	64D	64D	32D	64D	64D	**128D**
Score (en %)	74,3	74,5	82,7	74,4	89	91	**97,26**

Nous pouvons remarquer que la configuration optimale pour la détection des rires est assez proche de celle des applaudissements. Il s'agit également d'une **analyse spectrale** sur des trames de **1024 échantillons** avec un lissage de **500 millisecondes**. Par contre, le nombre de **lois gaussiennes** est de **128** avec des matrices de covariance diagonales. Nous obtenons un taux de classification correcte de 97,26 %, au sens de l'outil d'évaluation du NIST, pour la détection des rires qui semble excellent.

TAB. 5.4 – *Résultats de tests sur la détection des rires avec une taille de fenêtre constante (1024 points).*

Paramétrisation	**SPL**	SPL	SPLΔ	SPL	SPL
Taille fenêtre	**1024**	1024	1024	1024	1024
Nbe lois gaussiennes	**128D**	64D	64D	1P	4P
Score (en %)	**97,26**	95,98	95,75	56,47	14,4

Observons plus précisément les rires retrouvés. Nous avons :

- pour l'étiquetage manuel : 652 secondes (359 segments dont 135 significatifs, c'est-à-dire supérieurs à 500 millisecondes),
- pour l'étiquetage automatique : 212 secondes d'étiquetage correct (221 segments dont 102 significatifs).

Le taux de reconnaissance des rires est de 32,52 % (212/652). Ce score est très faible : ceci signifie que sur trois segments de rires recherchés, nous n'en trouvons qu'un ! Là encore il faut relativiser ce score car le nombre d'événements importants (segments significatifs) retrouvés est assez élevé (102 sur 135).

La figure 5.5 présente les résultats d'une classification rires/non-rires sur la ligne 2. La ligne 3 correspond à l'étiquetage manuel. Notre système de détection des rires est correct sans être excellent. Quelles en sont les raisons ?

Le signal correspondant aux rires est difficile à modéliser à cause de sa forte variabilité. Lors de l'indexation automatique, les signaux de faible amplitude ne sont pas détectés, des séquences de musique sont considérées comme des rires et des décalages de frontières sont présents. Dans une tâche de structuration, savoir qu'il y a des rires sur de la musique ne nous est d'aucune aide, comme nous le verrons dans le chapitre 7, nous avons donc décidé d'ignorer ces rires.

La méthode conservée est la suivante : lorsque une étiquette « rires » est insérée dans le même intervalle de temps qu'une étiquette « musique », elle est supprimée. La figure 5.6 présente un exemple de fusion (ligne 3) entre une classification musique/non-musique (ligne 1) et rires/non-rires (ligne 2). Cette méthode permet d'augmenter légèrement les résultats mais n'est pas suffisante.

La fusion des résultats entre le système de classification musique/non-musique (cf. chapitre 3) et le système de détection de rires donne un taux de classification correcte de 97,45 % au sens de l'outil d'évaluation NIST.

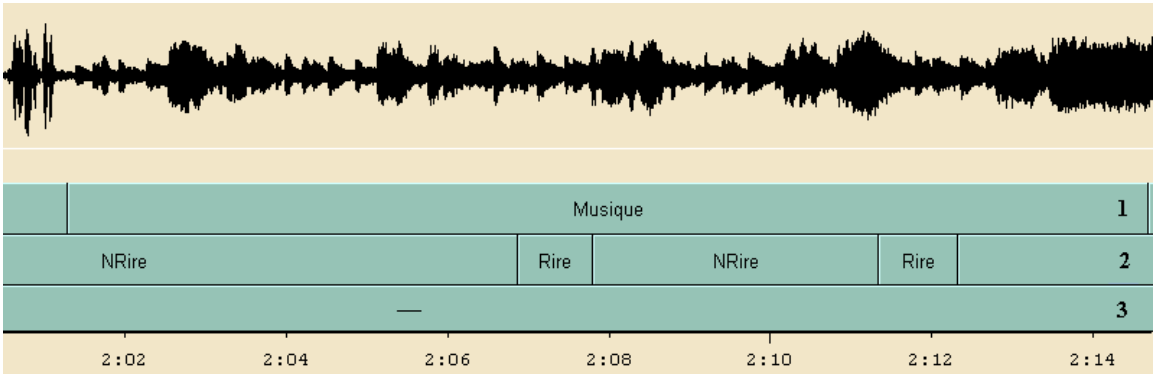

FIG. 5.6 – Suppression des étiquettes « rires » (ligne 3) grâce à l'apport d'information du système de classification musique/non-musique (ligne 1) sur le système de classification rires/non-rires (ligne 2). La durée des deux segments « Rire » est d'environ 1 s.

5.3.3 Le locuteur cible

Sur les 3 heures d'émission de l'apprentissage, il n'y a que 45 minutes de parole environ dont 25 minutes correspondent au présentateur Jacques Chancel. Nous utilisons les 45 minutes de parole afin d'apprendre le modèle du « monde ».

Le modèle du locuteur est obtenu en adaptant le modèle du « monde », avec les 3 premières minutes d'étiquetage qui correspondent au présentateur.

À la suite d'une série d'expérimentation, nous avons pris :
- **64 lois** pour les deux MMG,
- un **fenêtrage de 256 points** avec recouvrement sur la moitié,
- **18 MFCC** : 8 coefficients cepstraux, l'énergie et les dérivées premières,
- un **lissage d'une seconde**.

Le taux de classification correcte obtenu, en utilisant l'outil NIST, est de 92,89 % en utilisant une classification parole/non-parole manuelle au préalable, c'est-à-dire parfaite : la détection du locuteur cible ne se fait que sur des zones de parole, et sur toutes les zones de parole !

Dans le cas « réel », où nous utilisons la classification parole/non-parole automatique, issue du système présenté dans le chapitre 3, le taux de classification correcte baisse légèrement à 89,74 %. Il convient de noter que sur ce fichier de test « GE2 », la classification parole/non-parole obtient 94,99 % de taux de classification correcte et que les erreurs sont principalement des insertions de parole sur de la musique.

La figure 5.7 est un exemple de détection du présentateur à partir d'un étiquetage manuel parole/non-parole puis d'un étiquetage automatique.

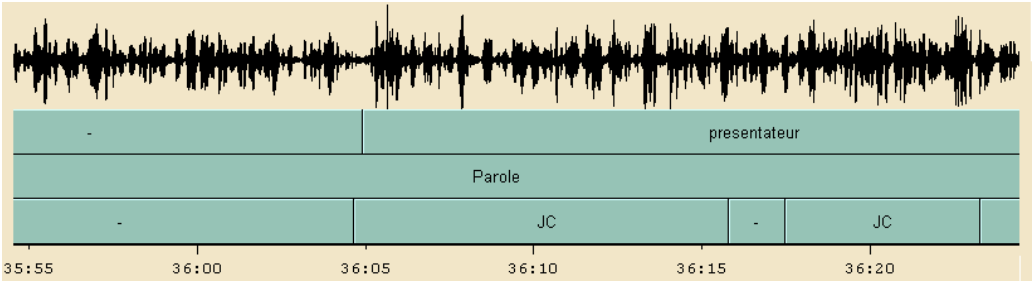

FIG. 5.7 – Détection manuelle (ligne 1) et automatique (ligne 3) du locuteur cible par rapport à une détection de parole automatique (ligne 2) sur un extrait du fichier « GE2 » du corpus « le Grand Échiquier ». « JC » et « présentateur » désignent Jacques Chancel.

5.4 Conclusion

Au cours de ce chapitre, plusieurs systèmes de détection ont été présentés, afin de détecter des rires, des applaudissements et un locuteur cible : le présentateur de l'émission. Ces systèmes, mis en œuvre à partir de mélanges de lois gaussiennes, sont fondés sur une analyse spectrale du continuum sonore pour les sons clés, applaudissements et rires, et sur une analyse cepstrale pour le locuteur.

Pour une première étude, les résultats sont intéressants. Le système de détection des applaudissements est excellent car il n'y a pratiquement pas d'insertions et tous les segments significatifs sont retrouvés.

La détection du présentateur Jacques Chancel, le locuteur cible choisi, est encourageante avec un taux de classification correcte supérieur à 90 %. Cette détection nous sera très utile pour la structuration (cf. chapitre 7).

Par contre, le système de détection des rires est moins bon car les insertions sont nombreuses et ceci malgré une détection musique/non-musique au préalable ; cette mise en œuvre permet d'augmenter légèrement les résultats mais n'est pas suffisante.

Paramétrisation (analyse spectrale) et modélisation par des MMG ne sont sans doute pas adaptés à ce type de signaux. Les données d'apprentissage sont sans doute également en quantité insuffisante.

José Arias pourra apporter une réponse complémentaire à ces questions car il effectue une thèse sur une modélisation fondée sur des SVM (Support Vector Machine) et travaille sur cette problématique [Ari04].

Chapitre 6

Les mots clés

Sommaire

6.1 Introduction

6.1.1 Problématique

La reconnaissance de mots clés est un cas particulier de celle des sons clés. Cette thématique doit prendre une part beaucoup plus importante parmi les méthodes d'indexation sonore car une transcription totale des zones de parole est une opération très coûteuse en temps de calcul, en adaptation et plus difficile.

Le rôle de la détection de mots clés est important et complémentaire aux outils issus de la vidéo. Lorsque les méthodes de détection de séquences ou de reconnaissance de visages se heurtent à une dégradation visuelle, il est alors intéressant d'utiliser les mots clés afin d'obtenir une information sur la scène. Cette information peut guider l'interprétation.

Afin d'arriver à une bonne efficacité, la reconnaissance de mots clés doit pouvoir s'adapter aux situations les plus diverses : robustesse face au bruit, indépendance vis-à-vis des locuteurs et du canal de transmission... De plus, elle doit offrir un temps de latence le plus faible possible afin de garder tout son intérêt face à des systèmes de transcription automatique complet.

Notre travail n'a pas pour objectif l'amélioration des techniques de détection de mots clés. Il s'agit ici d'une première étude, fondée sur des méthodes assez classiques et notre but est de montrer comment on peut améliorer la structuration des documents sonores et audiovisuels par ajout de nouvelles informations telles que les mots clés. Ces mots clés ne seront pas spécifiques à un corpus (ou un type de corpus) comme c'est généralement le cas ; ils devront rester généralistes afin de représenter par exemple le type d'une émission ou la transition d'une émission à une autre. Ainsi, nous faisons une détection de mots clés en vue d'identifier des thèmes.

6.1.2 Bref historique

La nécessité d'un outil permettant la détection de mots clés n'est pas apparue récemment. Déjà Wilpon [Wil90], par une approche Markovienne, a mis en évidence ce besoin à travers une étude sur une base de données téléphoniques.

Depuis, de nombreux systèmes de reconnaissance de mots clés ont été proposés. Ils sont fondés sur trois types de stratégies : la programmation dynamique (**DTW** : **D**ynamic **T**ime **W**arping) [Bez93], les réseaux de neurones [Mor91] et les MMC [Ros90]. Certains sont fondés sur la combinaison de ces architectures [Cer93]. Les MMC restent la modélisation la plus

employée compte tenu de leur performance (en apprentissage et en reconnaissance) et de leur efficacité (coût de calcul faible).

L'utilisation de MMC en détection de mots clés implique la notion de modèles poubelles, les modèles complémentaires des modèles de mots clés, ou la définition de mesures de confiance.

Les modèles poubelles

L'un des précurseurs fut Rose [Ros91] qui proposa d'utiliser les mots clés pour le tri de messages vocaux. Il fit une étude assez approfondie sur les **modèles « poubelles »** que Wilpon [Wil90] et Higgins [Hig85] avaient introduits. Rappelons qu'un modèle « poubelle », dans le cadre des MMC, est un modèle de tous les mots qui ne sont pas des mots clés : il représente non seulement les mots hors vocabulaire mais aussi les bruits, les faux départs, les hésitations.

Il y a diverses formes de modèles poubelles :

- les **modèles dynamiques** [Boi93] : le MMC est fait d'un seul état et la probabilité d'émission d'une trame est la moyenne des N meilleures vraisemblances de cette trame, calculées pour chacun des états des modèles mots clés atteints à cet instant. L'état dynamique modélise tout le vocabulaire mais avec une moins bonne qualité pour les mots clés. Ce procédé permet de réduire la charge de calcul inhérente aux modèles poubelles classiques.

- les **modèles lexicaux** au niveau des syllabes [EM98]. Deux types de modèles poubelles sont possibles : soit chaque modèle poubelle représente une syllabe, soit un modèle poubelle contient toutes les syllabes ayant la même fréquence d'apparition dans le corpus d'apprentissage. Pour construire le modèle poubelle, El Méliani regroupe les syllabes ayant des fréquences d'apparition semblables. Les avantages de ce procédé sont de permettre un traitement plus rapide, de ne pas nécessiter de modèle acoustique propre au modèle poubelle, et d'effectuer un apprentissage des modèles acoustiques indépendamment de la tâche à réaliser.

- les **modèles phonétiques** : Zhang [Zha01] propose cinq modèles poubelles. Le premier est appris sur les mots contenants un ou deux phonèmes, le deuxième sur les mots de trois phonèmes. Le dernier est appris sur les mots comportants plus de six phonèmes. Le choix de cinq modèles poubelles est motivé par le fait qu'un seul modèle poubelle ne peut pas couvrir correctement tous les mots hors vocabulaire.

En comparaison des systèmes de transcription automatique, la détection de mots clés permet de résoudre les problèmes d'hésitations, de faux départs, de phrases tronquées et de phrases grammaticalement incorrectes. Mais des problèmes de faux rejets et de fausses acceptations apparaissent. L'intérêt de l'utilisation des modèles poubelles est de diminuer le nombre de fausses acceptations.

Une alternative aux modèles poubelles sont les **anti-modèles** [Suk96] et [Mor00]. À chaque modèle M_Φ du phonème Φ correspond un anti-modèle $\overline{M_\Phi}$ qui modélise les erreurs de substitutions et de fausses alarmes. Cette représentation s'apparente à notre notion de classe/non-classe.

Les mesures de confiance

Les mesures de confiance ont été très étudiées [Chi92], [Cox96], [Cam97]. À chaque mot clé détecté correspond une mesure de confiance et lors de la phase de vérification, cette mesure permet l'acceptation ou le rejet des différentes détections des mots clés. Cette méthode oblige de trouver un seuil permettant de limiter les fausses acceptations.

L'étude des techniques existantes nous a conduit à proposer un système simple de détection de mots clés. Ce système de détection de mots clés, fondé sur des MMC de nature phonétique, est décrit au cours de ce chapitre à travers les phases de paramétrisation, d'apprentissage et de reconnaissance. Nous nous attardons sur la phase de paramétrisation dans la mesure où nous voulons sélectionner des représentations robustes du signal, pour être le plus indépendant possible des conditions d'enregistrement. Différents tests sont effectués afin de déterminer la paramétrisation adéquate d'une part et d'évaluer les performances de notre système dans une tâche de structuration d'autre part.

6.2 Le système de détection de mots clés

Nous avons choisi d'utiliser une approche phonétique centrée sur un ensemble de modèles de phonèmes : les modèles de mots clés sont décrits par la suite des modèles de phonèmes qui le composent et le modèle poubelle permet de représenter toutes les suites de phonèmes possibles.

Il s'en suit que la réalisation de notre système se compose de quatre modules (cf. figure 6.1) :

- le **pré-traitement acoustique** au cours duquel la paramétrisation du signal est effectuée. Deux paramétrisations sont utilisées : les **MFCC** et les **PLP** (**P**erceptual **L**inear **P**redictive).
- l'**actualisation du modèle global** de notre système (lexique, modèle de langage).
- l'**apprentissage** qui utilise l'algorithme de Baum-Welch afin d'estimer les paramètres des modèles de Markov cachés (nombre d'états, transitions...) correspondant aux unités de base choisies.
- la **reconnaissance** fondée sur l'algorithme de Viterbi afin de déterminer le chemin optimal. Lors de cette phase, un lexique et une grammaire sont nécessaires.

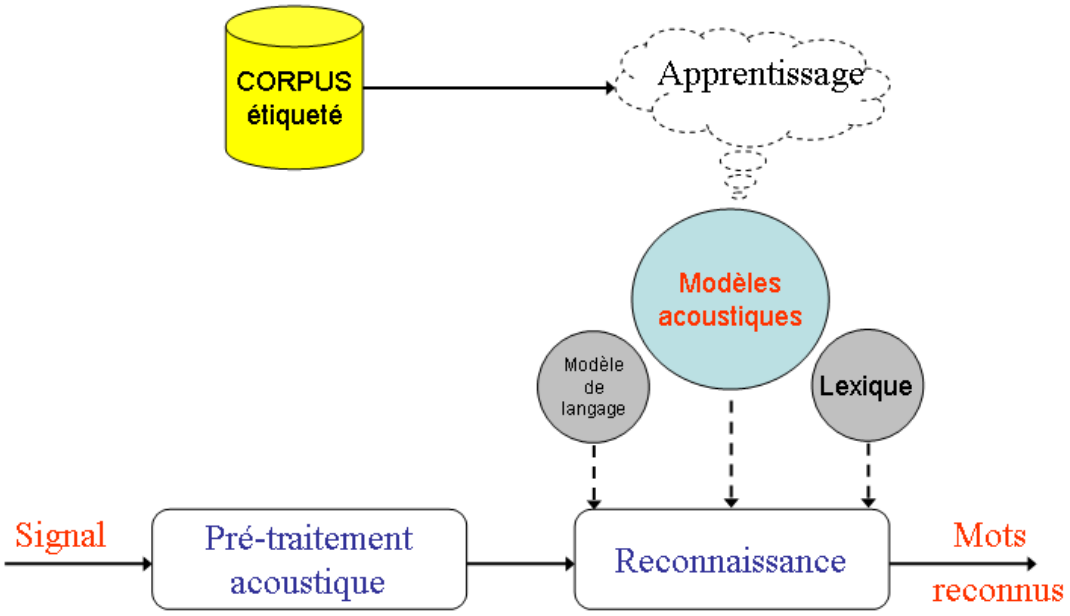

FIG. 6.1 – Le système de détection de mots clés.

6.2.1 Pré-traitement acoustique

Le rôle de l'analyse acoustique est de trouver un jeu de paramètres qui caractérisent au mieux, à chaque instant, le signal de parole afin d'être robuste au bruit, le plus indépendamment possible des variabilités du signal.

Notre système utilise l'analyse cepstrale (les MFCC décrits dans la section 1.2.4) et l'analyse perceptuelle (les PLP). Ces deux techniques, fondées sur la transformée de Fourier, utilisent des fenêtres de 16 ms avec un recouvrement sur la moitié (8 ms). L'ajout de dérivées premières et secondes permet d'extraire la dynamique acoustique et respecte la plasticité du signal de parole (rallonger la durée des voyelles n'altère pas l'intelligibilité).

Nous avons étudié l'analyse perceptuelle puisque dans nombre d'études, cette paramétrisation semble plus robuste et que nos conditions environnementales sont difficiles. L'analyse perceptuelle est fondée sur la modélisation autorégressive du signal. Nous en rappelons brièvement les principes ci-après.

6.2.1.1 Analyse par codage prédictif linéaire (LPC)

Appelée **LPC** (**L**inear **P**redictive **C**oding), cette méthode se fonde sur les connaissances de production de la parole et fait l'hypothèse d'un modèle linéaire [Fan60]. Le conduit vocal est le plus souvent modélisé par un filtre autorégressif excité soit par un bruit blanc (pour les frications), soit par un peigne de Dirac (pour les sons voisés). L'analyse par prédiction linéaire suppose que les échantillons x_n du signal de parole sont corrélés, et que le signal peut être prédit par :

$$\widehat{x_n} = -\sum_{i=1}^{p} a_i.x_{n-i} \tag{6.1}$$

avec a_i les coefficients de prédiction et p l'ordre de prédiction.

L'erreur de prédiction est alors :

$$e_n = x_n - \widehat{x_n} \tag{6.2}$$

Les coefficients a_i peuvent être déterminés par minimisation de l'erreur quadratique en faisant l'hypothèse de pseudo-stationnarité du signal sur la fenêtre analysée :

$$E = E(e_m^2) = \sigma^2 \tag{6.3}$$

L'annulation des dérivées partielles de E par rapport aux coefficients a_i ($a_0 = 1$) et l'utilisation de la matrice d'autocorrélation R conduisent à la résolution d'un système d'équations dits de Yule-Walker réduites de $p+1$ équations et $p+1$ inconnues :

$$\begin{bmatrix} R_0 & R_1 & \cdots & R_p \\ R_1 & R_0 & \ddots & R_{p-1} \\ \vdots & \ddots & \ddots & \vdots \\ R_p & \cdots & R_1 & R_0 \end{bmatrix} . \begin{bmatrix} 1 \\ a_1 \\ \vdots \\ a_p \end{bmatrix} = \begin{bmatrix} \sigma^2 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \tag{6.4}$$

Pour résoudre ce système, la méthode d'autocorrélation préserve le caractère Toëplitz de la matrice R, en prenant un estimateur ergodique.

$$r_{|i-j|} = \sum_{m=-\infty}^{+\infty} x_{m-i}.x_{m-j} \tag{6.5}$$

L'algorithme de Levinson [Lev47] est une version rapide de la résolution des équations 6.4 en effectuant une récurrence sur l'ordre du modèle.

6.2.1.2 Analyse par prédiction linéaire perceptuelle (PLP)

La prédiction linéaire perceptuelle est une exploitation des connaissances du système auditif humain pour paramétrer la parole : il s'agit de rendre compte des mécanismes psychoacoustiques de l'oreille humaine [Her90].

L'analyse par PLP repose sur l'analyse par prédiction linéaire où la matrice d'autocorrélation est calculée par transformée de Fourier inverse du module au carré de la transformée de Fourier du signal. On introduit au niveau du spectre de puissance des bandes critiques. Cette intégration se fait avec un banc de 17 filtres dont les fréquences centrales sont espacées linéairement selon l'échelle Bark (cf. figures 6.2 et 6.3).

Cette intégration se justifie par le fait que le système auditif se comporte comme un banc de filtres dont les bandes, appelées « bandes critiques », se chevauchent et dont les fréquences centrales s'échelonnent continûment. Chaque bande critique correspond à l'écartement en fréquence nécessaire pour que deux harmoniques soient discriminées dans un son complexe périodique. L'échelle Bark, comme l'échelle Mel, reproduit approximativement la sensibilité de l'oreille. Par rapport au signal de parole, cette échelle est beaucoup plus performante que l'échelle Hertz.

Toujours dans le domaine spectral, on effectue ensuite une pré-accentuation à l'aide d'un filtre du premier ordre pour rendre compte des courbes d'isosonie. La sonie ou intensité perçue est liée physiquement à la pression acoustique : c'est la contrepartie perceptive de l'amplitude. On applique une phase de compression d'intensité en sonie, en approximant l'échelle perceptuelle par la fonction racine cubique.

À la suite de ces opérations dans le domaine spectral, la transformée de Fourier inverse fournit une estimation de la matrice d'autocorrélation R. Les PLP sont les coefficients cepstraux déduits de l'analyse LPC obtenue à partir de cette estimation de R.

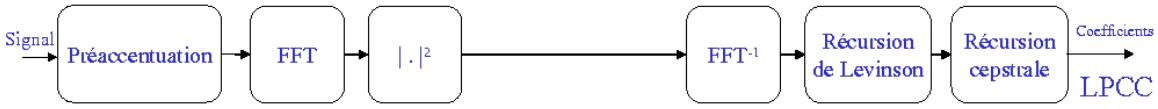

FIG. 6.2 – Processus de création des coefficients LPC.

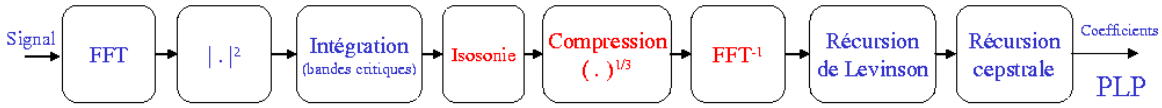

FIG. 6.3 – Processus de création des coefficients PLP.

Par rapport à une représentation classique (spectrogramme, prédiction linéaire), l'analyse PLP effectue une normalisation, les expériences montrant que les différences entre locuteurs sont presque annulées. Les PLP représentent les formes grossières du spectrogramme plutôt que les détails ; ils doivent ainsi être plus robustes au bruit [Her85].

6.2.2 Les Modèles de Markov Cachés (MMC)

Depuis leur introduction en traitement de la parole ([Bak75] et [Jel76]), les MMC ont pris une place importante, au point que la quasi-totalité des systèmes de reconnaissance de la parole les utilise.

Nous présentons sommairement ces MMC puisqu'ils sont maintenant maîtrisés par l'ensemble de la communauté « parole » et que les références sont multiples [Ros90], [Mat01], [Raz04]... Nous décrivons la plate-forme utilisée et l'utilisation des MMC dans notre système.

6.2.2.1 Présentation des MMC

Les MMC sont utilisés par les systèmes de reconnaissance de la parole pour faciliter l'identification des mots représentés par le signal acoustique. Un MMC décrit la réalisation d'une concaténation de variables élémentaires qui représentent la séquence de paramètres acoustiques extraits d'un énoncé humain. Un MMC est semblable à une machine à états finis dans laquelle les transitions et les émissions sont stochastiques. Chaque état représente un son et émet de manière probabiliste un vecteur d'observations. Les transitions représentent les différentes possibilités d'enchaîner les sons.

La classification consiste à définir pour chaque classe un Modèle de Markov Caché. La fonction de décision est fondée sur le critère du maximum de vraisemblance.

Le MMC est moins performant, trame à trame, qu'un réseau de neurones mais son gros avantage est l'intégration de la dimension temporelle. Cette notion explique pourquoi il est souvent employé en reconnaissance de parole.

Un modèle est constitué d'un nombre fini d'états prédéterminés, et est défini traditionnellement comme suit :

- un nombre N d'états numérotés de 1 à N,
- la loi de probabilité initiale Π,

$$\Pi = (\pi_1, \cdots, \pi_N) \tag{6.6}$$

avec $\pi_i \geq 0$ et $\sum_i \pi_i = 1$,

– la matrice de transition A, dont les éléments a_{ij} représentent la probabilité de passer de l'état i, à un instant quelconque, à l'état j à l'instant suivant :

$$\sum_{j=1}^{N} a_{ij} = 1 \tag{6.7}$$

avec $1 \leq i, j \leq N$.

– la matrice d'observation notée $B = [b_i]$ caractérise les lois d'observations sur chaque état. Les lois sont en général des mélanges de lois gaussiennes. Chaque (b_i) est alors paramétrée par $(\lambda_l, \mu_l, \Sigma_l,\ l = 1, \cdots, L)$.

L'apprentissage des paramètres d'un tel modèle est réalisé à l'aide de l'algorithme de Baum-Welch (cf. annexe G) et la reconnaissance, fondée sur la recherche du meilleur chemin, est effectuée grâce à l'algorithme de Viterbi (cf. annexe F).

6.2.2.2 La plate-forme HTK

La plate-forme **HTK** (Hidden Markov Model Toolkit), développée par l'Université de Cambridge [You94] fournit tous les outils logiciels nécessaires à la réalisation de systèmes fondés sur des MMC (cf. annexe E).

Les modèles peuvent aussi bien représenter les mots ou tout type d'unité sub-lexicale (phonème, triphone), et leur topologie est librement configurable. Les densités de probabilité d'émission, qui sont associées aux états, sont décrites par des lois multigaussiennes. Les modèles sont initialisés avec l'algorithme de Viterbi, puis réestimés par l'algorithme de Baum-Welch. Le décodage est réalisé par l'algorithme de Viterbi, sous la contrainte d'un réseau syntaxique défini par l'utilisateur.

Notre choix s'est porté vers cette plate-forme pour plusieurs raisons.

– Contrairement à Julius[10] et Sirocco[11] qui ne sont que des moteurs de reconnaissance de la parole, HTK permet de construire la chaîne complète : apprentissage et reconnaissance.
– L'ensemble des outils, écrit en langage C, possède une documentation très détaillée sur leur utilisation et les principes de leur implémentation : ceci permet d'intégrer de manière efficace les modifications souhaitées.

[10] http ://julius.sourceforge.jp/en/
[11] http ://www.irisa.fr/sirocco/

– HTK est un système largement répandu dans le monde de la recherche : ceci peut servir à évaluer (ou comparer) de manière plus précise les résultats.
– Par l'intermédiaire de listes de diffusions, une assistance aux problèmes techniques nous est offerte.

6.2.2.3 Modélisation phonétique

Depuis leurs premières utilisations, les chaînes de Markov cachées ont montré qu'elles étaient appropriées à modéliser de nombreux types d'unités de parole ([Bak75], [Bak76]). Idéalement, il faudrait pouvoir associer à chaque phrase possible un MMC. En pratique, ceci est impossible car le nombre de modèles et le volume d'apprentissage par voie de conséquence seraient trop importants. La situation s'aggrave dans le cas de grands vocabulaires, et associer à chaque mot son modèle distinct pose là encore des problèmes d'apprentissage. C'est pourquoi la reconnaissance de grands vocabulaires est toujours effectuée à partir des MMC d'unités comme les phonèmes, les diphones ou les triphones. L'approche fréquemment utilisée consiste à prendre un modèle à 3 états par phonème, en faisant l'hypothèse que l'état du milieu modélise la partie stationnaire du phonème et les états extérieurs modélisent la coarticulation avec les phonèmes voisins (cf. figure 6.4). Le nombre de phonèmes choisi est généralement de l'ordre de 35 pour le français.

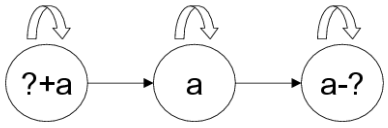

FIG. 6.4 – Exemple d'un modèle de phonème à trois états : le phonème est « a ».

Un autre modèle souvent utilisé est le triphone ou allophone ou phonème en contexte ; il tient compte de deux phonèmes, le précédent et le suivant pour fixer le contexte (cf. figure 6.5). On en dénombre environ 7500 et leur utilisation nécessite un entraînement sur des bases de données conséquentes.

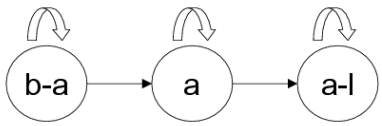

FIG. 6.5 – Exemple d'un modèle de triphone : le triphone est « b-a-l ».

Les modèles des mots sont alors obtenus par la concaténation de plusieurs MMC. Ces MMC correspondent à la suite des unités caractérisant les mots en question. La collecte d'énormes corpus d'apprentissage est évitée et il est possible d'ajouter de nouveaux mots dans le lexique à tout moment (cf. figure 6.6). Aucun étiquetage supplémentaire n'est a priori nécessaire ; il suffit juste de déterminer la ou les séquences d'unités correspondant aux nouveaux mots et de construire le MMC les modélisant.

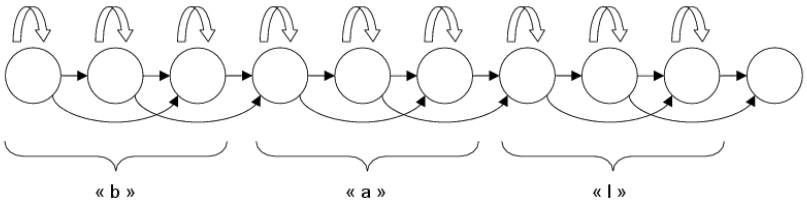

FIG. 6.6 – Exemple de MMC d'un mot obtenu par concaténation de phonèmes : le mot est « bal ».

Nous nous inspirons de cette approche pour définir le modèle de notre détecteur de mots clés.

6.2.3 Le modèle de mots clés

Nous choisissons comme **unité de base le phonème** et nous optons pour une approche de type **modèle « poubelle »** afin de représenter toutes les phrases possibles. Ce modèle correspond à tous les enchaînements de phonèmes, y compris ceux représentant les mots clés.

Le modèle de chaque **mot clé** correspond à la **concaténation des modèles de phonèmes** le composant, ces modèles étant ceux du modèle poubelle. Le choix de l'approche par concaténation de phonèmes, pour la constitution de nos modèles de mots clés permet une forte flexibilité. En effet, l'ajout d'un nouveau mot clé ne nécessite aucun apprentissage supplémentaire : il suffit uniquement de décomposer les différentes prononciations du nouveau mot en suites de phonèmes.

Afin de favoriser le passage vers les modèles de mots clés et pénaliser le modèle poubelle, des poids w_k, fonction du nombre d'états du modèle k, sont utilisés. Les différentes valeurs de favorisation w_k seront choisies à la suite d'une phase d'expérimentation.

Les mots clés sont réunis en **thèmes**. Chacun des thèmes est représenté par un ensemble de mots clés (cf. figure 6.7). Le nombre de mots clés par thème peut être différent et un mot clé peut appartenir à plusieurs thèmes. Chaque passage par un thème est quantifié et la décision finale sur le choix du thème d'un reportage se fait par un vote majoritaire.

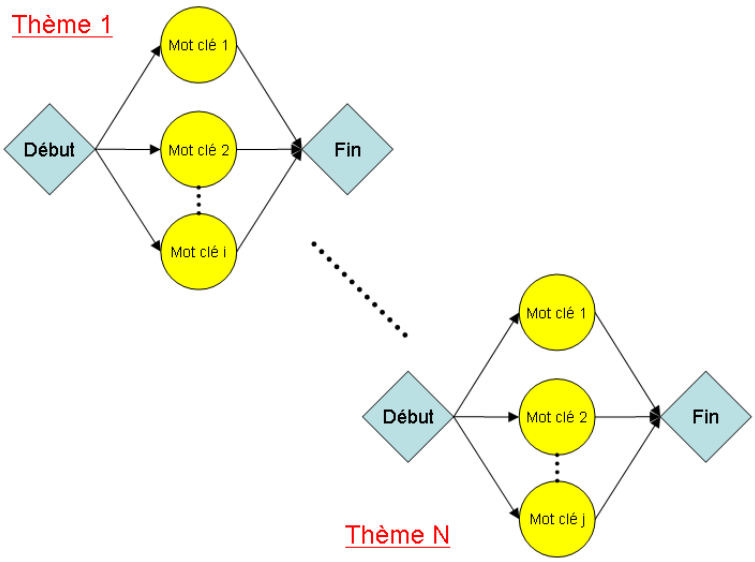

FIG. 6.7 – Réseaux correspondant à la décomposition des thèmes en mots clés.

Nous disposons également d'un **modèle de silence**. Tous ces modèles (mots clés, poubelle et silence) s'enchaînent selon la grammaire présentée sur la figure 6.8.

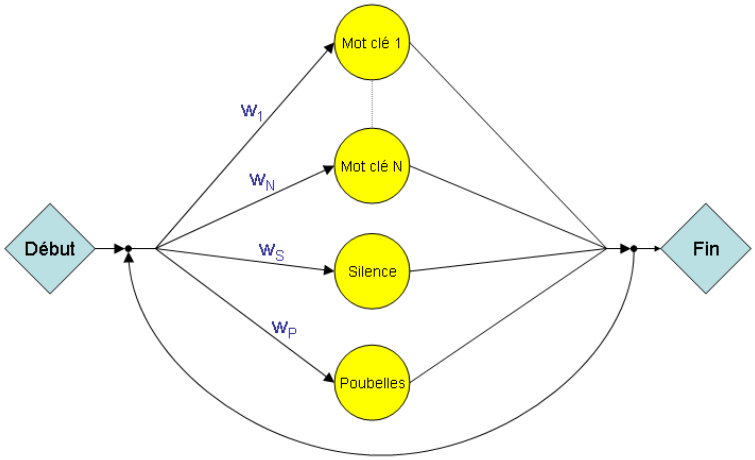

FIG. 6.8 – Grammaire de notre système de détection de N thèmes.

6.3 Expériences et résultats

6.3.1 Corpus

Notre base de données expérimentale est divisée en deux parties :

- le **corpus d'apprentissage** est composé d'environ 30 heures : 20 heures de France Inter et 10 heures de RFI. Ce corpus a été transcrit manuellement dans le cadre de la campagne d'évaluation ESTER, l'unité étant le phonème. Il nous sert à apprendre les modèles des phonèmes de notre système.

- le **corpus de test** est le même que celui utilisé lors de la section 3.4.1 : il s'agit d'enregistrements de RFI d'une durée de 6 heures. Sur ces 6 heures, 4 heures et 46 minutes correspondent à de la parole dont 3 heures et 12 minutes de français. Il est à noter que les enregistrements de RFI ont été réalisés sur des périodes totalement différentes.

6.3.2 Mise en œuvre

Afin de valider notre approche, 20 mots clés ont été choisis et classés en 5 thèmes :
- *politique* : politique, président, ministre, Europe et gouvernement ;
- *économie* : technologie, industrie, travail et entreprise ;
- *catastrophe* : génocide, attentat, victime, sécurité et militaire ;
- *sports* : championnat, victoire et football ;
- *météo* : dépression, précipitations, température.

En se référant à la section 6.2.3, le mot *victoire* est représenté par quatre formes prenant en compte les différentes prononciations et les silences possibles en fin de mots :
- v i k t w a R
- v i k t w a R sil
- v i k t w a R @
- v i k t w a R @ sil

Les prononciations de ces mots clés sont issues des ressources lexicales BDLEX développé à l'IRIT [DC98].

Dans notre corpus de test, nous avons dégagé manuellement 81 sujets traitant de nos cinq thèmes (cf. tableau 6.1). Le terme « sujet » désigne les différents reportages rencontrés. La longueur des sujets peut varier de 20 secondes à 5 minutes. À l'intérieur des sujets, au moins un mot clé est présent. Ces sujets sont délimités dans le corpus de test de manière manuelle. Dans le chapitre 7, cette délimitation sera automatisée grâce à une détection de jingles.

TAB. 6.1 – *Nombre de sujets par thèmes dans le corpus RFI.*

Thèmes	Nombre de sujets
politique	34
économie	14
catastrophe	9
sports	18
météo	6

Nous possédons 37 modèles de phonèmes. Chaque modèle a 3 états et le nombre de mélanges de lois gaussiennes par état varie de 10 à 32.

Dans le corpus de test sont présents 283 occurrences des 20 mots clés. Les autres mots (plus de 180000) sont des mots hors-vocabulaires. Cette disproportion rend difficile la détection. Cette détection est d'autant plus délicate que le corpus d'apprentissage correspond essentiellement à des émissions d'informations alors que le corpus de test est beaucoup plus varié : reportages, interviews, publicités...

Nous utilisons la classification parole/non-parole (cf. chapitre 3) comme pré-traitement. Seules les zones de parole détectées sont présentées au détecteur de mots clés.

6.3.3 Évaluation

Une partie de l'évaluation a eu pour objectif de fixer les différents choix posés au niveau de l'apprentissage et du type de paramétrisation. Nous avons testé 5 configurations différentes :

- Système 1 : 12 MFCC + Énergie + dérivées premières et MMC avec un mélange de 10 lois gaussiennes par état,
- Système 2 : 12 MFCC + Énergie + dérivées premières + dérivées secondes et MMC avec un mélange de 10 lois gaussiennes par état,
- Système 3 : 12 MFCC + Énergie + dérivées premières + dérivées secondes et MMC avec un mélange de 32 lois gaussiennes par état,

– Système 4 : 12 PLP + Énergie + dérivées premières + dérivées secondes et MMC avec un mélange de 32 lois gaussiennes par état,
– Système 5 : 12 PLP + Énergie + dérivées premières + dérivées secondes et dérivées troisièmes + MMC avec un mélange de 32 lois gaussiennes par état,
– Système 6 : 12 MFCC + dérivées premières + dérivées secondes + dérivée Énergie et MMC avec un mélange de 32 lois gaussiennes par état.

Les résultats, en terme d'accuracy au sens d'HTK (cf. section 2.4.3.1), pour chacun des systèmes sont présentés dans le tableau 6.2. Le meilleur taux est de 56 % pour un système fondé sur la paramétrisation MFCC. Contrairement à notre attente, les PLP ne se sont pas montrés plus performants que les MFCC.

TAB. 6.2 – *Comparaison des systèmes de détection de mots clés.*

Configuration	Accuracy
Système 1	42,57 %
Système 2	54,55 %
Système 3	56,04 %
Système 4	40,20 %
Système 5	46,70 %
Système 6	**56,62 %**

Ces taux ne sont pas excellents pour une tâche de détection de mots clés. Mais il convient de préciser plusieurs choses :

– les taux peuvent être améliorés en utilisant des phonèmes en contexte tels que les triphones,
– ce système est utilisé sur un corpus de test (magazines, interviews, informations) assez différent de celui d'apprentissage (uniquement des informations),
– le choix de la modélisation permet de choisir n'importe quel mot clé lors de la phase de reconnaissance (le mot clé étant décrit comme une suite de phonèmes),
– le système est cinq fois plus rapide que le temps réel : moins de douze minutes sont nécessaires à l'analyse d'un fichier d'une heure (processeur Pentium 4 2,6 GHz, mémoire vive 2 Go).

Caractérisation des émissions

Notre but n'est pas de rechercher des mots clés afin de concurrencer la transcription automatique totale mais d'arriver à extraire le thème de chacun de nos sujets.

Les résultats de la détection des thèmes sont très bons : nous avons retrouvé 74 des 81 thèmes soit 91,4 % de taux de reconnaissance (cf. tableau 6.3).

TAB. 6.3 – *Résultats de la détection des sujets dans le corpus RFI.*

Thèmes	Nombre de sujets (manuel)	Nombre de sujets correctement reconnus (automatique)
politique	34	**33**
économie	14	**10**
catastrophe	9	**8**
sports	18	**17**
météo	6	**6**

Quasiment toutes les erreurs sont dues à l'insertion de mots clés correspondant au thème « politique » (cf. tableau 6.4).

TAB. 6.4 – *Matrice de confusion des thèmes sur le corpus RFI.*

Thèmes	politique	économie	catastrophe	sports	météo
politique	**33**	0	1	0	0
économie	4	**10**	0	0	0
catastrophe	1	0	**8**	0	0
sports	1	0	0	**17**	0
météo	0	0	0	0	**6**

6.4 Conclusion

Dans ce chapitre, nous avons présenté un système de détection de mots clés permettant de classer des reportages selon des thèmes. Ce système est fondé sur des modèles de Markov cachés de nature phonétique. Les modèles de mots clés correspondent à la concaténation des modèles de phonèmes qui les composent et les modèles poubelles représentent tous les enchaînements de phonèmes possibles.

Cette approche est dynamique et flexible car le choix des mots clés n'entre pas en compte lors de l'apprentissage des modèles de phonèmes. Un ajout de mot clé ne nécessite pas un apprentissage supplémentaire, il suffit de le décomposer en suites de phonèmes.

Bien que les résultats de la détection de mots clés ne soient pas excellents, seulement 56,62 % d'accuracy au sens d'HTK, la détection de thèmes qui en découle est performante : 91,4 % de taux de reconnaissance. Notons que le corpus d'apprentissage n'est composé que de journaux télévisés alors que le corpus de test est plus diversifié : reportages, interviews...

Il ne s'agit ici que d'une première étude sur la détection de mots clés visant à montrer, si besoin en était, son intérêt en vue d'une structuration de documents audiovisuels et sa faisabilité. Notre système peut être amélioré de nombreuses façons, nous pourrons par exemple utiliser une modélisation par phonèmes en contexte qui donne de meilleurs résultats dans la littérature. Nous proposons lors du chapitre 7 une structuration faisant intervenir ce type d'approche.

Troisième partie

Vers une structuration audiovisuelle

Chapitre 7

Réflexions sur une structuration audiovisuelle

Sommaire

7.1 Introduction

7.1.1 Structuration et indexation automatique

Zhang [Zha97] présente les principes que doivent respecter les systèmes de recherche d'information multimédia en général et donc les systèmes d'indexation de documents audiovisuels. Pour que les descripteurs, issus de l'indexation automatique, soient utilisables par un système de requêtes, une analyse structurelle doit être ajoutée à tout système d'information. Les interfaces de navigation répondant aux systèmes de requêtes sont alors optimisées.

Pour Zhang [Zha97], l'analyse structurelle d'un document englobe les résultats issus de deux traitements :

- une segmentation temporelle du document à indexer, qui entraîne la notion d'unité de base,
- une extraction de contenu de ces unités issues.

Jusqu'à maintenant, la structuration temporelle d'émissions audiovisuelles est essentiellement étudiée au travers de l'analyse du flux visuel. On distingue classiquement deux niveaux de structure dans le document que sont le plan et la séquence.

Définitions

- Une séquence est un ensemble de plans constituant une unité narrative définie selon l'unité de lieu et d'action.
- Un plan correspond à ce qui est tourné en une seule prise de caméra.

Faire un découpage en séquences est une macrosegmentation du document.

Ce type d'analyse a donné lieu à de nombreuses études dont celle d'Aigrain qui, exploitant les règles de production, identifie les changements de séquences ([Aig97], [Rui02]).

Notre objectif est de montrer, au travers de quelques exemples, qu'une structuration temporelle de documents audiovisuels peut être atteinte à partir d'annotations automatiques du flux audio.

Nous nous plaçons dans l'optique de recherche :

- d'**événements caractéristiques** au niveau sonore, révélateurs d'un schéma de production.
  Exemples : Un son clé détermine-t-il une séquence ?
  Un jingle joue-t-il le rôle de logo visuel en caractérisant une publicité, voire une marque ?

- d'**enchaînements temporels** possibles entre les différents événements, révélateurs aussi de la structure.
  Exemple : Quelle est la signification des applaudissements ?

L'étude conjointe des flux audio et vidéo doit conduire à la définition d'une macrosegmentation d'un document audiovisuel beaucoup plus riche et intéressante pour l'indexation et la recherche d'informations. Ceci relève de nos perspectives à court et moyen terme.

7.1.2 Analyse audio

Lorsque l'on parle d'apport du flux sonore à une structuration audiovisuelle, on pense évidemment à la transcription automatique de la parole. Mais cette technique pose d'énormes problèmes encore à l'heure actuelle :

- le temps de calcul reste important : 10 à 20 fois le temps réel (le temps réel est possible mais dégrade fortement les résultats),
- le traitement est spécifique au corpus utilisé, à une tâche : des modèles appris sur un corpus de journaux télévisés conduisent à des performances intéressantes sur ce type de documents mais les résultats sont moins bons (voire très mauvais) sur des documentaires ou des émissions moins contraintes acoustiquement,
- les taux d'erreurs sont encore élevés et une étape de correction n'est pas envisageable car ce temps est en général supérieur au temps de transcription manuelle total.

Heureusement, cette approche n'est pas la seule : une recherche de composantes primaires telles la parole, la musique (partie I) ou les sons clés (partie II) peut fournir des informations très utiles, comme le montrent d'ores et déjà certaines études [Kij03].

Les analyses audio peuvent bien entendu servir de pré-traitement, de post-traitement ou d'aide aux analyses d'image classiquement utilisées en indexation (segmentation en plans, détection de visages...). Chang [Cha96] utilise les mots clés des commentaires audio de matchs de football américain pour repérer les moments importants de la retransmission.

Réciproquement, [Li96] et [Lie99] extraient des caractéristiques sonores de plans déjà extraits d'un document.

7.1.3 Organisation

Nous nous plaçons ici dans le cadre d'une première analyse de scène par les composantes sonores primaires : la recherche de parole, de musique, de jingles, d'applaudissements et de rires. Ces informations de bas niveau ne sont pas directement exploitables pour la structuration de documents audiovisuels. Pour accéder à une information de plus haut niveau, il faut d'une part les regrouper, et d'autre part voir leurs impacts sur les autres informations sonores ou sur les informations issues de l'image ou du texte.

Nous avons examiné ces deux stratégies.

- En premier lieu, nous présentons les fusions possibles de nos systèmes en vue d'une structuration et notamment l'enchaînement temporel de nos différentes composantes. Il s'agit ici de tester sommairement l'impact de nos outils.
- Ensuite, en se plaçant dans un cadre beaucoup plus prospectif, nous réfléchissons à l'apport de nos outils aux techniques existantes que ce soit en audio ou en vidéo. Notamment, nous évaluons les manques de nos travaux plutôt monomédia (fondés sur le son) qui nous amènent à étudier, développer des outils du média vidéo et envisager leur fusion ou leur corrélation.

7.2 Structuration : nos apports

Nous proposons deux exemples de fusion possible de nos outils de segmentation sonore en vue d'une structuration de plus haut niveau :

- une recherche de motif est effectuée sur une collection d'émissions,
- une étude de structuration du journal télévisé est décrite. Ce domaine fait l'objet de nombreux travaux.

7.2.1 Détection de motif dans une collection d'émissions

Lorsqu'une seule émission est analysée, un traitement très spécifique peut être effectué. Suivant la durée de l'émission, un traitement manuel peut même être réalisé et s'avérer plus rapide.

Par contre, quand nous sommes en présence de plusieurs émissions, comme la collection du « Grand Échiquier » présentée dans la section 5.3.1 dont nous possédons 54 émissions, il est nécessaire de changer de stratégie. Les traitements ne peuvent être qu'automatiques vue la durée du corpus (plus de 160 heures). Le niveau structurel doit rester assez grossier afin de correspondre à toute la collection d'émissions considérée.

L'émission « le Grand Échiquier » est une émission de variétés où le présentateur Jacques Chancel interroge plusieurs invités sur leur vie artistique. Comme beaucoup d'émissions de divertissement, cette émission possède une alternance d'interviews et de spectacles. Les interviews sont en général entre le présentateur et le ou les invités qui participent au spectacle. Le spectacle est en général une œuvre musicale mais il peut s'agir aussi de sketchs humoristiques, d'extraits d'émissions précédentes ou d'extraits de films.

L'étude de cette collection a permis de définir un motif, c'est-à-dire un enchaînement récurrent de caractéristiques communes à chacune des émissions. Ce motif structure les émissions en parties homogènes ; en général, il s'agit du passage d'un invité à un autre. Le motif est le suivant :
présentateur / *[applaudissements]* / **spectacle** / *[applaudissements / spectacle]* / **applaudissements / présentateur**.

Ceci signifie que dans cette collection, un spectacle (chanson, morceau de musique, sketch, extrait de film...) est introduit par le présentateur et est suivi par des applaudissements. À la fin de ceux-ci, le présentateur reprend la parole. Des applaudissements peuvent éventuellement précéder la composition artistique ou la « découper ».

Nous avons choisi de rechercher cette structuration sur le même fichier de test, émission « GE2 » que, lors du chapitre 5 pour des raisons évidentes : nous possédons déjà la « vérité terrain » correspondant aux détections d'applaudissements et du présentateur pour cette émission. Lors de cette émission, nous avons répertorié une succession de dix de ces motifs.

Afin de retrouver le motif en question, de manière automatique, nous allons appliquer une stratégie dite « aveugle ». Trois classifications sont effectuées indépendamment les unes des autres (cf. figure 7.1) :

– une détection de musique permettant de repérer les chansons,
– une détection de parole, pré-traitement pour la recherche du présentateur,
– une détection des applaudissements.

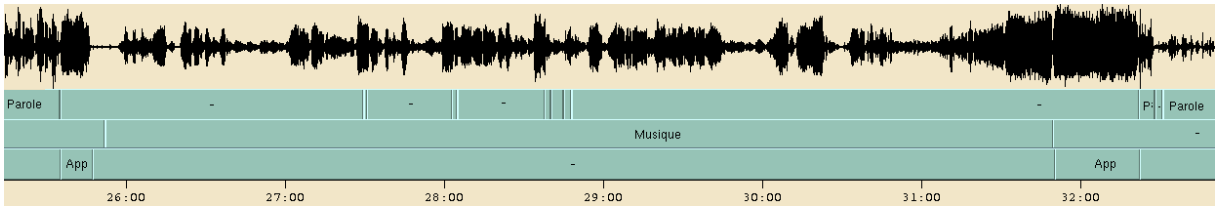

FIG. 7.1 – Exemple de recherche de motif sur 7 minutes de l'émission « GE2 » de la collection du « Grand Échiquier » à travers les détections automatiques de parole (ligne 1), de musique (ligne 2) et d'applaudissements (ligne 3).

Une détection du présentateur Jacques Chancel est effectuée sur les zones de parole détectée : il s'agit d'une classification présentateur/non-présentateur.

Notons, que les modèles du présentateur, du non-présentateur (le « monde »), des applaudissements et des non-applaudissements sont les mêmes que ceux développés lors du chapitre 5. Ils ont été appris sur une autre émission : « GE1 ». Rappelons aussi, que pour les détections de parole et de musique, notre système PMB de fusion ne nécessite pas d'apprentissage (cf. chapitre 3). Il n'y a donc eu aucun apprentissage sur GE2.

La figure 7.2 est un exemple de résultat obtenu par une mise en commun de tous les résultats de détection.

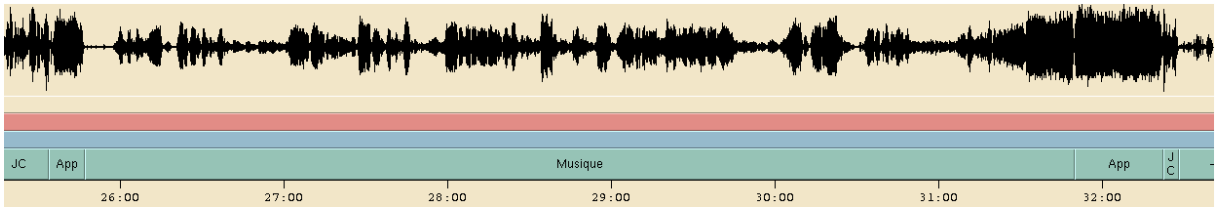

FIG. 7.2 – Exemple de résultat obtenu par fusion des différentes détections sur un extrait de 7 minutes de l'émission « GE2 » du « Grand Échiquier ». « JC » représente Jacques Chancel.

Les résultats sont excellents sur l'émission « GE2 ». Les 10 motifs cherchés sont retrouvés et ceci bien que les détections de parole et de musique soient imparfaites : quelques légers décalages et des insertions de parole sur la musique sont observés.

L'ensemble de cette structuration peut être faite sur l'ensemble des 54 émissions de cette collection sans aucun autre apprentissage ou intervention manuelle.

Comme nous l'avons déjà signalé, le motif recherché n'est pas spécifique au « Grand Échiquier » mais commun à nombre d'émissions de plateaux. La recherche d'un tel motif sur une autre émission ne nécessite alors qu'un étiquetage de 2 ou 3 minutes du nouveau présentateur de l'émission afin de créer son modèle par adaptation du modèle du « monde » (cf. chapitre 5). Les autres détections ne nécessitent pas d'apprentissage.

7.2.2 Structuration d'un journal télévisé

Le journal télévisé est un excellent exemple de besoin en structuration afin d'avoir accès à un contenu précis. L'utilisateur (téléspectateur) pourrait ainsi naviguer uniquement sur les informations qui lui semblent utiles ou intéressantes ! Le « 6 minutes » de M6 est un exemple très caractéristique de ce type d'émission. Ce journal télévisé est assez bref et chaque sujet d'information quotidienne, ou chaque reportage, est séparé des autres par l'intermédiaire d'un jingle. Deux types de jingles sont utilisés (cf. figure 7.3). Le premier type de jingle, « Jingle générique », correspond aux génériques de début et de fin d'émission, et le second, « Jingle M6 », délimite les différentes actualités. Une des difficultés que nous rencontrerons est liée au fait que le second jingle est joué différemment (l'ordre des notes diffère) suivant l'avancement de l'émission.

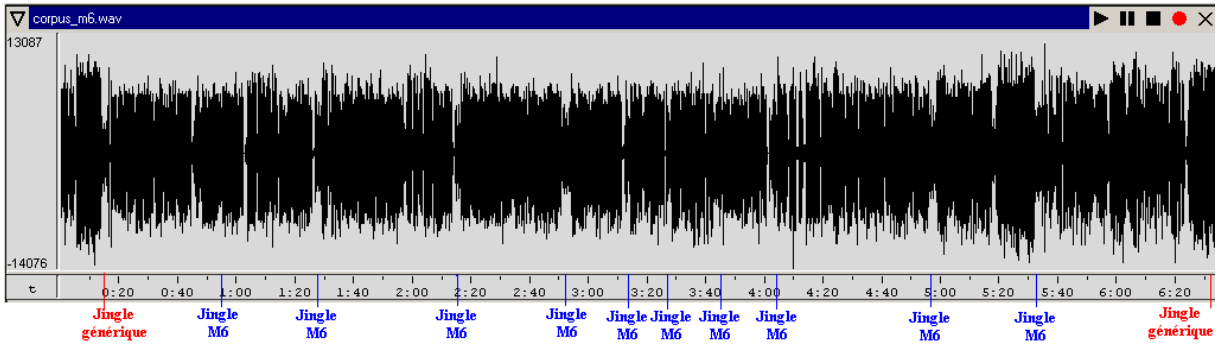

FIG. 7.3 – Repérage des jingles dans le « 6 minutes » de M6.

Sur chaque unité comprise entre deux jingles, une recherche de thèmes par mots clés doit permettre d'identifier le contenu. Le choix des thèmes et des mots clés est très important : ils sont forcément très liés au type de l'émission, à savoir le journal télévisé.

Pour un journal télévisé, les mots peuvent être :
- ministre, politique, général et président pour couvrir le thème « *politique* »,
- économique et tourisme pour couvrir le thème « *économie* »,
- victoire, championnat, tour de France et Formule 1 pour couvrir le thème « *sport* »,
- inondation, victime, attentat et catastrophe pour couvrir le thème « *catastrophe* ».

Un « mini thésaurus » doit être défini pour chacun des thèmes répertoriés.

Thèmes comme mots clés ne sont pas forcément présents dans chaque émission du « 6 minutes », ils sont simplement recherchés comme des indicateurs potentiels.

Le choix du nom du thème importe peu ici : il ne s'agit pas de faire une étude sur la détection de thème mais plutôt d'avoir une idée de ce que représente chacun des segments ou reportages. Il appartient à la communauté travaillant sur les données textuelles de nous fournir par la suite les thèmes et thésaurus associés.

Sur la figure 7.4, un étiquetage manuel de chacun des segments d'une occurrence du « 6 minutes », délimités par les jingles, est effectué ; il nous servira de vérité « terrain » lors de notre expérimentation.

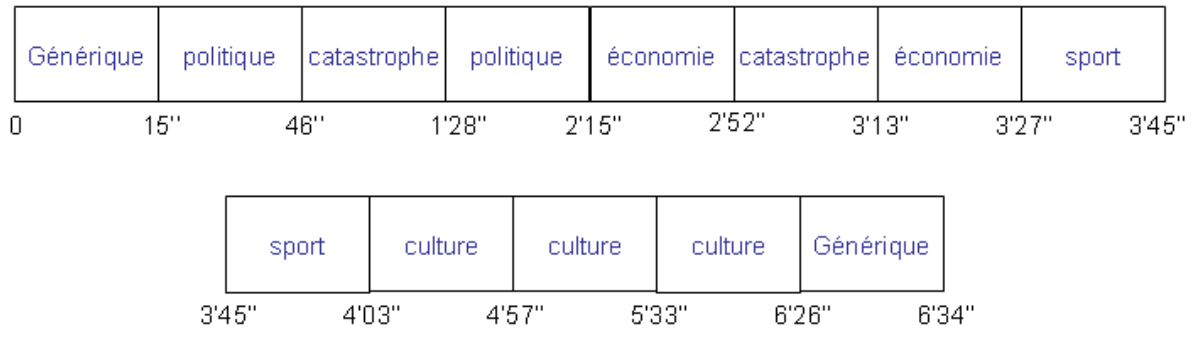

FIG. 7.4 – Etiquetage manuel en thèmes d'une occurrence du « 6 minutes » de M6 (l'échelle de durée n'est pas respectée).

Tous les segments annotés ne correspondent pas à des thèmes ! Nous avons ajouté deux types de segments :

- les segments annotés « Générique » correspondent à ce qui se trouve au début et à la fin de l'émission : ils sont délimités par le jingle « JG »,
- les segments annotés « culture » correspondent à des zones contenant de la musique pendant une durée d'au moins 5 secondes.

Afin d'arriver à une structuration automatique du journal « 6 minutes » comparable à l'étiquetage manuel effectué auparavant, plusieurs pré-traitements sont nécessaires. Les détections préalables de parole et de musique sont nécessaires pour deux raisons majeures : la détection de mots clés ne se fait que sur les zones de parole et un segment contenant de la musique est considéré comme un reportage culturel (musique, loisirs, divertissements...).

En conclusion, la structuration automatique du journal de M6 s'effectue en appliquant la procédure suivante :

- **Détection de jingles**

  Nous avons extrait manuellement deux jingles de référence, « JG » pour le jingle du générique et « J » pour le jingle M6, afin de retrouver les deux types de jingles présents dans le « 6 minutes ». La signature de ces deux jingles est effectuée à travers la première occurrence de chacun d'eux, dans une production du « 6 minutes ».

  La détection est ensuite automatique, sans apprentissage, et s'effectue par une comparaison spectrale tout au long du flux (cf. chapitre 4). Le premier et le dernier segments, délimités par les jingles, sont alors annotés « Générique ». Par défaut, les segments inférieurs à une seconde sont lissés. La figure 7.5 présente les résultats que nous avons obtenus sur une émission du « 6 minutes ».

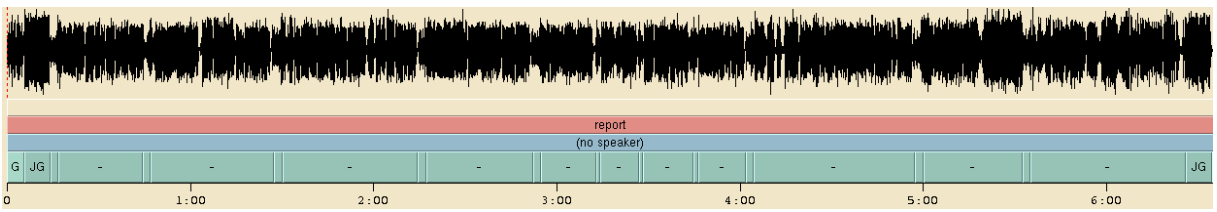

FIG. 7.5 – Détection automatique de jingles sur le « 6 minutes » de M6 avec « G » pour « Générique », « JG » pour « jingle_M6_generique » et « J » pour « jingle_M6 » (petits intervalles).

– **Détections de parole et de musique**

L'utilisation de notre système robuste de classification PMB permet à cette tâche de ne nécessiter aucun apprentissage (cf. chapitre 3). La détection est réalisée sur tous les segments sauf le premier et le dernier déjà étiquetés. Les segments contenant de la musique sont annotés « musique » (au moins cinq secondes de musique sont nécessaires). Les autres, s'ils contiennent de la parole, sont annotés « parole » (cf. figure 7.6).

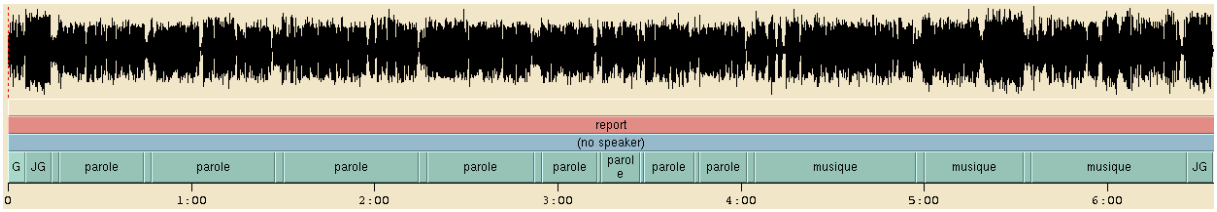

FIG. 7.6 – Détection de parole et de musique sur le « 6 minutes » de M6.

– **Détection de mots clés**

De par la stratégie choisie pour modéliser les mots clés, concaténation de phonèmes, cette tâche s'effectue également sans nouvel apprentissage sur les segments correspondant à de la parole (cf. chapitre 6). Les segments sont annotés par le thème correspondant aux mots clés dont la liste a été donnée précédemment (cf. figure 7.7).

Si il y a ambiguïté, le thème le plus représentatif est affecté au segment analysé : le thème qui possède le nombre de mots détectés le plus important, sur cet intervalle, est choisi. Si aucun mot clé n'est détecté, le segment est étiqueté « divers ». En cas d'égalité entre deux thèmes, l'annotation des deux thèmes est effectuée.

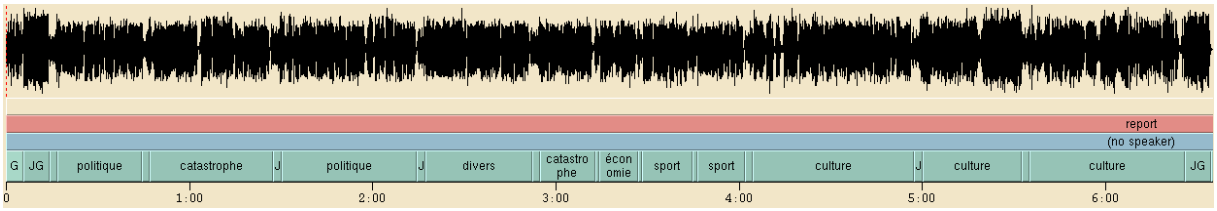

FIG. 7.7 – Détection de thèmes sur les zones de parole du « 6 minutes » de M6.

La comparaison entre les étiquetages manuel et automatique donne des conclusions intéressantes (cf. figures 7.4 et 7.7). Tous les jingles sont repérés, aucune omission ni aucune insertion n'est observée, le découpage du journal en reportages est parfait : les génériques sont bien positionnés.

La détection de musique donne de très bons résultats et permet d'annoter correctement les segments de culture.

La détection des mots clés est imparfaite mais nous permet néanmoins de donner une étiquette valide, en termes de « thème », à chacun des segments. La seule erreur présente correspond à aucune détection de mot clé et le thème « divers » est affecté à un segment de type « économie » (cf. tableau 7.1).

TAB. 7.1 – *Comparaison des détections automatique et manuelle de mots clés et de thèmes pour une émission du « 6 minutes ». Les erreurs sont en gras.*

Thème (manuel)	Thème (auto)	Mots clés (manuel)	Mot clé (auto)
Politique	Politique	ministre	ministre
Catastrophe	Catastrophe	catastrophe(3) inondation	catastrophe(3) inondation **victime**
Politique	Politique	général(2) ministre attentat	**général** ministre attentat
Économie	**Divers**	économique(2) tourisme	
Catastrophe	Catastrophe	victime	**victime(2)**
Économie	Économie	économique	économique
Sport	Sport	tour de France	tour de France
Sport	Sport	Formule 1	Formule 1

Cet exemple de structuration du journal télévisé le « 6 minutes » est très intéressant car, une fois que l'on possède une occurrence des jingles séparateurs des reportages de l'émission, la totalité de la chaîne de traitement est automatique et ne nécessite aucun apprentissage. Le choix des mots clés, comme nous l'avons déjà dit, est primordial afin de couvrir au mieux les thèmes recherchés ; cette étude n'est actuellement pas de nos compétences.

7.3 Structuration : perspectives

Les éléments de structuration que nous avons proposés reposent sur des outils monomédias, fondés uniquement sur la composante sonore, alors que nous traitons des documents multimédias (audiovisuels). Il semble indispensable de réfléchir sur les apports mutuels de l'analyse de la composante vidéo et de celle de la composante audio. Nous allons présenter l'apport possible de techniques classiques de vidéo aux deux exemples de structuration présentés dans la section précédente : recherche de motif et structuration d'un journal télévisé. Ensuite, le besoin d'une fusion audiovisuelle plus performante où il ne s'agit plus de faire une fusion « aveugle » des différentes sorties des outils d'analyse, est abordé. Enfin, une réflexion est menée sur la nécessité d'obtenir une macrosegmentation entièrement automatisée où les motifs sont découverts de manière non supervisée avant d'être localisés.

7.3.1 Apports de la vidéo

Alors que les deux exemples de structuration développés dans la section précédente sont obtenus à partir d'une analyse audio, une analyse vidéo aurait pu conduire à un résultat comparable. Les outils vidéo concernés sont :

- la détection de logos,
- l'extraction de texte,
- la reconnaissance de l'intervenant.

Cette coexistence amène à s'interroger sur la complémentarité des outils audio et vidéo, sur leur corrélation et leur fusion.

7.3.1.1 Détection de logos

Cette détection correspond à une détection d'objets visuels dans les images. Elle peut s'effectuer à partir de la détection d'incrustations dans la vidéo [Dem00], et suppose alors que les incrustations sont des zones rectangulaires. La détection s'effectue au moyen de filtres morphologiques appliqués aux dérivées temporelles.

Une autre méthode consiste à segmenter l'image. Soffer [Sof98] suggère de trouver pour chaque composante du logo, la composante de l'image qui est la plus proche dans l'espace des caractéristiques choisies (couleur, texture, forme, agencement spatial...). Le degré d'inclusion du logo dans l'image est alors défini comme la moyenne des distances correspondantes, sur toutes les composantes du logo.

Dans le cas de la structuration de journal télévisé, l'extraction du logo nous renseigne sur la chaîne émettrice et l'origine des images présentées.

Pour le « 6 minutes » de M6, le logo « M6 » (cf. figure 7.8) est présent en même temps que nos jingles (« Jingle M6 » et « Jingle Générique »), c'est-à-dire au début de l'émission, à la fin de l'émission et entre chaque reportage (cf. figure 7.9).

FIG. 7.8 – Le « logo M6 ».

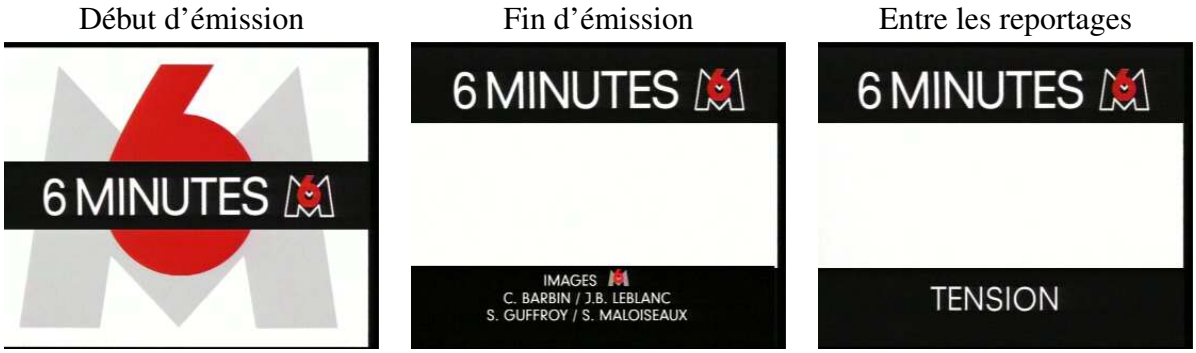

FIG. 7.9 – Apparition du « logo M6 » durant le « 6 minutes » de M6.

Les deux informations « Logo » et « Jingle » sont synchrones et de même nature. Les détections vidéo et audio reposent pour l'essentiel sur des fonctions à seuil et/ou des phases de détection et d'identification. Une analyse simultanée des deux flux doit permettre une sorte de relaxation des méthodes et des seuils.

Par exemple, si la détection de jingles fournit un certain nombre de candidats potentiels avec l'attribut d'un score de confiance, la détection de logo ne s'effectuera que sur les zones, en terme de temps, où les jingles potentiels sont présents. Cette détection sera plus ou moins précise suivant la valeur de confiance du jingle. Une fois le traitement effectué, l'information recueillie sur la vidéo (présence ou absence de logos) dirigera la phase d'identification du jingle.

7.3.1.2 Extraction de texte

L'extraction de texte dans les documents audiovisuels repose sur l'analyse de chaque image représentative d'un plan. Une binarisation de l'image est effectuée et un traitement permet de compenser les effets de distorsion et rétablir l'orientation du texte. Une méthode de détection des régions textuelles est appliquée à l'image binaire [Kim97]. Les régions détectées sont ensuite segmentées en régions de taille plus petite de l'ordre de la ligne de texte puis du caractère [Lu95]. Un traitement OCR est finalement effectué ; l'utilisation de dictionnaires ou de grammaires permet de corriger les caractères mal identifiés.

Un outil d'extraction de texte trouve des applications évidentes et très intéressantes pour nos deux exemples de structuration :

– pour la collection d'émission, le titre de l'émission, les noms du présentateur et de(s) invité(s) sont ainsi repérés (cf. figure 7.10),

FIG. 7.10 – Apport de l'extraction de texte sur l'émission le « Grand Échiquier ».

– pour le journal de M6 l'information extraite est complémentaire de la détection de thèmes. Un mot de type « mot clé » (« tension » dans l'exemple de la figure 7.11) apparaît en même temps que le logo « M6 » et l'audition du jingle. Durant les reportages, des noms de personnes et de lieux sont incrustés (« Izo dance'n effeckt » et « Aéroport d'Orly » dans la figure 7.11).

L'analyse vidéo apporte des informations complémentaires qui permettent d'enrichir le contenu des séquences détectées en audio :

– si, comme l'indique le générique du « Grand Échiquier » (cf. figure 7.10), l'orchestre national de France est attendu, de grandes plages musicales doivent être retrouvées,
– si le nom de l'intervenant ou la localisation du reportage est extrait de la vidéo, la recherche de thème du reportage en sera facilitée.

FIG. 7.11 – Apport de l'extraction de texte sur le « 6 minutes », journal de M6.

7.3.1.3 Reconnaissance de l'intervenant

La reconnaissance des individus est un sujet actuellement très étudié en analyse d'images. Il a atteint une certaine maturité après trente ans de recherche et il est au cœur de nombreuses applications. Cette tâche est étroitement liée à la reconnaissance de visages.

Différentes stratégies ont vu le jour. Zhao [Zha03] présente un état de l'art très complet sur la reconnaissance de visages. D'autres travaux sont liés à la reconnaissance de l'intervenant à travers son costume [Jaf04].

La reconnaissance de l'intervenant par la vidéo peut permettre d'améliorer ou de confirmer la détection du présentateur par la piste sonore (cf. figure 7.12) et réciproquement. Il s'agit là encore de fusionner les informations en sortie des détecteurs audio et vidéo ; ce travail fait l'objet d'études récentes [Alb02].

FIG. 7.12 – Détection de Jacques Chancel, le présentateur du « Grand Échiquier » par une analyse de la vidéo.

Des exemples immédiats de complémentarité certaine des deux flux d'information peuvent être exhibés ; ils sont liés à une défaillance de l'une des deux analyses :

– la reconnaissance d'un intervenant par la composante vidéo devient très intéressante dès lors que l'analyse de la piste sonore échoue. Il est actuellement très difficile voire impossible de détecter un chanteur par les techniques classiques de reconnaissance de locuteur alors que sur un intervalle de musique assez long, de la durée d'une chanson, le personnage, principalement et aisément détecté par l'analyse vidéo, correspond naturellement à l'interprète (cf. figure 7.13).

FIG. 7.13 – Détection d'un intervenant caractéristique : le chanteur.

– réciproquement la détection de l'intervenant par reconnaissance de visage devient très difficile dès que celui-ci se présente de profil ou que le plan devient panoramique. Sa présence sera révélée grâce à l'analyse audio.

Néanmoins, cette coopération est, la plupart du temps, loin d'être évidente et pose plus de problèmes qu'elle ne semble en résoudre. L'un des principaux est lié à la non synchronisation des apparitions d'une même personne sur la bande audio et sur la bande vidéo :

– il n'est pas rare d'entendre une personne (que ce soit l'intervenant principal ou non) alors que les scènes vidéo alternent des prises de vue de l'environnement (le public par exemple), des zooms sur des éléments particuliers (une personne du public par exemple), ou bien sur des plans sur le locuteur en question.
– dans un dialogue, il est fréquent que chaque prise de parole ne coïncide pas avec un changement de plan. Ce déphasage est même certainement souhaitable au niveau de la production du document en termes de qualité pour le spectateur.

Il apparaît clairement que la corrélation entre les analyses vidéo et audio dans ce cadre nécessite une réflexion beaucoup plus élaborée sur leur complémentarité réelle.

7.3.2 Macrosegmentation automatique

Lors de la section 7.2.1, nous avons repéré un motif sonore récurrent dans la collection du « Grand Échiquier », à savoir : *présentateur / [applaudissements] / spectacle / [applaudissements / spectacle] / applaudissements / présentateur*. Le choix de ce motif s'est effectué après une intervention humaine, à savoir l'écoute de deux émissions du « Grand Échiquier ». Il s'est avéré que ce motif a pu être retrouvé tout au long des émissions et a été ainsi validé.

Il serait intéressant de pouvoir trouver de manière automatique le motif récurrent d'une émission. Le principe serait le suivant :

- annotation automatique à partir des outils d'analyse audio et vidéo,
- recherche automatique des suites récurrentes dans la succession des annotations,
- inférence d'un motif,
- structuration du document à partir du motif trouvé.

La deuxième étape s'apparente à une détection d'invariants audiovisuels et fait l'objet de recherche récentes [Hai04].

Cette étude est d'autant plus importante que le motif ou l'invariant recherché n'est pas forcément quelque chose de facile à trouver manuellement.

7.4 Conclusion

Au cours de ce dernier chapitre, nous avons examiné deux scénarios au cours desquels il est indispensable d'enchaîner les différents outils d'analyse sonore que nous avons présentés lors des chapitres précédents.

La structuration d'une émission télévisuelle de plateau, le « Grand Échiquier » faisant partie d'une collection, a été réalisée. Le motif récurrent, *présentateur / [applaudissements] / spectacle / [applaudissements / spectacle] / applaudissements / présentateur*, apparaissant 10 fois lors de l'émission, a été bien identifié et localisé. Cette tâche a fait intervenir les détections de parole, de musique, d'applaudissements et du présentateur cible : Jacques Chancel.

La structuration d'un journal télévisé, le « 6 minutes », a été effectuée également. Les détections de jingles entraînent un découpage du journal en reportages et une détection de thèmes, à travers les mots clés, caractérise chacun des reportages. Les détections de parole et de musique sont alors un pré-traitement indispensable.

Ce travail de structuration, par l'intermédiaire de ces deux scénarios, nous a conduit à une réflexion sur l'apport de la vidéo dans notre analyse essentiellement audio et sur la fusion en général. Mais celle-ci ne doit pas être envisagée sous le simple angle de la fusion des informations issues des analyses audio et vidéo. Pour dépasser ce raisonnement, il faut considérer l'information acquise lors du traitement d'un média comme une connaissance a priori pour l'analyse d'un autre média. Le problème de la reconnaissance de l'intervenant illustre bien ces propos.

Conclusion et perspectives

Dans le contexte de l'indexation sonore, nous avons étudié différentes composantes primaires, permettant une structuration audiovisuelle. Pour chacune de ces unités bas niveau, un détecteur automatique est développé afin de les extraire du continuum sonore.

Pour les spécialistes de l'audio, les composantes primaires correspondent souvent à la parole et à la musique. Afin de réaliser une telle détection, de nombreuses études existent car c'est le pré-traitement quasi indispensable de toute indexation sonore. Le système original que nous avons développé est fondé sur une fusion de quatre paramètres : la modulation de l'énergie à 4 Hertz, la modulation de l'entropie, le nombre de segments issus d'une segmentation automatique et la durée de ces mêmes segments. Les résultats obtenus sont très bons, plus de 90 % de classification correcte, mais surtout le système est très robuste. Nous l'avons utilisé sur des corpus très hétérogènes tout en obtenant des scores équivalents (cf. tableau 1). Il ne nécessite aucun nouvel apprentissage et/ou adaptation de ses seuils contrairement aux approches classiques fondées sur une analyse cepstrale et des modèles de mélanges de lois gaussiennes.

D'autres composantes primaires correspondent à des sons clés : nous avons étudié les jingles, les rires, les applaudissements, le locuteur cible et les mots clés.

Notre détecteur de jingles est excellent. Bien que la méthode soit assez simple, mesure de distance dans le domaine spectral, sur 10 heures de tests et 132 jingles à retrouver, nous n'avons observé que 2 omissions et aucune fausse alarme. Les erreurs apparaissent dans des conditions très particulières, par exemple là où la parole recouvre entièrement le jingle. Cette étude est d'autant plus intéressante qu'elle ne nécessite aucun apprentissage, seulement une occurrence du jingle à reconnaître.

Une première étude sur la détection de sons caractéristiques, tels les applaudissements et les rires, a permis de montrer la faisabilité d'une méthode fondée sur une analyse spectrale et des MMG. Dans le cas des applaudissements, les résultats sont excellents car les sons caractéristiques, utiles en vue d'une structuration, sont tous détectés. Par contre, les rires posent beaucoup plus de problèmes et nous obligent à revoir leur modélisation. Se basant sur les mêmes techniques, une recherche de locuteur cible, en l'occurrence le présentateur d'une émission, a été réalisée. Cette méthode est très encourageante avec plus de 90 % de classification correcte : 2 à 3 minutes de signal, correspondant à la personne cible, sont nécessaires pour appliquer la recherche à toutes les émissions faisant intervenir cette même personne.

Une dernière unité a été recherchée : le mot clé. La reconnaissance de mots clés est un sujet très étudié depuis une quinzaine d'années. Pour de nombreux chercheurs, elle doit concurrencer la transcription automatique complète très coûteuse. Nous ne nous sommes pas placés dans cette optique, mais plutôt dans le sens où le mot clé est une composante primaire supplémentaire

qui doit apporter une information lors de la structuration du document traité. Pour cela, la détection, comme pour les autres unités bas niveau, doit être robuste et applicable à tout type de corpus. Les modèles statistiques n'ont nécessité aucun apprentissage et, bien que les résultats de détection de mots clés ne soient pas excellents, la recherche de thèmes qui en découle est performante.

Le tableau 1 permet d'avoir une vision de l'ensemble des expériences réalisées au cours de cette étude. Les besoins en apprentissage et en adaptation sont présentés pour chacun des outils ainsi que les résultats sur les différents corpus.

TAB. 1 – *Ensemble des expériences et apprentissages réalisés avec chacune de nos méthodes. Les croix représentent l'apprentissage. Les valeurs en italique correspondent au volume d'apprentissage utilisé. Pour les classification PMB, la première valeur est le résultat de la détection de parole et la seconde celle de la détection de musique.*

Méthode (*APP*/RECO)	PMB (Base)	PMB (Fusion)	Jingle	App	Rires	JC	Thèmes (mots clés)
MULTEXT (*35min*) Base perso (*35min*)		X X					
AGIR (*6h*/12h)	X 93,9 % 91 %						
RFI (*2h*/6h)	X 90,9 % 87 %	90,5 % 89 %	100 %				91,4 % (56,6 %)
ESTER (*30h*/5h)		97,6 % 89,2 %					X
FERIA (*3h*/6h)		94,6 % 93,9 %		X 98,6 %	X 97,3 %	X 89,7 %	
France 3 (*15min*) M6 (15min) CANAL+ (30min) France Info (1h) Pubs (1h30min)			X 100 % 100 % 91,7 % 97 %				

À la suite des détections de ces différentes composantes primaires, deux études en structuration ont été effectuées. La première, une détection de motif sur une collection d'émissions, permet de mettre en parallèle nos différentes détections afin d'extraire un enchaînement récurrent dans des émissions issues de la collection le « Grand Échiquier ». Il s'agit de la séquence : présentateur / [applaudissements] / spectacle / [applaudissements / spectacle] / applaudissements / présentateur. Durant notre émission de test, les 10 occurrences de ce motif sont toutes détectées.

La seconde étude est une structuration du journal télévisé le « 6 minutes » en thèmes. La segmentation en reportages s'effectue par une détection de jingles et le thème attribué est le résultat d'une détection de mots clés. Un pré-traitement, détections de parole et musique, est également nécessaire. Les résultats sont prometteurs : un seul reportage est mal annoté.

Ces premiers travaux de structuration sont encourageants, mais il est fort dommage de se limiter à l'analyse d'un seul media (le son dans notre cas), alors que nous exploitons des bases de données audiovisuelles. C'est pourquoi nous avons développé quelques pistes de réflexion sur l'apport de l'analyse vidéo.

Un couplage simple des analyses audio et vidéo représente nos perspectives à court terme. L'extraction de texte, la détection de logos et la reconnaissance d'intervenants sont les études que nous mènerons en priorité. Cette recherche sera facilitée du fait de notre environnement : des travaux sont actuellement en cours au sein de l'équipe SAMoVA sur l'analyse de la vidéo.

Cette immersion dans l'analyse de la vidéo doit nous permettre de mieux cerner les complémentarités entre l'audio et la vidéo et d'appréhender à moyen terme le vrai problème du traitement audiovisuel. Il est impératif de savoir répondre plus précisément à des questions classiques du type :

- qu'est ce qu'une information audiovisuelle ? Qu'est ce qu'une indexation audiovisuelle ? La présence d'un personnage est une illustration simple de ce type d'information, voire d'index. Peut-on généraliser cette démarche ?
- qu'est ce qu'une analyse audiovisuelle ? Sachant que, comme nous avons essayé de le montrer, une analyse audiovisuelle ne signifie pas une simple fusion d'informations issues d'une analyse audio et d'une analyse vidéo.

Ces axes de recherche sont très prometteurs et si des réponses positives apparaissent, leur potentiel applicatif est très large.

Annexe A

Le logiciel Transcriber

Sommaire

A.1 Présentation

Transcriber est un logiciel d'aide à l'annotation de signaux de parole. Il permet de segmenter des enregistrements de longue durée, de les transcrire et de marquer les tours de parole (changement de locuteur), la segmentation thématique et les conditions acoustiques. Il est plus spécialement conçu pour créer des corpus nécessaires au développement de systèmes en traitement de la parole.

Transcriber est distribué en logiciel libre sous licence GNU (http ://www.gnu.org/) : licence publique générale.

Transcriber - Copyright (C) 1998-2001, DGA
http ://www.etca.fr/CTA/gip/Projets/Transcriber/
Auteur : C. Barras
Coordinateurs :
Edouard Geoffrois, DGA/DCE/CTA/GIP
Mark Liberman et Zhibiao Wu, LDC

A.2 Utilisation

Nous utilisons ce logiciel pour faire l'étiquetage manuel des fichiers sonores servant à l'apprentissage de nos Modèles de Mélanges de lois gaussiennes pour le système de base PMB (cf. chapitre 2) et pour les systèmes de détection des applaudissements et des rires (cf. chapitre 5).

Ce logiciel nous permet aussi de visualiser les résultats de classification automatique de chacun de nos systèmes.

Description de nos entrées/sorties :

- entrée : fichier son (format wav, 16 bits, 16 kHz, mono),
- sortie : fichier xml (correspondant à la Définition de Type de Documents -DTD- de Transcriber).

A.2.1 Exemple d'étiquetage

La figure A.1 est un exemple d'affichage effectué par le logiciel sur un fichier de 2 minutes. L'étiquetage concerne la discrimination parole/non-parole.

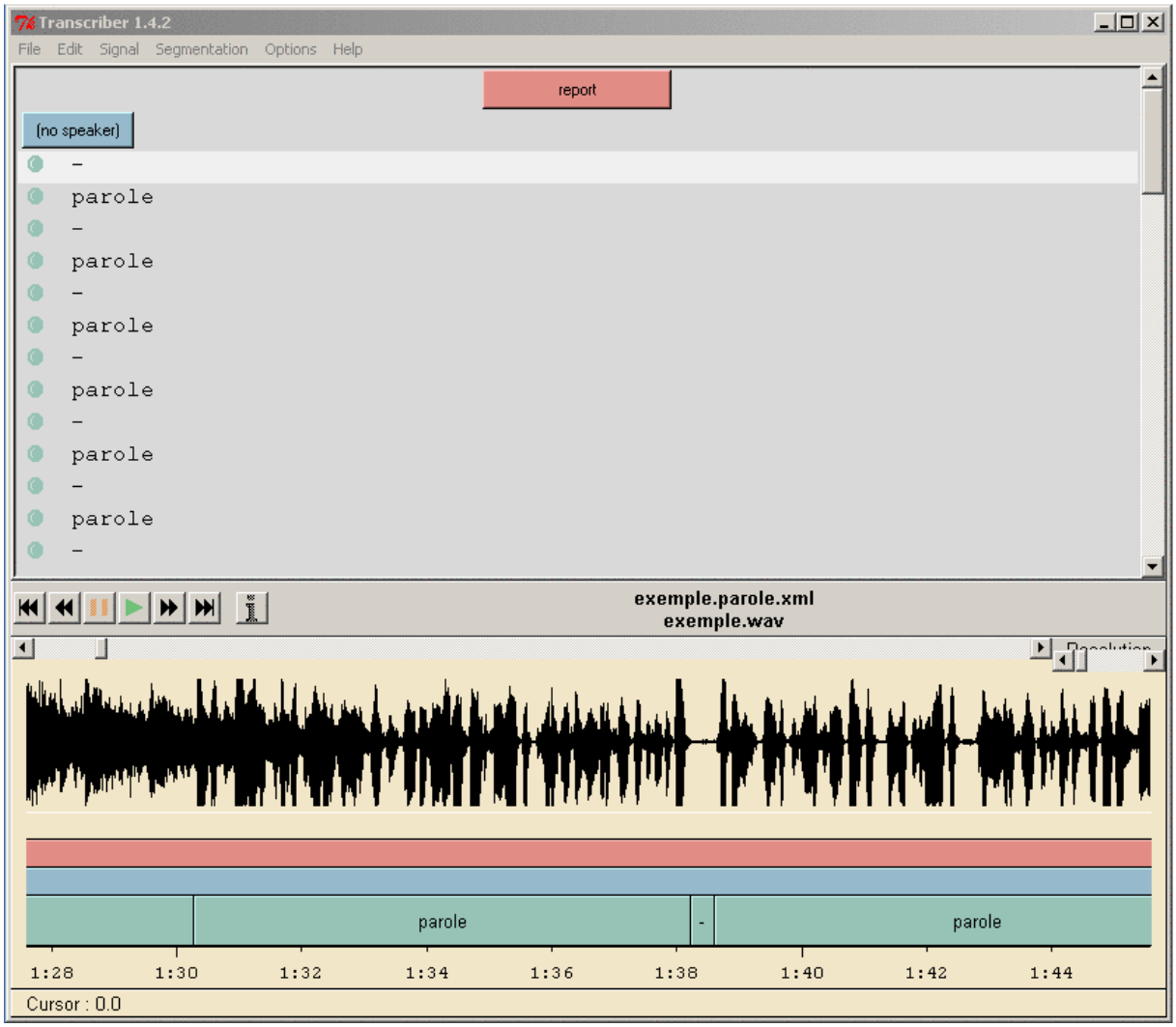

FIG. A.1 – Exemple d'étiquetage manuel par le logiciel Transcriber.

A.2.2 Exemple de fichier de transcription

Voici le fichier de transcription (xml) correspondant au signal précédent (exemple.wav) et à la DTD (trans-13.dtd).

< ?xml version="1.0" encoding="ISO-8859-1" ?>
< !DOCTYPE Trans SYSTEM "trans-13.dtd">
<Trans scribe="Julien" audio_filename="exemple" version="2" version_date="041004">
<Episode>
<Section type="report" startTime="0" endTime="120.000">
<Turn startTime="0" endTime="120.000">
<Sync time="0"/>
-
<Sync time="12.09"/>
parole
<Sync time="13.023"/>
-
<Sync time="28.284"/>
parole
<Sync time="42.277"/>
-
<Sync time="90.285"/>
parole
<Sync time="98.233"/>
-
<Sync time="98.606"/>
parole
<Sync time="107.002"/>
-
<Sync time="107.301"/>
parole
</Turn>
</Section>
</Episode>
</Trans>

Annexe B

Résultats complémentaires pour la détection de parole et de musique

Sommaire

B.1 Présentation

Cette annexe présente des résultats d'expériences concernant la détection de parole et de musique par le système présenté dans le chapitre 3.

Différents corpus sont utilisés. Le premier (RAIVES), utilisé dans le chapitre 3 afin de montrer la robustesse des paramètres, est plus détaillé ici. Les deux autres (ESTER et FERIA) décrivent l'homogénéité des résultats.

B.2 Corpus projet RAIVES

Ce corpus a déjà été présenté dans la section 3.4.1. Il s'agit d'enregistrements de RFI. Différents types d'émissions sont présentes : des informations et des reportages notamment. Les résultats sont très bons pour les informations et un peu moins bons pour les reportages (cf. tableau B.1) à cause d'un fond musical très variable.

TAB. B.1 – *Taux de classification correcte en parole/non-parole et en musique/non-musique pour le corpus RAIVES.*

Émissions	Durée	Performances (parole)	Performances (musique)
Information 1	24 min	96,88 %	91,19 %
Information 2	24 min	97,67 %	96,91 %
Information 3	24 min	96,22 %	89,96 %
Information 4	1 h	97,21 %	89,60 %
Information 5 (anglais)	1 h	96,86 %	88,71 %
Information 6 (espagnol)	1 h	97,03 %	91,32 %
Reportage 1	25 min	88,05 %	83,21 %
Reportage 2	24 min	84,31 %	79,77 %
Reportage 3	1 h	91,53 %	85,38 %
Total	**6 h**	**90,5 %**	**89 %**

B.3 Corpus campagne d'évaluation ESTER

Il s'agit d'un corpus de 4 heures et 40 minutes composé d'émission de radio issues de France-Inter et de RFI. Ces émissions ont servi de test « à blanc » pour la campagne d'évaluation ESTER. Elles sont principalement composées d'informations : ce qui explique les excellents résultats en détection de parole. Les résultats en détection de musique sont comparables au corpus précédent (cf. tableau B.2).

TAB. B.2 – *Taux de classification correcte en parole/non-parole et en musique/non-musique pour le corpus ESTER.*

Émissions	Durée	Performances (parole)	Performances (musique)
France-Inter1	1 h	98,56 %	98,03 %
France-Inter2	1 h	99,48 %	97,75 %
France-Inter3	20 min	97,19 %	96,09 %
France-Inter4	20 min	95,91 %	89,10 %
RFI1	1 h	96,74 %	81,74 %
RFI2	1 h	94,67 %	76,91 %
Total	**4 h 40 min**	**97,58 %**	**89,18 %**

B.4 Corpus projet FERIA

Ce corpus est décrit dans la section 5.3.1. Comme pour les sons clés, nous n'avons utilisé que 2 émissions pour l'évaluation des résultats (parmi les 54), car l'étiquetage manuel est fastidieux. Ces 2 émissions du « Grand Échiquier » (émissions de type « plateau ») ont la particularité de contenir de longs passages musicaux et beaucoup d'interviews. Les résultats sont meilleurs dans l'ensemble (cf. tableau B.3) et notamment pour la détection de musique, car les zones de musique sont bien marquées : il s'agit de chansons et il n'y a pas de fond musical qui pose problèmes comme dans les autres corpora.

TAB. B.3 – *Taux de classification correcte en parole/non-parole et en musique/non-musique pour le corpus FERIA.*

Émissions	Durée	Performances (parole)	Performances (musique)
Émission1	3 h 10 min	94,99 %	94,03 %
Émission2	3 h 10 min	94,28 %	93,75 %
Total	**6 h 20 min**	**94,64 %**	**93,89 %**

Annexe C

Algorithme VQ (Quantification Vectorielle)

Sommaire

C.1 Objectif

La quantification vectorielle consiste à extraire un « dictionnaire » de « prototypes » (ensemble des centroïdes) d'un grand ensemble représentatif de données. Le dictionnaire doit respecter le mieux possible leur répartition dans l'espace.

La première version de l'algorithme de construction du dictionnaire pour la quantification est connue sous le nom de Lloyd [Llo57] et fut utilisée pour la quantification scalaire. Cet algorithme a ensuite été généralisé pour la classification automatique et la reconnaissance des formes sous le nom d'algorithme des « K-means » ou méthode des « nuées dynamiques » [Did76].

C.2 Algorithme des K-means

$(y_n),\ 0 \leq n \leq N$ représente un nuage de points (observations) de R^d, d est la distance euclidienne et la taille du dictionnaire K est fixée.

1. **Initialisation**
   Soit un dictionnaire D_0 de taille K.

2. **Construction de la partition**
   À la tième itération, le dictionnaire est noté :

$$D_t = \{D_{i,t}\}_{i=1,\cdots,K} \tag{C.1}$$

   La partition qui minimise l'erreur de quantification associée à D_t est composée des classes :

$$C_{i,t} = \{y_n / d(y_n, D_{i,t}) \leq d(y_n, D_{j,t}),\ j \neq i\} \tag{C.2}$$

   L'erreur de quantification vaut :

$$Dis_t = \frac{1}{N} \sum_{n=1}^{N} \left[\min_{i=1}^{K} d(y_n, \mu_{i,t}) \right] \tag{C.3}$$

   où μ_{it} est le centroïde de C_{it}.

3. **Test d'arrêt**
   Si $(Dis_{t-1} - Dis_t)/Dis_t < \epsilon$ alors l'algorithme est terminé, et le dictionnaire recherché est D_{t+1} composé des nouveaux centroïdes, soit :

$$D_{i,t+1} = \mu_{i,t} \tag{C.4}$$

   Sinon $t = t + 1$ et l'algorithme est repris à l'étape 2.

Puisque cet algorithme n'est que localement optimal, le choix du dictionnaire de départ est important. Une variante très utilisée de l'algorithme de Lloyd est l'algorithme LBG [Lin80] : il procède hiérarchiquement et réalise une sorte d'initialisation itérative au cours de la construction.

C.3 Algorithme LBG (Linde, Buzo, Gray)

Le but est de construire un dictionnaire de taille K, où $K = 2^p$.

1. **Initialisation**
   Le centre de gravité de l'ensemble d'apprentissage est calculé.
   Soit d_0 ce vecteur. Le dictionnaire est constitué de d_0, $p = 0$.

$$D_0 = \{d_0\}, \quad |D_0| = 2^p \tag{C.5}$$

2. **Éclatement « Splitting »**
   Tous les éléments d en nombre 2^k du dictionnaire sont « éclatés » en deux vecteurs. Ceci se fait par exemple en transformant chaque d en $d+\epsilon$ et $d-\epsilon$, où ϵ est un vecteur aléatoire de variance adaptée aux points du nuage associés à d.

3. **Convergence**
   L'algorithme de Lloyd (cf. section précédente) est appliqué sur le dictionnaire des 2^{k+1} éléments ainsi constitués. Après convergence, un dictionnaire de 2^{k+1} éléments « optimal » est obtenu.

4. **Arrêt**
   $k = k + 1$.
   Si $k > k_0$ fixé à l'avance, alors l'algorithme prend fin, sinon le processus est itéré (2).

   Le test d'arrêt peut se faire aussi par rapport à un seuil minimal sur la distorsion des données d'apprentissage par rapport au dictionnaire, comme dans le cas de l'algorithme de Lloyd.

Annexe D

Algorithme EM (Expectation Maximisation)

Sommaire

D.1 Rappels

L'expression de la vraisemblance d'une observation y de l'ensemble d'apprentissage, supposée réalisation d'un modèle de mélanges de lois gaussiennes, est donnée par :

$$\sum_{k=1}^{N} \nu_k . N(y, \mu_k, \Sigma_k) \tag{D.1}$$

avec :

$$N(y, \mu_k, \Sigma_k) = \frac{1}{(2\pi)^{d/2} |\Sigma_k|^{1/2}} \, exp^{\left[-\frac{1}{2}(y-\mu_k)^t \Sigma_k^{-1}(y-\mu_k)\right]} \tag{D.2}$$

et :
N le nombre de composantes du mélange,
ν_k le poids de chaque composante,
μ_k la moyenne de chaque composante,
Σ_k la matrice de covariance associée.

L'algorithme EM est fondé sur la vraisemblance de chaque vecteur observé par rapport à chaque composante gaussienne du modèle.

D.2 Algorithme de base

1. **Initialisation (t=0)**

   – Initialisation des moyennes μ_k par N points extraits aléatoirement de l'ensemble des observations X. $X = \{x_1, \cdots, x_N\}$.

   – Initialisation de toutes les matrices de covariance Σ_k à la matrice unité I_p.

   – Initialisation équiprobable des poids des composantes : $\nu_k = 1/N$.

   OU

   – Utilisation de l'algorithme VQ (Quantification Vectorielle) présenté dans l'annexe C pour l'initialisation.

2. **Itération (t)**

   Pour tout $k = 1, \cdots, N$

   – *Phase d'estimation*

   Calcul de la probabilité P_{nk} que le vecteur y_n soit généré par la loi gaussienne k.

$$P_{nk} = \frac{\frac{\nu_k}{(2\pi)^{d/2}|\Sigma_k|^{1/2}}\ exp^{\left[-\frac{1}{2}(y_n-\mu_k)^t\Sigma_k^{-1}(y_n-\mu_k)\right]}}{\sum_{k'=1}^{K}\frac{\nu_{k'}}{(2\pi)^{d/2}|\Sigma_{k'}|^{1/2}}\ exp^{\left[-\frac{1}{2}(y_n-\mu_{k'})^t\Sigma_{k'}^{-1}(y_n-\mu_{k'})\right]}} \tag{D.3}$$

   – *Phase de maximisation*

   Réestimation des paramètres à partir des probabilités P_{nk} :

$$\overline{\nu_k} = \frac{1}{N}\sum_{n=1}^{N} P_{nk} \tag{D.4}$$

$$\overline{\mu_k} = \frac{\sum_{n=1}^{N} P_{nk} y_n}{\sum_{n=1}^{N} P_{nk}} \tag{D.5}$$

$$\overline{\Sigma_k} = \frac{\sum_{n=1}^{N} P_n k(y_n - \overline{\mu_k})(y_n - \overline{\mu_k})^t}{\sum_{n=1}^{N} P_{nk}} \tag{D.6}$$

   – *Incrémentation de t à $t+1$ et retour à la phase d'estimation*

3. **Arrêt de l'algorithme**

   Calcul de la vraisemblance des observations (y_n).

   Si la variation de la vraisemblance descend en dessous d'un seuil fixé alors l'estimation est terminée sinon l'estimation est reprise à l'étape 2.

Annexe E

Outils HTK

Sommaire

E.1 Introduction

HTK est une boîte à outils portable pour manipuler des Modèles de Markov Cachés. L'enchaînement des différents outils permet de construire un système de reconnaissance (ou un système de détection de mots clés par exemple) complet (cf. figure E.1) :

- le calcul des paramètres du signal,
- l'apprentissage des modèles,
- la reconnaissance.

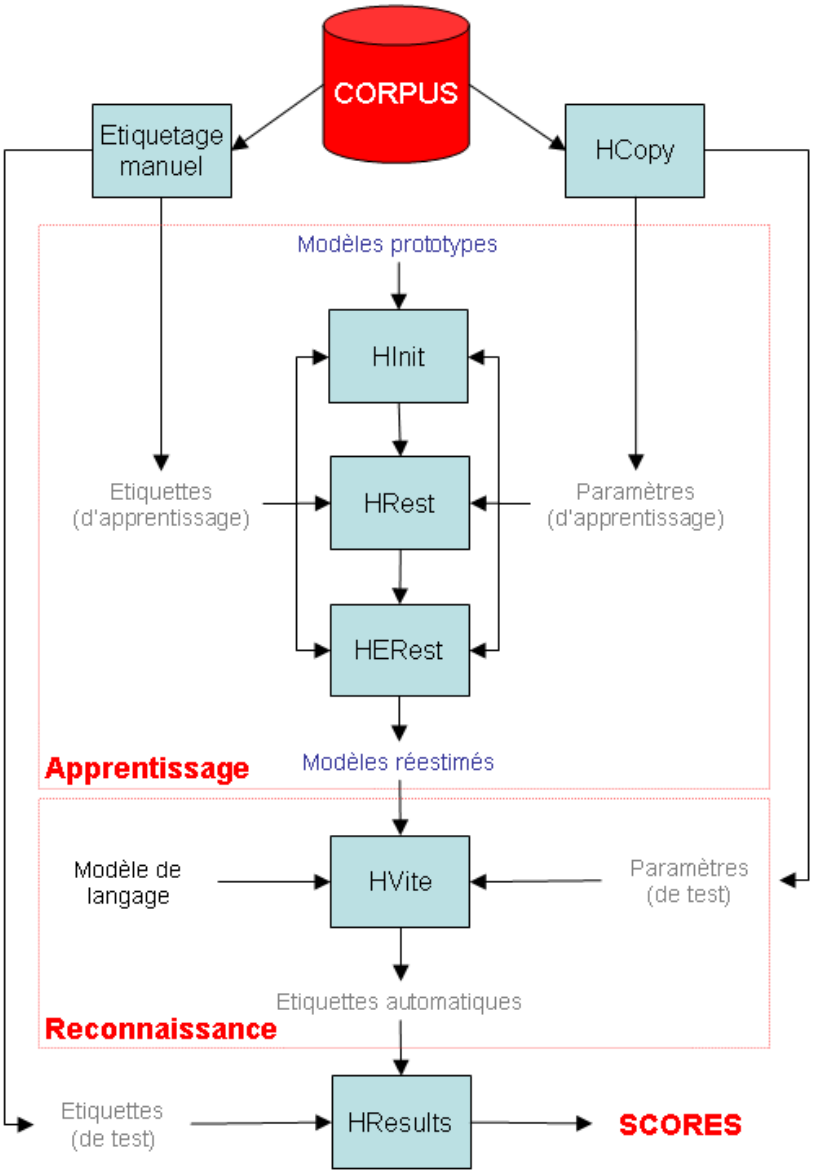

FIG. E.1 – Construction d'un système de reconnaissance à l'aide de HTK.

E.2 Paramétrisation

Cette phase sert à représenter le signal par des vecteurs de paramètres. Cette représentation est obtenue en utilisant l'outil **HCopy** qui va produire en particulier les coefficients PLP ou MFCC ainsi que l'énergie et les différentes variantes (dérivées premières, secondes, normalisations...).

E.3 Apprentissage des modèles

E.3.1 Présentation des modèles

Pour chaque unité acoustique, il faut définir un modèle prototype contenant la topologie choisie :

- le nombre d'états du modèle,
- les transitions possibles,
- la loi de probabilité associée à chaque état.

Remarques :

- *les états initial et final ont la particularité de ne pas émettre d'observation, mais de servir uniquement à la connexion des modèles de parole continue,*
- *les modèles peuvent avoir des prototypes différents.*

Les probabilités d'émission correspondent à des MMG, caractérisés par leur moyenne et leur matrice de covariance.

E.3.2 Étiquetage

Avant l'apprentissage des modèles, il est nécessaire de préparer les données d'apprentissage en calculant les paramètres du signal d'une part (cf. section E.2) et en étiquetant les phrases d'apprentissage d'autre part. Ces phrases doivent, bien sûr, être étiquetées en fonction des unités acoustiques modélisées : il faut que tous les segments correspondent aux unités. Cette tâche fastidieuse (segmentation manuelle en général) peut être limitée à une fraction de l'ensemble d'apprentissage. Cette tâche peut également être réalisée grâce à l'outil **HCompv** qui va parcourir un fichier de données, calculer la moyenne et la variance et fixer les valeurs des lois gaussiennes du MMC. Par l'intermédiaire de la campagne d'évaluation française **ESTER**, nous avons pu récupérer des corpus alignés phonétiquement.

E.3.3 Initialisation et réestimation des modèles (cf. annexe G)

Pour chaque unité acoustique, l'outil **HInit** initialise les probabilités d'émission des états du modèle par la procédure des « k-moyennes segmentales » fondée sur l'algorithme de Viterbi.

L'estimation des paramètres d'un modèle est affinée avec **HRest**, qui applique l'algorithme optimal de Baum-Welch jusqu'à la convergence et réestime les probabilités d'émission et de transition.

Ensuite, il est possible d'appliquer plusieurs itérations de l'outil **HERest** qui réestime simultanément l'ensemble des modèles sur de la parole continue non segmentée.

L'amélioration des modèles peut être effectuée par augmentation du nombre de lois gaussiennes servant à estimer la probabilité d'émission d'une observation dans un état.

E.4 Reconnaissance (cf. annexe F)

Le module de décodage de la parole **HVite** utilise l'algorithme de Viterbi pour trouver la séquence d'états la plus probable correspondant aux paramètres observés dans un modèle composite, et en déduire les unités acoustiques correspondantes. Le modèle composite autorise la succession des modèles acoustiques en fonction d'une syntaxe choisie préalablement. Celle-ci peut se situer au niveau phonétique ou lexical : les mots du lexique peuvent être définis par la concaténation d'unités sub-lexicales.

Le résultat du décodage est comparé à l'étiquetage manuel (la référence) par un alignement dynamique réalisé par **HResults**, afin de répertorier le nombre d'étiquettes identifiées, insérées, omises ou substituées et ainsi de calculer le taux de reconnaissance et l'accuracy.

Annexe F

Reconnaissance par l'algorithme de Viterbi

Sommaire

F.1 Reconnaissance

L'**algorithme de Viterbi** [Vit67] permet de rechercher la séquence d'états cachés la plus probable dans un MMC et calcule la probabilité d'émission le long de ce chemin.

La variable $\delta(t,i)$ est définie comme la probabilité maximale que les observations jusqu'à l'instant t aient été émises par le modèle de Markov caché λ en suivant un chemin dont l'état à l'instant t est l'état d'indice i :

$$\delta(t,i) = \max_{q_1 \cdots q_{t-1}} P(o_1 \cdots o_t, q_1 \cdots q_t = i|\lambda) \tag{F.1}$$

avec $O = (o_1 \cdots o_T)$ la suite d'observations et i le numéro de l'état atteint par le processus caché q à l'instant t.

Une récurrence sur le temps, en parallèle pour tous les états, s'applique, à laquelle s'ajoute la mémorisation du meilleur chemin au travers de la fonction . L'algorithme suivant est donné pour un modèle gauche-droit où seul l'état 1 est considéré comme état initial et l'état N comme l'état final ; en reprenant les notations de la section 6.2.2.1, on obtient :

– Initialement (état d'indice 1) :

$$\delta(1,i) = \begin{cases} b_1(o_1), & i = 1 \\ 0, & 1 < i \leq N \end{cases} \tag{F.2}$$

– Par récurrence sur les valeurs de $\delta(t,i)$:
Pour t variant de 2 à T,
Pour j variant de 1 à N,

$$\delta(t,j) = \max_{1 \leq i \leq N} \delta(t-1,i) a_{ij} b_j(o_t) \tag{F.3}$$

$$\Psi(t,j) = \arg \max_{1 \leq i \leq N} \delta(t-1,i) a_{ij} \tag{F.4}$$

– La probabilité du meilleur chemin est donnée par :

$$P(O, \widetilde{Q}|\lambda) = \delta(T,N) \tag{F.5}$$

et le meilleur chemin est :

$$\begin{aligned} \widetilde{Q} &= (\widetilde{q_1} \cdots \widetilde{q_T}) \\ \widetilde{q_T} &= N \end{aligned} \tag{F.6}$$

et :

$$\widetilde{q_{t-1}} = \Psi(t, \widetilde{q}_t) \tag{F.7}$$

La segmentation du signal fournie par l'algorithme de Viterbi sert aussi à l'initialisation des modèles d'apprentissage (cf. annexe G).

Annexe G

Apprentissage par l'algorithme de Baum-Welch

Sommaire

G.1 Introduction

Si O désigne une suite d'observations supposées émises par un modèle de Markov caché, λ, l'estimation des paramètres définissant λ par la méthode du maximum de vraisemblance consiste à choisir ces paramètres afin de rendre maximale la probabilité d'émission des observations O par le modèle :

$$\widetilde{\lambda} = \arg\max_{\lambda} P(O|\lambda) \tag{G.1}$$

Résoudre ceci de manière analytique est impossible ; les formules de Baum-Welch [Bau72] permettent une réestimation des paramètres $A = (a_{ij})$ et $B = (b_j(.))$ du modèle $\widetilde{\lambda_{i+1}}$ de telle sorte que le nouveau modèle $\widetilde{\lambda_{i+1}}$ vérifie :

$$P(O|\widetilde{\lambda_{i+1}}) \leq P(O|\widetilde{\lambda_i}) \tag{G.2}$$

La convergence vers un optimum local est démontrée. Les valeurs initiales des paramètres A et B deviennent primordiales afin d'assurer une convergence correcte et rapide vers le maximum global. L'algorithme de Viterbi sert à l'initialisation des modèles.

Dans la mesure où tous les modèles de Markov cachés utilisés en parole sont de type gauche-droit, il est nécessaire d'avoir plusieurs suite d'observations pour réestimer un modèle. Nous noterons $O = (o^1, \cdots, o^r, \cdots, o^R)$ l'ensemble de ces suites, ensemble d'apprentissage du modèle λ.

G.2 Initialisation des modèles

Les probabilités d'émission sont initialisées à partir de segmentations successives des suites d'observations :

- à la première itération, la segmentation est obtenue à partir d'un alignement uniforme des suites d'observations sur le modèle,
- aux itérations suivantes, la segmentation est issue de l'alignement des suites sur le modèle par l'algorithme de Viterbi.

À la suite de chaque segmentation, l'ensemble des observations associées à chaque trame permet d'estimer les paramètres de la loi d'émission correspondante. Si cette loi est un MMG, la procédure itérative des « k means » en fournit l'initialisation [Rab89].

G.3 Réestimation des modèles

La réestimation des paramètres du modèle λ est fondée sur le comptage du nombre moyen de transitions observées entre les états i et j.

Soient :

– la variable directe $\nu_t^r(i)$ définie comme la probabilité que la suite o^r d'observations jusqu'à l'instant t ait été émise par le modèle λ à N états, et que l'état à cet instant soit l'état d'indice i :

$$\nu_t^r(i) = P(o_1^r \cdots o_t^r, q_t = i|\lambda) \tag{G.3}$$

– la variable rétrograde $\beta_t^r(i)$ définie comme la probabilité que les observations o^r après l'instant t soient émises en partant de l'état d'indice i :

$$\beta_t^r(i) = P(o_{t+1}^r \cdots o_T^r|q_t = i, \lambda) \tag{G.4}$$

La probabilité de transition est réestimée par :

$$\widetilde{a_{ij}} = \frac{\sum_{r=1}^{R} \frac{1}{P_r} \sum_{t=1}^{T_r-1} \nu_i^r(t) a_{ij} b_j(o_{t+1}^r) \beta_j^r(t+1)}{\sum_{r=1}^{R} \frac{1}{P_r} \sum_{t=1}^{T_r} \nu_i^r(t) \beta_i^r(t)} \tag{G.5}$$

T_r est la longueur de la suite o^r et $P_r = P(o^r|\lambda)$.

Il est important de remarquer que la structure du modèle est conservée au cours de la réestimation : les transitions initialement interdites entre deux états le restent :

$$a_{ij} = 0 \Rightarrow \widetilde{a_{ij}} = 0 \tag{G.6}$$

Au niveau de la réestimation des probabilités d'émission, nous nous plaçons dans le cas de densités de probabilités continues représentées par un mélange de M lois gaussiennes multidimensionnelles. Le vecteur de moyenne associée à l'état i et à la gaussienne m du mélange est recalculé :

$$\widetilde{\mu_{jm}} = \frac{\sum_{r=1}^{R} \sum_{t=1}^{T_r} L_{jm}^r(t) o_t^r}{\sum_{r=1}^{R} \sum_{t=1}^{T_r} L_{jm}^r(t)} \tag{G.7}$$

avec $L_{jm}^r(t)$ qui représente la probabilité que l'observation o_t^r soit émise par la composante m de la loi d'émission b_j.

La matrice de covariance est réestimée de manière semblable :

$$\widetilde{\Sigma_{jm}} = \frac{\sum_{r=1}^{R}\sum_{t=1}^{T_r} L_{jm}^r(t)(o_t^r - \widetilde{\mu_{jm}})(o_t^r - \widetilde{\mu_{jm}})'}{\sum_{r=1}^{R}\sum_{t=1}^{T_r} L_{jm}^r(t)} \tag{G.8}$$

ainsi que le poids :

$$\widetilde{\nu_{jm}} = \frac{\sum_{r=1}^{R}\sum_{t=1}^{T_r} L_{jm}^r(t)}{\sum_{r=1}^{R}\sum_{t=1}^{T_r} L_{j}^r(t)} \tag{G.9}$$

Bibliographie

[Aig97] P. Aigrain, P. Joly, et V. Longueville. Medium Knowledge-Based Macro-Segmentation of Video into Sequences. *In Intelligent Multimedia Information Retrieval*, pages 159–173, 1997.

[Alb02] A. Albiol, L. Torrest, et E. J. Delp. Combining Audio and Video for Video Sequence Indexing Applications. Dans *IEEE International Conference on Multimedia and Expo*, tome 2, pages 353–356. Naples, Italy, juillet 2002.

[Alt03] H. Altinçay et M. Demirekler. Speaker Identification by Combining Multiple Classifiers Using Dempster-Shafer Theory of Evidence. *Speech Communication*, 2003.

[Ama01] R. Amaral, T. Langlois, H. Meinedo, J. Neto, N. Souto, et I. Trancoso. The Development of a Portuguese Version of a Media Watch System. Dans *European Conference on Speech Communication and Technology*, tome 4, pages 2689–2692. Aalborg, Denmark, septembre 2001.

[ANS01] ANSI. ISO/IEC 15938-4 Information Technology - Multimedia Content Description Interface - Audio. Rapport technique, MPEG, 2001.

[AO88] R. André-Obrecht. A New Statistical Approach for Automatic Speech Segmentation. *IEEE Transactions on Audio, Speech, and Signal Processing*, 36(1) :29–40, janvier 1988.

[AO93] R. André-Obrecht. *Segmentation et parole ?*. Thèse d'état, IRISA, 1993.

[AO97] R. André-Obrecht et B. Jacob. Direct Identification vs. Correlated Models to Process Acoustic and Articulatory Informations in Automatic Speech Recognition. Dans *IEEE International Conference on Audio, Speech and Signal Processing*, pages 989–992. Munich, Germany, 1997.

[Ari04] J. A. Arias. *Méthodes à vecteurs de support et indexation sonore*. Rapport de DEA, IRIT, Université Paul Sabatier, Toulouse III, juin 2004.

[Ata83] B. Atal. Efficient Coding of LPC Parameters by Temporal Decomposition. Dans *IEEE International Conference on Audio, Speech and Signal Processing*, pages 81–84. Boston, USA, avril 1983.

[Bak75] J. K. Baker. The DRAGON system - An overview. *IEEE Transactions on Audio, Speech, and Signal Processing*, 23(1) :24–29, 1975.

[Bak76] R. Bakis. Continuous Speech Recognition via Centisecond Acoustic States. Dans *91st Meeting of the Acoustical Society of America*. Washington, USA, avril 1976.

[Bau72] L. Baum. An Inequality and Association Maximization Technique in Statistical Estimation for Probabilistic Functions of a Markov Process. *Inequalities*, 3(1) :1–8, 1972.

[Bes98] L. Besacier. *Un modèle parallèle pour la reconnaissance automatique du locuteur*. Thèse de doctorat, Université d'Avignon, avril 1998.

[Bez93] O. Bezie et P. Lockwood. Beam Search And Partial Traceback In The Frame-Synchronous Two-Level Algorithm (TLBS). Dans *IEEE International Conference on Audio, Speech and Signal Processing*, pages 511–514. Minneapolis, USA, avril 1993.

[Bim88] F. Bimbot, G. Cholet, P. Deleglise, et C. Montacie. Temporal Decomposition and Acoustic-phonetic Decoding of Speech. Dans *IEEE International Conference on Audio, Speech and Signal Processing*, pages 425–428. Singapore, novembre 1988.

[Boi87] R. Boite et M. Kunt. *Traitement de la parole*. Presses Polytechniques Romandes, 1987.

[Boi93] J. M. Boite, H. Boulard, B. D'hoore, et M. Haesen. A New Approach Toward Keyword Spotting. Dans *European Conference on Speech Communication and Technology*, pages 1273–1276. Berlin, Germany, septembre 1993.

[Bur98] C. Burges. A Tutorial on Support Vector Machines for Pattern Recognition. *Data Mining and Knowledge Discovery*, 2(2) :121–167, 1998.

[Cae79] J. Caelen. *Un modèle d'oreille ; analyse de la parole continue ; reconnaissance phonémique*. Thèse de doctorat, IRIT, Université Paul Sabatier, Toulouse III, 1979.

[Cal89] Calliope. *La parole et son traitement automatique*. Masson, Paris, France, 1989.

[Cam97] J. Caminero, L. Hernandez-Gomez, C. De la Torre, et C. Martin. Improving Utterance Verification using Hierarchical Confidence Measures in Continuous Natural Numbers Recognition. Dans *IEEE International Conference on Audio, Speech and Signal Processing*, tome 2, pages 891–894. Munich, Germany, avril 1997.

[Cam98] E. Campione et J. Véronis. A Multilingual Prosodic Database. Dans *International Conference on Spoken Language Processing*, pages 3163–3166. Sydney, Australia, décembre 1998.

[Car99] M. J. Carey, E. J. Parris, et H. Lloyd-Thomas. A Comparison of Features for Speech, Music Discrimination. Dans *IEEE International Conference on Audio, Speech and Signal Processing*, pages 149–152. Phoenix, USA, mars 1999.

[Car00a] J. Carrive, F. Pachet, et R. Ronfard. CLAViS - A Temporal Reasoning System for Classification of Audiovisual Sequences. Dans *Content-Based Multimedia Information Access Conference (RIAO)*. College de France, Paris, France, avril 2000.

[Car00b] M. Carré et P. Pierrick. Indexation audio : un état de l'art. *Annales des télécommunications*, 55(9-10) :507–525, 2000.

[Cer93] J. A. Cercadillo et A. H. Gomez. Grammar Learning and Word Spotting using Recurrent Neural Networks. Dans *European Conference on Speech Communication and Technology*, pages 21–24. Berlin, Germany, septembre 1993.

[Cha96] Y.-L. Chang, W. Zeng, I. Kamel, et R. Alonso. Integrated Image and Speech Analysis for Content-Based Video Indexing. Dans *International Conference on Multimedia Computing and Systems*, pages 306–313. Hiroshima, Japan, juin 1996.

[Cha02] O. Chapelle, V. Vapnik, O. Bousquet, et S. Mukherjee. Choosing Multiple Parameters for Support Vector Machines. *Machine Learning*, 46(1-3) :131–159, 2002.

[Chi92] B. Chigier. Rejection and Keyword Spotting Algorithms for a Directory Assistance City Name Recognition Application. Dans *IEEE International Conference on Audio, Speech and Signal Processing*, tome II, pages 93–96. San Francisco, USA, avril 1992.

[Cor95] C. Cortes et V. Vapnik. Support-Vector Networks. *Machine Learning*, 20(3) :273–297, 1995.

[Cox96] S. Cox et R. C. Rose. Confidence Measures for the Switchboard Database. Dans *IEEE International Conference on Audio, Speech and Signal Processing*, pages 511–514. Atlanta, USA, mai 1996.

[Cri00] N. Cristianini et J. Shawe-Taylor. *An Introduction to Support Vector Machines and other Kernel-Base Learning Methods*. Cambridge University Press, 2000.

[Dav52] K.-H. Davis, R. Biddulph, et S. Balashek. Automatic Recognition of Spoken Digits. *Journal of the Acoustical Society of America*, 24 :637–642, 1952.

[DC98] M. De Calmès et G. Pérennou. BDLEX : a Lexicon for Spoken and Written French. Dans *International Conference on Language Resources and Evaluation*, pages 1129–1136. Granada, Spain, 1998.

[Del96] B. Delyon, A. Morin, et S. Dufour. Reconnaissance de parole. Rapport technique, INRIA, 1996.

[Dem77] A. P. Dempster, N. M. Laird, et D. B. Rubin. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society*, 39 (Series B) :1–38, 1977.

[Dem00] C.-H. Demarty. *Segmentation et Structuration d'un Document Vidéo pour la Caractérisation et l'Indexation de son Contenu Sémantique*. Thèse de doctorat, Ecole des Mines de Paris, janvier 2000.

[Did76] E. Diday et J. C. Simon. *Digital Pattern Recognition*, chapitre Clustering analysis, pages 47–94. Springer-Verlag, New-York, USA, 1976.

[Dub94] D. Dubois et H. Prade. La fusion d'informations imprécises. *Traitement du signal*, 11(6), 1994.

[Dud01] R. O. Duda, P. E. Hart, et D. G. Stork. *Pattern Classification (2nd Edition)*. Wiley-Interscience Publication, 2001.

[EM98] R. El Meliani et D. O'Shaughnessy. Specific Language Modelling for New-Word Detection in Continuous-Speech Recognition. Dans *IEEE International Conference on Audio, Speech and Signal Processing*, pages 321–324. Seattle, USA, mai 1998.

[Fan60] G. Fant. *The Acoustic Theory of Speech Production*. The Hague : Mouton, 1960.

[Fon00] L. Fontaine, C. Sénac, N. Vallès-Parlangeau, et R. André-Obrecht. Indexation de la bande sonore : les composantes parole/musique. Dans *Journées d'Étude sur la Parole*, pages 65–68. Aussois, France, juin 2000.

[Foo97] J. Foote. A Similarity Measure for Automatic Audio Classification. Dans *AAAI, Spring Symposium on Intelligent and Use of Text, Image, Video, and Audio Corpora*, pages 138–147. Stanford University, mars 1997.

[Foo00] J. Foote. Automatic Audio Segmentation using a Measure of Audio Novelty. Dans *IEEE International Conference on Multimedia and Expo*, pages 452–455. New-York, USA, 2000.

[Fre04] C. Fredouille, D. Matrouf, G. Linares, et P. Nocera. Segmentation en macro-classes acoustiques d'émissions radiophoniques dans le cadre d'ESTER. Dans B. Bel et I. Marlien, rédacteurs, *Journées d'Etude sur la Parole*, tome 1, pages 225–228. AFCP, Fès, Maroc, avril 2004.

[Gau94] J. L. Gauvain et C. H. Lee. Maximum A Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains. *IEEE Transactions on Speech and Audio Processing*, 2(2) :291–298, avril 1994.

[Gau99] J. L. Gauvain, L. Lamel, et G. Adda. Systèmes de processus légers : concepts et exemples. Dans *International Workshop on Content-Based Multimedia Indexing*, pages 67–73. GDR-PRC ISIS, Toulouse, France, octobre 1999.

[Gra04] G. Gravier, J. F. Bonastre, E. Geoffrois, S. Galliano, K. Mc Tait, et K. Choukri. ESTER, une campagne d'évaluation des systèmes d'indexation automatique d'émissions radiophoniques en français. Dans B. Bel et I. Marlien, rédacteurs, *Journées d'Etude sur la Parole*, tome 1, pages 253–256. AFCP, Fès, Maroc, avril 2004.

[Gun97] S. R. Gunn, M. Brown, et K. M. Bossley. Network Performance Assessment for Neurofuzzy Data Modelling. Dans X. Liu, P. Cohen, et M. Berthold, rédacteurs, *Intelligent Data Analysis*, tome 1208 de *Lecture Notes in Computer Science*, pages 313–323. London, UK, août 1997.

[Hai04] S. Haidar, P. Joly, et B. Chebaro. Detection Algorithm of Audiovisual Production Invariant. Dans *Workshop on Adaptative Multimedia Retrieval (AMR)*, pages 156–169. Valencia, Spain, août 2004.

[Her85] H. Hermansky, B. A. Hanson, et H. Wakita. Perceptually Based Linear Predictive Analysis of Speech. Dans *IEEE International Conference on Audio, Speech and Signal Processing*, tome 1, pages 85–88. Tampa, USA, mars 1985.

[Her90] H. Hermansky. Perceptual Linear Predictive (PLP) Analysis of Speech. *Journal of the Acoustical Society of America*, 87(4) :1738–1752, avril 1990.

[Hig85] A. L. Higgins et R. E. Wohlford. Keyword Recognition using Template Concatenation. Dans *IEEE International Conference on Audio, Speech and Signal Processing*, pages 1233–1236. Tampa, USA, mars 1985.

[Hou85] T. Houtgast et 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*, 77(3) :1069–1077, 1985.

[Jaf04] G. Jaffré et P. Joly. Costume : A New Feature for Automatic Video Content Indexing. Dans *Coupling approaches, coupling media and coupling languages for information retrieval (RIAO)*, pages 314–325. Avignon, France, avril 2004.

[Jan96] F. Janez. *Fusion d'informations définies sur des référentiels non-exhaustifs différents*. Thèse de doctorat, Université d'Angers, 1996.

[Jel76] F. Jelinek. Continuous Speech Recognition by Statistical Methods. *Proceedings of the IEEE*, 64(4) :532–556, avril 1976.

[Joh70] N. L. Johnson et S. Kotz. *Continuous Univariate Distributions*. Wiley-Interscience Publication, New-York, USA, 1970.

[Kij03] E. Kijak, G. Gravier, L. Oisel, et P. Gros. Audiovisual Integration for Tennis Broadcast Structuring. Dans *International Workshop on Content-Based Multimedia Indexing*, pages 421–428. GDR-PRC ISIS, Rennes, France, septembre 2003.

[Kim96] D. Kimber et L. Nilcox. Acoustic Segmentation for Audio Browsers. Dans *Interface Conference*. Sydney, Australia, juillet 1996.

[Kim97] H.-K. Kim. *Détection automatique des mouvements de caméra et des régions de textes pour la structuration et l'indexation de documents audiovisuels*. Thèse de doctorat, IRIT, Université Paul Sabatier, Toulouse III, mars 1997.

[Koh88] T. Kohonen, G. Barna, et C. Ronald. Statistical Pattern Recognition with Neural Networks : Benchmarking Studies. Dans *Conference on Neural Networks*, pages 61–68. San Diego, USA, juillet 1988.

[Kor99] J. Koreman, B. Andreeva, et H. Strik. Acoustic parameters versus Phonetic Features in ASR. Dans *International Congress of Phonetic Sciences*, pages 549–553. San Francisco, USA, août 1999.

[Lam97] L. Lamel et J. L. Gauvain. Speaker Recognition with the Switchboard Corpus. Dans *IEEE International Conference on Audio, Speech and Signal Processing*, tome II, pages 1067–1070. Munich, Germany, avril 1997.

[Lam00] L. Lamel et J. L. Gauvain. Speaker Verification over the Telephone. *Speech Communication*, 31(2-3) :141–154, 2000.

[Lam04] R. Lamy, D. Moraru, B. Bigi, et L. Besacier. Premiers pas du CLIPS sur les données d'évaluation ESTER. Dans B. Bel et I. Marlien, rédacteurs, *Journées d'Etude sur la Parole*, tome 1, pages 301–304. AFCP, Fès, Maroc, avril 2004.

[Ler00] P. Leray, H. Zaragoza, et F. d'Alché Buc. Pertinence des mesures de confiance en classification. Dans *Congrès de Reconnaissance des Formes et Intelligence Artificielle*, pages 267–276. Paris, France, février 2000.

[Lev47] N. Levinson. The Wiener RMS (Root Mean Square) Error Criterion in Filter Design and Prediction. *Journal of Mathematical Physics*, 25(4) :261–278, janvier 1947.

[Li96] W. Li, S. Gauch, J. Gauch, et K. M. Pus. VISION : a Digital Video Library. Dans *International Conference on Digital Libraries*, pages 19–27. Bethesda, USA, mars 1996.

[Lie99] R. Lienhart, S. Pfeiffer, et W. Effelsberg. Scene Determination Based on Video and Audio Features. Dans *Inernational Conference on Multimedia Computing and Systems*, pages 685–690. Florence, Italy, juin 1999.

[Lin80] Y. Linde, A. Buzo, et R. Gray. An Algorithm for Vector Quantizer Design. *IEEE Transactions on Communication*, 28(1) :84–95, janvier 1980.

[Llo57] S. P. Lloyd. Least Squares Quantization in PCM, 1957. Unpublished Bell Labs Technical Note (1957).

[Llo82] S. P. Lloyd. Least Squares Quantization in PCM. *IEEE Transactions Information Theory (Special Issue on Quantization)*, 28(2) :129–137, mars 1982.

[Lu95] Y. Lu. Machine Printed Character Segmentation - An Overview. *Pattern Recognition*, 28(1) :67–80, 1995.

[Lu01] L. Lu, H. Jiang, et H. Zhang. A Robust Audio Classification and Segmentation Method. Dans *ACM International Conference on Multimedia*, pages 203–211. Ottawa, Canada, septembre 2001.

[Mar02] J. Mariani, rédacteur. *Analyse, Synthèse et Codage de la Parole, Traitement du langage Parlé 1*. Hermes, 2002.

[Mat01] D. Matrouf et J. L. Gauvain. Utilisation des modèles de markov cachés pour le débruitage. *Traitement du signal*, 18(3) :213–218, septembre 2001.

[Mau03] J. Mauclair. *Fusion de paramètres pour une classification Parole/Musique/Bruit*. Rapport de DEA, IRIT, Université Paul Sabatier, Toulouse III, juin 2003.

[Mei01] S. Meignier, J. F. Bonastre, et S. Igounet. E-HMM Approach for Learning and Adapting Sound Models for Speaker Indexing. Dans *A Speaker Odyssey, The Speaker Recognition Workshop*, pages 175–180. Chinia, Crete, juin 2001.

[Mod89] R. Moddemeijer. On Estimation of Entropy and Mutual Information of Continuous Distributions. *Signal Processing*, 16(3) :233–246, 1989.

[Mok95] C. Mokbel, D. Jouvet, et J. Monné. Blind Equalization using Adaptative Filtering for improving Speech Recognition over Telephone. Dans *European Conference on Speech Communication and Technology*, pages 817–820. Madrid, Spain, septembre 1995.

[Mor91] D. P. Morgan, C. L. Scofield, et J. E. Adcock. Multiple Neural Network Topologies Applied To Keyword Spotting. Dans *IEEE International Conference on Audio, Speech and Signal Processing*, pages 313–316. Toronto, Canada, avril 1991.

[Mor00] N. Moreau, D. Charlet, et D. Jouvet. Confidence Measure and Incremental Adaptation for the Rejection of Incorrect Data. Dans *IEEE International Conference on Audio, Speech and Signal Processing*, tome 3, pages 1807–1810. Istanbul, Turkey, janvier 2000.

[MT03] K. Mc Tait et M. Adda-Decker. The 300k LIMSI German Broadcast News Transcription System. Dans *European Conference on Speech Communication and Technology*, pages 213–216. Geneva, Switzerland, septembre 2003.

[NIS00] NIST. *Speech Transcription Workshop*, mai 2000.

[Pin02] J. Pinquier, C. Sénac, et R. André-Obrecht. Indexation de la bande sonore : recherche des composantes parole et musique. Dans *Congrès de Reconnaissance des Formes et Intelligence Artificielle*, pages 163–170. Angers, France, janvier 2002.

[Pin03a] J. Pinquier, J.-L. Rouas, et R. André-Obrecht. A Fusion Study in Speech / Music Classification. Dans *IEEE International Conference on Audio, Speech and Signal Processing*. Hong-Kong, China, avril 2003.

[Pin03b] J. Pinquier, J.-L. Rouas, et R. André-Obrecht. Fusion de paramètres pour une classification automatique parole/musique robuste. *Technique et Science Informatiques (TSI)*, 22(7-8) :831–852, 2003.

[Pin04a] J. Pinquier et R. André-Obrecht. Audio Indexing : Primary Components Retrieval - Robust Classification in Audio Documents. *Multimedia Tools and Applications (MTAP)*, page à paraître, 2004.

[Pin04b] J. Pinquier, J. Arias, et R. André-Obrecht. Audio Classification by Search of Primary Components. Dans *The International Workshop on Multidisciplinary Image, Video, and Audio retrieval and Mining,*, pages CDRom, paper 11. Sherbrooke, Canada, octobre 2004.

[Rab89] L. R. Rabiner, J. G. Wilpon, et F. K. Soong. High Performance Connected Digit Recognition using Hidden Markov Models. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 37(8) :1214–1225, 1989.

[Raz04] J. Razik, D. Fohr, O. Mella, et N. Parlangeau-Vallès. Segmentation parole/musique pour la transcription automatique. Dans B. Bel et I. Marlien, rédacteurs, *Journées d'Etude sur la Parole*, tome 1, pages 417–420. AFCP, Fès, Maroc, avril 2004.

[Rey00] D. A. Reynolds, T. F. Quatieri, et R. B. Dunn. Speaker Verification using Adapted Gaussian Mixture Models. *Digital Signal Processing*, 10(1-3) :19–41, 2000.

[Ros62] F. Rosenblatt. *Principles of Neurodynamics*. Academic Press, 1962.

[Ros90] R. C. Rose et D. B. Paul. A Hidden Markov Model Based Keyword Recognition System. Dans *IEEE International Conference on Audio, Speech and Signal Processing*, pages 129–132. Albuquerque, USA, avril 1990.

[Ros91] R. C. Rose, E. I. Chang, et R. P. Lippmann. Techniques for Information Retrieval from Voice Messages. Dans *IEEE International Conference on Audio, Speech and Signal Processing*, tome 1, pages 317–320. Toronto, Canada, avril 1991.

[Ros00] S. Rossignol. *Segmentation et Indexation des signaux sonores musicaux*. Thèse de doctorat, Université Paris IV, IRCAM, juillet 2000.

[Rui02] R. Ruiloba et P. Joly. Addind the Concept of Video Editing Levels in Shot Segmentation. Dans *Information Processing and Management of Uncertainty in Knowledge-based Systems*, pages 501–506. Annecy, France, juillet 2002.

[Sau96] J. Saunders. Real-time Discrimination of Broadcast Speech/Music. Dans *IEEE International Conference on Audio, Speech and Signal Processing*, pages 993–996. Atlanta, USA, mai 1996.

[Sch97] E. Scheirer et M. Slaney. Construction and Evaluation of a Robust Multifeature Speech/Music Discriminator. Dans *IEEE International Conference on Audio, Speech and Signal Processing*, pages 1331–1334. Munich, Germany, avril 1997.

[Sof98] A. Soffer et H. Samet. Using Negative Shape Features for Logo Similarity Matching. Dans *International Conference on Pattern Recognition*, tome 1, pages 571–574. Brisbane, Australia, août 1998.

[Sua94] N. Suaudeau. *Un modéle probabiliste pour intgrer la dimension temporelle dans un systéme de reconnaisance automatique de parole*. Thèse de doctorat, IRISA, 1994.

[Suk96] R. A. Sukkar et C. Lee. Vocabulary Independent Discriminative Utterance Verification for Nonkeyword Rejection in Subword Based Speech Recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 4(6) :420–429, 1996.

[The99] S. Theodoridis et K. Koutroumbas. *Pattern Recognition*. Academic Press, San Diego, USA, 1999.

[Vap99] V. Vapnik. *The Nature of Statistical Learning Theory*. Springer, 1999.

[Vit67] A. J. Viterbi. Error Bounds for Convolutional Codes and an Asymptotically Optimum Decoding Algorithm. *IEEE Transactions on Information Theory*, 13(2) :260–269, avril 1967.

[Wil90] J. Wilpon, L. Rabiner, C. Lee, et E. Goldman. Automatic Recognition of Keywords in Unconstrained Speech using Hidden Markov Models. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 38(11) :1870–1878, novembre 1990.

[Wol99] E. Wold, T. Blum, D. Keislar, et J. Wheater. Classification, Search and Retrieval of Audio. Dans *CRC Handbook of multimedia computing*. CRC Press LLC, 1999.

[You94] S. Young. The HTK Hidden Markov Model Toolkit : Design and Philosophy. Rapport technique 152, Cambridge University Engineering Department, UK, 1994.

[Zha97] H. J. Zhang, C. Y. Low, S. W. Smoliar, et J. H. Wu. Video Parsing, Retrieval and Browsing - An Integrated and Content-Based Solution. *In Intelligent multimedia information retrieval*, pages 139–158, 1997.

[Zha98] T. Zhang, C. Kuo, et C. J. Hierarchical System for Content-Based Audio Classification and Retrieval. Dans *Conference on Multimedia Storage and Archiving Systems III*, tome 3527 de *SPIE*, pages 398–409. novembre 1998.

[Zha01] Y. Zhang et A. I. Rudnicky. Word Level Confidence Annotation using Combinations of Features. Dans *European Conference on Speech Communication and Technology*, pages 2105–2108. Aalborg, Denmark, septembre 2001.

[Zha03] W. Zhao, R. Chellappa, P. J. Phillips, et A. Rosenfeld. Face Recognition : A Literature Survey. *ACM Computing Surveys (CSUR)*, 35(4) :399–458, décembre 2003.

Résumé

Le développement croissant des données numériques et l'explosion des accès multimédia à l'information, sont confrontés au manque d'outils automatiques efficaces. Dans ce cadre, plusieurs approches relatives à l'indexation et la structuration de la bande sonore de documents audiovisuels sont proposées. Leurs buts sont de détecter les composantes primaires telles que la parole, la musique et les sons clés (jingles, sons caractéristiques, mots clés...).

Pour la classification parole/musique, trois paramètres inhabituels sont extraits : la modulation de l'entropie, la durée des segments (issue d'une segmentation automatique) et le nombre de ces segments par seconde. Les informations issues de ces trois paramètres sont ensuite fusionnées avec celle issue de la modulation de l'énergie à quatre hertz. Des expériences sur un corpus radiophonique montrent la robustesse de ces paramètres : notre système possède un taux de classification correcte supérieur à 90 %. Le système est ensuite comparé, puis fusionné à un système classique fondé sur des Modèles de Mélanges de lois Gaussiennes (MMG) et une analyse cepstrale.

Un autre partitionnement consiste à détecter des sons clés. La sélection de candidats potentiels est effectuée en comparant la « signature » de chacun des jingles au flux de données. Ce système est simple par sa mise en œuvre mais rapide et très efficace : sur un corpus audiovisuel d'une dizaine d'heures (environ 200 jingles) aucune fausse alarme n'est présente. Il y a seulement deux omissions dans des conditions extrêmes.

Les sons caractéristiques (applaudissements et rires) sont modélisés à l'aide de MMG dans le domaine spectral. Un corpus télévisuel permet de valider cette première étude par des résultats encourageants.

La détection de mots clés est effectuée de manière classique : il ne s'agit pas ici d'améliorer les systèmes existants mais de se placer toujours dans un besoin de structuration. Ainsi, ces mots clés renseignent sur le type des émissions (journal, météo, documentaire...).

Grâce à l'extraction de ces composantes primaires, les émissions audiovisuelles peuvent être annotées de manière automatique. Une réflexion est conduite quant à l'utilisation de ces composantes afin de trouver une structure temporelle à nos documents. La première étude permet une détection d'un motif récurrent dans une collection d'émissions, dites de plateau, alors que la seconde réalise la structuration en thèmes d'un journal télévisé. Quelques pistes de réflexions sur l'apport de l'analyse vidéo sont développées et les besoins futurs sont explorés.

Mots-clés: indexation sonore, structuration audiovisuelle, classification, énergie, entropie, segmentation, parole, musique, jingles, sons clés, applaudissements, rires, mots clés, thèmes.