

République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieure et de la Recherche Scientifique
Université des Sciences et de la Technologie Houari Boumedienne
Faculté d'Electronique et d'Informatique



THESE

Présentée pour l'obtention du Diplôme de DOCTORAT D'ETAT

EN : ELECTRONIQUE

Spécialité : Electronique des Systèmes

Par : Lamya FERGANI née FERHAT-HAMIDA

Thème :

Indexation et Classification de Signaux Audio

Soutenue publiquement le Décembre 2010, devant le Jury composé de :

Mme Rachida TOUHAMI	Professeur	USTHB Présidente
M. Amrane HOUACINE	Professeur	USTHB Directeur de thèse
Mme Mhania GUERTI	Maître de Conférence/A	ENP Examinatrice
M. Bachir BOUDRAA	Maître de Conférences/A	USTHB Examineur
M. Riadh BABA ALI	Maître de Conférences/A	USTHB Examineur
M. Halim SAYOUD	Maitre de Conférences A	USTHB Examineur

Résumé

Les signaux audio sont omniprésents dans notre vie quotidienne et constituent notre environnement naturel. Ces signaux nous permettent de scruter l'environnement, de communiquer et d'apprendre. Le flux audio est généralement l'enregistrement d'un signal acoustique obtenu à partir de plusieurs sources sonores. Le signal ainsi obtenu est constitué de plusieurs composantes dont les plus fréquentes sont la parole, la musique et le bruit. Cette dernière composante regroupe en fait toutes les composantes non identifiables à de la parole ou de la musique. Cet environnement est caractérisé par une forte variabilité acoustique, on peut ainsi rencontrer de larges segments de parole en bande élargie ou en bande téléphonique, de très brefs segments de jingles et des segments de quelques minutes de musique instrumentale ou de la voix chantée.

Le traitement des signaux musicaux représente un axe de recherche et de développement technologique à fort potentiel mobilisateur en temps-homme et en ressources matérielles et justifie l'intérêt grandissant de l'industrie des produits multimédia. Les genres musicaux sont des catégories créées par les humains pour caractériser des morceaux de musique. Ces caractéristiques sont typiquement liées à l'instrumentation, la structure rythmique, et à la teneur harmonique de la musique. Actuellement l'annotation du genre musical se fait manuellement. La classification automatique du genre musical peut aider ou remplacer l'opérateur humain dans ce processus et peut rendre ainsi la recherche de documents musicaux plus efficace. Dans ce cadre, notre thèse s'est focalisée sur l'analyse et la classification en genres de signaux musicaux.

Nous nous sommes attachés, en premier lieu, à établir une synthèse bibliographique détaillée des méthodes d'analyse des signaux audio, et plus spécifiquement musicaux. Ces signaux sont alors représentés dans un nouvel espace de dimension réduite : l'espace des descripteurs audio. Le premier maillon de la chaîne d'analyse de signaux musicaux est la segmentation du flux audio en classes sonores. Après une brève revue des méthodes de segmentation existantes, nous proposerons une méthode originale de segmentation en classes sonores. Le processus de segmentation est ensuite suivi d'une étape de séparation de sources sonores qui permettra, dans la classe sonore musique, de séparer la parole d'un bruit ambiant et la musique instrumentale de la voix chantée : deux méthodes seront proposées puis comparées à cet effet l'Analyse en Composantes Indépendantes (A.C.I.) et la Décomposition Empirique Modale (E.M.D.). Le cœur de notre travail est constitué d'une étude comparative de différentes méthodes de classification en genre de signaux musicaux. Plusieurs classifieurs tels que les SVM, les réseaux de neurones et les GMM seront utilisés. Différentes combinaisons de descripteurs seront également testées. Nous avons recherché la meilleure combinaison de descripteurs en vue d'optimiser la classification en genres de signaux musicaux.

La validation de nos méthodes repose sur des enregistrements de signaux musicaux organisés en bases de données thématiques et régionales. Notre contribution porte sur les deux aspects théoriques (proposition d'un nouvel algorithme de segmentation audio) et applicatifs (base de données de signaux audio algériens).

Sommaire

Chapitre I : Introduction Générale.....	10
Références bibliographiques.....	14
Chapitre II : Descripteurs de Signaux Audio et Musicaux.....	15
Introduction.....	16
Descripteurs Temporels.....	17
Descripteurs Energétiques.....	20
Descripteurs Harmoniques.....	23
Descripteurs Perceptuels.....	24
Descripteurs Ondelettes.....	29
Choix de filtres Ondelettes pour les signaux musicaux.....	32
Combinaison de descripteurs et réduction de dimensionnalité.....	39
Conclusions.....	39
Références bibliographiques	40
Chapitre III : Segmentation de Signaux Audio.....	42
Introduction.....	43
Etat de l'art.....	43
Une méthode hybride : ACP + Critère Alignement.....	44
L'algorithme de segmentation ACP_Alignement.....	46
Conclusions	49
Références Bibliographiques.....	50
Chapitre IV : Séparation de Sources Sonores.....	52
Introduction.....	53
Analyse en Composantes Indépendantes : Formulation générale.....	54
L'Algorithme FASTICA.....	57
La Décomposition Empirique Modale.....	58
Séparation de sources par DEM.....	61
Critères d'évaluation de la séparation de sources audio.....	63
Conclusions	65
Références bibliographiques.....	66

Chapitre V :Classification en genres de signaux musicaux.....	68
Introduction.....	69
Etat de l’art de la classification en genres musicaux.....	70
Formalisation probabiliste de la classification.....	71
Formalisation de la classification en termes de RdF.....	72
Exemples d’approches génératives.....	73
Exemples d’approches discriminatives.....	74
Classification SVM Multi-Labels.....	84
Conclusions.....	87
Références bibliographiques.....	89
Chapitre IV : Résultats et évaluations de performances.....	92
Introduction.....	93
Sur la segmentation de signaux audio.....	93
Sur la séparation de sources audio.....	100
Sur la classification en genres musicaux.....	105
Description de la base de données « Zohier ».....	105
Réseaux de Neurones, approche globale et descripteurs TSPMFCC.....	108
Effet du partitionnement de la base de données sur le taux de classif.....	114
Classification en genres et évaluations.....	115
Machines à Vecteurs de Supports, approche instantanée et descript.....	117
SVM et noyau RBF.....	118
SVM, noyau RBF et fusion de classes.....	121
SVM, approche globale et descripteurs TSPMFCC.....	122
SVM et descripteurs ondelettes.....	127
SVM Multilabels et descripteurs ondelettes.....	131
Comparaison de descripteurs et Conclusions.....	132
Conclusions.....	133
Chapitre IIV : Conclusion Générale et Perspectives.....	136
Bibliographie Globale.....	139

LISTE DES TABLEAUX

Tab.VI.1 : Description de la base de données

Tab.VI.2 : Données avec énergies non normalisés

Tab.VI.3 : Données avec énergies normalisées

Tab.VI.4 : Influence de la dimension des descripteurs :F-mesure (classe Bruit)

Tab.VI.5: Influence de la dimension des descripteurs :F-mesure (classe Silence)

Tab.VI.6 : Influence de la dimension des descripteurs :F-mesure (classe Musique)

Tab.VI.7 : Influence de la dimension des descripteurs :F-mesure (classe Parole)

Tab.VI.8 : Influence de la durée des fichiers de référence :F mesure (classe Bruit)

Tab.VI.9 : Influence de la durée des fichiers de référence :F mesure (classe Silence)

Tab.VI.10 : Influence de la durée des fichiers de référence :F mesure (classe Musique)

Tab.VI.11: Influence de la durée des fichiers de référence :F mesure (classe Parole)

Tab.VI.12 : Influence de la taille de la fenêtre de lissage :F mesure (classe Bruit)

Tab.VI.13 : Influence de la taille de la fenêtre de lissage :F mesure (classe Silence)

Tab.VI.14 : Influence de la taille de la fenêtre de lissage :F mesure (classe Musique)

Tab.VI.15 : Influence de la taille de la fenêtre de lissage :F mesure (classe Parole)

Tab.VI.16 : Matrice de confusion pour le fichier *Test1.wav*

Tab.VI.17 : Matrice de confusion pour le fichier *Test2.wav*

Tab.VI.18 : Critères ISR et DSR pour les deux méthodes FASTICA et EMD

Tab.VI.19: Récapitulatif de la base de données

Tableau VI.20: Moyenne et Variance des paramètres temporels et spectraux

Tableau VI.21 : Moyenne des 9 premiers coefficients MFCC

Tableau VI.22 : Variance des 9 premiers coefficients MFCC

Tableau VI.23: Moyenne des 9 premiers coefficients LPC

Tableau VI.24 : Variance des 9 premiers coefficients LPC

Tableau VI.25 : Taux de classification en fonction du nombre de neurones de la couche cachée-VTS5MFCC

Tableau VI.26 : Taux de classification en fonction du nombre de neurones de la couche cachée - VTS5LPC

Tableau VI.27: Taux de classification en fonction du nombre d'itérations- VTS5MFCC.

Tableau VI.28 : Taux moyens de classification en fonction du nombre d'itérations- VTS5LPC.

Tableau VI.29: Taux de classification en fonction du pas d'apprentissage- VTS5MFCC

Tableau VI.30 : Taux moyens de classification en fonction du pas d'apprentissage-VTS5LPC.

Tableau VI.31: Effet du partitionnement sur le taux de classification - VTS5MFCC et VTS5LPC.

Tableau VI.32.a: Matrice de confusion

Tableau VI.32.b : Taux moyens - Différents vecteurs acoustiques

Tableau VI.33 : Evolution du taux de classification en fonction de $(C, \hat{u})$, sans normalisation.

Tableau VI.34 : Matrice de confusion (phase de test), sans normalisation.

Tableau VI.35: Evolution du taux de classification en fonction de $(C, \hat{u})$, avec normalisation.

Tableau VI.36 : Matrice de confusion (Phase de test), avec normalisation

Tableau VI.37 : Taux de classification en fonction de D

Tableau VI.38 : Matrice de confusion (Phase de test), Noyau Polynomial

Tableau VI.39 : Evolution du taux de classification en fonction de $(C, \hat{u})$.

Tableau VI.40 : Matrice de confusion (Phase test)

Tableau VI.41 : Evolution du taux de classification en fonction de $(C, \hat{u})$, 26 descripteurs.

Tableau VI.42: Matrice de confusion (Phase test).

Tableau VI.43 : Evolution du taux de classification en fonction de $(C, \hat{u})$, descripteurs TSPMFCC-IND

Tableau VI.44: Matrice de confusion (Phase test).

Tableau VI.45 : Taux de classification en fonction de $(C, \hat{u})$, MCPEA, Ondelette Meyer

Tableau VI.46 : Matrice de confusion (Phase test)

Tableau VI.47 Evolution du taux de classification en fonction de $(C, \hat{u})$, cas Multilabels

Tableau VI.48 : Matrice de confusion (Phase test)

LISTE DES FIGURES

Figure II.1.a : Son voisé (TPZ faible)

Figure II.1.b : Son non-voisé (TPZ élevé)

Figure II.2 : Durée effective

Figure II.3 : Modélisation d'une enveloppe ADSR

Fig.II.4.a : Skewness=0

Fig.II.4.b : Skewness<0

Fig.II.4.c : Skewness>0

Fig.II.5.a: Kurtosis=3

Fig.II.5.b: Kurtosis=1.8

Fig.II.5.c : Kurtosis=6

Figure II.6 : Energie du spectre en fonction de la fréquence (Roulement Spectral en rouge).

Figure II.7 : Spectre du signal brut et Spectre de la partie harmonique.

Figure II.8 Représentation des pics d'un signal (en lignes continues) et de ses harmoniques (en pointillés).

Figure II.9 : Sonie spécifique

Figure II.10 : Etapes de calcul des coefficients MFCC.

Figure II.11. : Répartition des filtres triangulaires sur une échelle linéaire (Hz) et sur une échelle Mel.

Figure II.12. : Dilatation d'une ondelette.

Figure II.13 : Translation d'une ondelette.

Figure II.14 : Analyse-Synthèse d'un signal (Algorithme de Mallat).

Figure II.15 : Ondelette de Haar

Figure II.16 : Ondelette complexe de Morlet

Figure II.17 : Ondelette chapeau mexicain

Figure II.18 : Ondelette de Meyer et sa transformée de Fourier

Figure II.19 : Ondelettes de Daubechies et réponses en fréquences pour différentes valeurs de p

Figure II.20 : Ondelettes Symlets pour différentes valeurs de p

Figure II.21 : Ondelettes Coiflets pour différentes valeurs de p

Figure II.22 : Ondelettes splines et fonctions d'échelle associées

Fig.III.1: Exemple de données linéairement séparables

Fig.III.2: Le paramètre α en fonction de l'alignement dans un cas séparable

Fig.III.3: Influence du paramètre α sur l'hyperplan séparateur

Fig.III.4: Exemple de données linéairement non séparables

Fig.III.5: Le paramètre α en fonction de l'alignement dans un cas non séparable

Fig.IV.1: Modèle général d'un système de séparation de sources

Fig.IV.2: Organigramme SAS par approche DEM

Fig.V.1: Schéma général de la classification

Fig.V.2: Structure d'un réseau de neurones

Fig.V.3: Exemple de séparation des classes avec différents hyperplans

Fig.V.4: Hyperplan optimal et vecteurs de supports

Fig.V.5: Différentes valeurs de α

Fig.V.6: Classification 1 parmi K

Fig.VI.1 .a et b : Source originale (Parole et Musique)

Fig.VI.2 .a et b : Sources estimés par FASTICA

Fig.VI.3 .a et b : Sources estimés par EMD

Fig.VI.4 .a et b : Evolution des critères DSR et ISR en fonction du RSB (Parole+musique+bruit)

Fig.VI.5 .a et b : Evolution des critères DSR et ISR en fonction du RSB (Parole+bruit)

Figure VI.6: Taux de classification en fonction du pas d'apprentissage-VTS5MFCC.

Figure VI.7: Taux de classification en fonction du pas d'apprentissage - VTS5LPC

Figure VI.8: Taux de classification pour différentes partitions de la base de données

Figure VI.9: Taux de Classification

Figure VI.10 : Taux moyens $\bar{\gamma}$ Différents vecteurs acoustiques

Figure VI.11: Evolution du taux de classification en fonction du nombre de descripteurs.

Figure VI.12 : Influence du type de descripteurs

Figure VI.13 : Influence du type d'ondelette

Figure VI.14 : Influence du type de noyau

Figure VI.15 : Influence de la normalisation

Figure VI.16 : Influence du type de vecteur acoustique

Figure VI.17 : Comparaison entre les différents jeux de descripteurs

Introduction Générale

Avec l'avènement des nouvelles technologies dans le domaine des réseaux et du stockage de données, on assiste à un surplus d'informations. En plus des documents textuels, le son, la vidéo et l'image prennent une part croissante dans les documents mis à la disposition du grand public. Actuellement, avec la présence de milliers de chaînes de télévision et de chaînes de radio, beaucoup d'organismes numérisent et archivent leurs émissions de radio et de télévisions, permettant ainsi aux auditeurs de consulter les émissions ultérieurement à partir d'internet ou à partir de bases de données spécialisées [1,2,3].

L'accès direct à l'information recherchée nécessite l'existence des nouveaux outils capables de décrire la structure et le contenu de ces documents. Étant donné le volume de documents existants, seuls des outils de traitements automatiques ou semi-automatiques (avec une faible intervention humaine) sont envisageables. Ces outils doivent être capables d'extraire et de représenter les informations contenues dans un document multimédia, en plus de simplifier la recherche des informations. De ce fait, le problème de traitement automatique des documents multimédia représente aujourd'hui un axe de recherche privilégié focalisant l'intérêt des chercheurs en traitement de signal [4].

Un document sonore est constitué de nombreuses composantes ou sources, dont les plus communes sont la parole et la musique. Évidemment, ces deux principales sources de sons ne sont que des composantes primaires, auxquelles il convient de rajouter les multiples sources de bruits [5].

L'objectif de cette thèse est de s'intéresser à la composante musicale d'un fichier audio. Les raisons et les arguments objectifs à cela sont multiples :

- La récente explosion de la quantité de musique accessible sur Internet a créé de nouveaux intérêts dans le domaine de la distribution de musique, aussi bien pour les utilisateurs qui recherchent des moyens pratiques de gérer et d'utiliser leurs collections personnelles, que pour les distributeurs, qui cherchent à donner accès au plus grand nombre à la musique qui les intéresse.
- Cette gigantesque quantité d'information ne pouvant être gérée directement, l'utilisation de méta-données, c'est-à-dire de données qui décrivent elles-mêmes les données recherchées, devient incontournable.
- Création d'applications musicales, comme la génération automatique de programmes musicaux ou de nouvelles compositions à partir de ces méta-données.

Le but de la classification en genres de signaux musicaux est de faciliter l'accès à la musique à un public d'utilisateurs non spécialistes. Elle permet en outre la gestion des collections musicales personnelles ou publiques et la recherche de nouveaux morceaux de musique (sur Internet), en fonction des goûts et des aspirations de l'utilisateur, dans des catalogues musicaux numériques de grande taille. Plusieurs laboratoires de Recherche et Développement d'entreprises spécialisées telles que Sony CSL et Nokia développent régulièrement de nouveaux produits dans ce domaine [6,7,8].

Les genres musicaux sont des catégories créées par les humains pour caractériser des morceaux de musique. Ces caractéristiques sont typiquement liées à l'instrumentation, la structure rythmique, et à la teneur harmonique de la musique. Les genres hiérarchiques sont généralement employés pour structurer les grandes collections de musique disponibles sur le WEB. Actuellement l'annotation du genre musical se fait manuellement. La classification automatique du genre musical peut aider ou remplacer l'opérateur humain dans ce processus ; et peut rendre ainsi la recherche documentaliste musical plus efficace. En outre, la classification automatique des genres musicaux fournit un cadre pour le développement et l'évaluation des dispositifs de n'importe quel type de système d'analyse des signaux musicaux.

Les genres musicaux n'ont aucune définition ou critères strictes puisqu'ils proviennent d'une interaction complexe entre le public, l'économie de marché, l'histoire et la culture. Cette observation a poussé quelques chercheurs à suggérer la définition d'un nouvel arrangement de classification de genre musical et ce dans le seul but d'améliorer les systèmes de recherche documentaliste musical. Toutefois même avec les genres musicaux courants, les membres d'un genre particulier partagent certaines caractéristiques typiques liées à l'instrumentation, la structure rythmique et à la teneur en lancement de la musique [9,10,11].

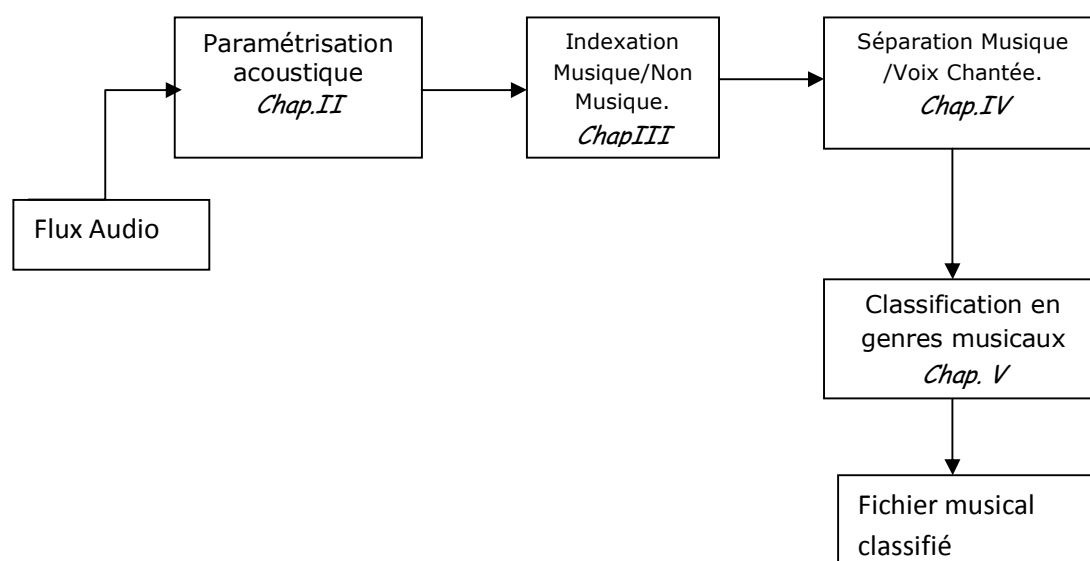
Les musiques d'un même genre présentent généralement une sorte de similarité d'instruments utilisés ou du rythme. Deux morceaux de musique d'un même genre sont généralement considérées comme étant plus similaires que deux morceaux des genres différents. Par nature, les termes « similarité musicale » ou « classification en genre de la musique » contiennent une part importante de la subjectivité. Il est connu que deux chansons similaires pour une personne ne le sont pas nécessairement pour une autre personne. Le problème posé est de classer automatiquement les signaux audio dans une hiérarchie des genres musicaux. Plus spécifiquement, nous proposons trois ensembles de dispositifs pour représenter la texture du timbre, le contenu rythmique et le contenu du lancement [12,13].

Bien qu'il y ait eu un travail significatif dans le développement des dispositifs pour la reconnaissance de la parole et la discrimination entre la musique et le discours, il y a eu relativement peu de travail dans le développement des dispositifs spécifiquement conçus pour le signal musical. Aussi notre travail s'inscrit dans cette problématique et traite de l'ensemble des aspects y afférant.

Après une synthèse bibliographique détaillée de l'état de l'art des méthodes d'analyse et de représentation des signaux audio, à savoir les descripteurs acoustiques caractéristiques des signaux musicaux, les méthodes de segmentation parole/musique ou de musique/Non musique seront traitées en détail dans le chapitre trois. Ces méthodes constituent un premier maillon de la chaîne de traitement des signaux musicaux, et conditionne souvent les performances des étages ultérieurs. Le chapitre quatre est dédié aux méthodes de séparation en classes de sources sonores, comme par exemple séparer la parole d'un bruit ambiant ou la musique instrumentale de la voix chantée. Le chapitre suivant constitue le cœur de notre travail dans lequel nous mettons en œuvre une étude comparative de méthodes de classification en genre de signaux musicaux. La partie expérimentale est le sixième chapitre,

il permet de valider nos méthodes en considérant plusieurs enregistrements de signaux musicaux divers et variés organisés en bases de données thématiques et régionales.

Le schéma fonctionnel ci-après résume les étapes des traitements réalisés sur un fichier audio stream afin d'aboutir à un fichier audio classifié en genre musical.



Références Bibliographiques du Chapitre

- [1] Carré M. et Philipe P., *indexation audio , un état de l'art*, in Annales des Télécommunications, septembre 2000.
- [2] J. T. Foote., *Content-Based Retrieval of Music and Audio*. In *Proc. SPIE*, vol3229, pp. 138-147, 1997.
- [3] J. Foote., *An overview of audio information retrieval*. In *Multimedia Systems* 7, pp 2-10. ACM, January 1999.
- [4] J. Foote., *Visualizing Music and Audio using Self-Similarity*. In *Proc. ACM Multimedia* 99, pp 77-80,1999.
- [5] S. Foster, W. Schloss, and A. J. Rockmore., *Towards an intelligent editor of digital audio: Signal processing methods . Computer Music Journal*, 6(1):42-51, 1982.
- [6] J.S. Downie., *Music Information Retrieval*, volume 37 of Annual Review of Information Science and Technology, chapter 7, pages 295-340. 2003.
- [7] S. Lippens, J. Martens, M. Leman, B. Baets, H. Meyer, and G. Tzanetakis., *A comparison of human and automatic musical genre classification*. In Proceedings of the IEEE International Conference on Audio, Speech and Signal Processing (ICASSP), volume 4, pages 233-236, May 2004.
- [8] G. Tzanetakis., and P. Cook., *Musical Genre Classification of Audio Signals*, IEEE Transactions on Speech and Audio Processing, vol. 10, no. 5, pp. 293-302, 2002.
- [9] J. J. Aucouturier, and F. Pachet., *Representing Musical Genre: A State of the Art* Journal of New Music Research, vol. 32, no. 1, pp. 1-12, 2003.
- [10] C. McKay. *Automatic Genre Classification of MIDI Recordings*, M.A. Thesis, McGill University, Canada, 2004.
- [11] G. Tzanetakis and K. Lemstrom, *Music Information Retrieval in Polyphonic Music* Encyclopedia of Library and Information Sciences. M. Bates, M.N. Maack and M. Drake (eds), Taylor and Francis 2008
- [12] A. P. Klapuri, A. J. Eronen, and J. T. Astola, *Analysis of the meter of acoustic musical signals*, in *IEEE Trans. on Audio, Speech and Lang. Proc.*, vol. 14, no. 1, pp. 342-355, 2006.
- [13] S. Hainsworth, *Techniques for the automated analysis of musical audio*, Ph.D. dissertation, Dept. Eng., Cambridge University, 2004.
- [14] Harb H., *Classification du signal sonore en vue d'une indexation par le Contenu des documents multimédia* Thèse de doctorat de l'école centrale de Lyon . septembre 2004 .

Descripteurs de Signaux Audio et Musicaux

Introduction

Le traitement numérique de signaux audio nécessite l'utilisation de descripteurs compacts. En effet, décrire un fichier sonore consiste à extraire des informations pertinentes et non redondantes nécessaires à la réalisation des traitements sur la chaîne d'analyse. Il s'agit donc d'associer à un fichier sonore de grande taille un nombre limité de descripteurs, qui cependant concentrent le maximum d'informations sur ce dernier.

Les descripteurs musicaux sont extrêmement variés, et on peut les classer suivant divers critères : leur origine, leur niveau d'abstraction et leur subjectivité [20]; Les descripteurs culturels sont des données liées au contexte dans lequel a été créé un son ou un extrait musical. Ils sont en général textuels (nom de l'artiste, titre, etc.), et extractibles par recherche dans des bases de données sur internet. Cependant, ces descripteurs ne donnent aucune information sur les caractéristiques sonores intrinsèques à la musique. Dans ce chapitre, afin de décrire véritablement les propriétés sonores de passages musicaux indépendamment de leur contexte culturel, nous nous intéresserons uniquement aux descripteurs extraits à partir de l'étude de signaux acoustiques.

L'ensemble des descripteurs acoustiques peut être divisé en plusieurs niveaux de description, mesurant l'abstraction des propriétés qu'ils évaluent par rapport à celles du signal acoustique original [6,8,9] :

- Des descripteurs de **bas niveau** expriment des propriétés sonores significatives directement observables ou calculables à partir du signal audio, mais qui n'ont pas forcément de signification évidente au sens musical, comme par exemple les caractéristiques des enveloppes d'énergie ou du spectre du signal ;
- Des descripteurs de **haut niveau** expriment des propriétés qui ont une signification musicale plus évidente et compréhensibles par les utilisateurs, mais nécessitent le calcul de modèles acoustiques plus complexes.

Différentes techniques existent pour extraire de façon automatique une information perceptive à partir de signaux acoustiques. Ainsi, certains descripteurs simples peuvent être extraits directement à partir du signal audio en utilisant des algorithmes spécifiques, tandis que d'autres descripteurs de haut niveau nécessitent non seulement l'extraction de caractéristiques primaires à partir du signal (les descripteurs de bas niveau), mais également la construction d'un modèle permettant de combiner ces paramètres pour obtenir le descripteur désiré. Ce dernier schéma est classique dans l'état de l'art de la description musicale [18,19]: on calcule tout d'abord un grand nombre de descripteurs de bas niveau à partir du signal audio, on choisit ensuite les plus pertinents en utilisant diverses méthodes de sélection de fonctions plus ou moins complexes pour produire un descripteur de plus haut niveau. L'agrégation des valeurs caractéristiques sur des fenêtres d'observation locales successives pour obtenir une valeur globale pour le signal entier, peut se faire suivant deux approches différentes et qui conduisent à deux types de descripteurs :

- Des descripteurs **globaux** qui rendent compte des propriétés générales du document audio. Ce sont des descripteurs calculés sur toute la durée du signal et qui utilisent les statistiques des descripteurs (moyenne, variance, variables statistiques d'ordre supérieur, corrélations, etc.)
- Des descripteurs **locaux (ou instantanés)** qui traduisent des propriétés estimées localement sur le signal. Ils sont souvent extraits à partir d'un processus de fenêtre glissante consistant à sélectionner répétitivement un bloc d'échantillons de signal de taille constante décalé d'un incrément fixe. L'agrégation est faite de manière séquentielle et on obtient un vecteur de dimension le nombre de fenêtres.

Tous les descripteurs sont modélisés dans le but d'être intégrés dans des applications musicales utilisant du contenu musical original. Ils doivent alors satisfaire à diverses contraintes d'exploitation, principalement : leur généralité, leur modularité, leur qualité, et leur efficacité. La généralité d'un descripteur implique qu'il soit applicable à tout extrait musical concerné. La modularité d'un descripteur implique qu'il puisse être intégré dans diverses applications informatiques, sans problème de compatibilité. La qualité d'un descripteur est sa capacité à donner une valeur descriptive correcte pour tout extrait musical. Enfin, l'efficacité d'un descripteur est la capacité à calculer sa valeur en un temps minimum compatible avec l'utilisation des applications musicales.

Nous nous intéressons, dans le cadre de ce chapitre, aux descripteurs acoustiques de bas niveau. Les deux approches globale et instantanée seront exposées et l'agrégation des descripteurs se fera par concaténation de différentes combinaisons de ces derniers. Les descripteurs que nous avons décidé de développer sont rangés en descripteurs **temporels, spectraux, énergétiques, harmoniques et perceptuels** ainsi que les descripteurs **ondelettes**.

II.1. DESCRIPTEURS TEMPORELS [23]

Ils sont calculés à partir de l'enveloppe temporelle du signal et peuvent être locaux ou globaux.

II.1.1 Taux de passage par zéro TPZ (ou Zéro Crossing Rate ZCR)

Le Taux de Passage par Zéro (TPZ) est le nombre de fois où le signal change de signe sur une unité de temps. Ce paramètre a été utilisé avec succès dans la classification de signaux sonores (voisé/non-voisé) mais aussi dans la discrimination parole/musique. Il est défini par la relation suivante :

$$TPZ = \frac{1}{2N} \sum_{i=1}^{N-1} |sign(x_n[i]) - sign(x_n[i+1])| \quad (II.1)$$

Où N est le nombre d'échantillons du signal sur la fenêtre de durée T .

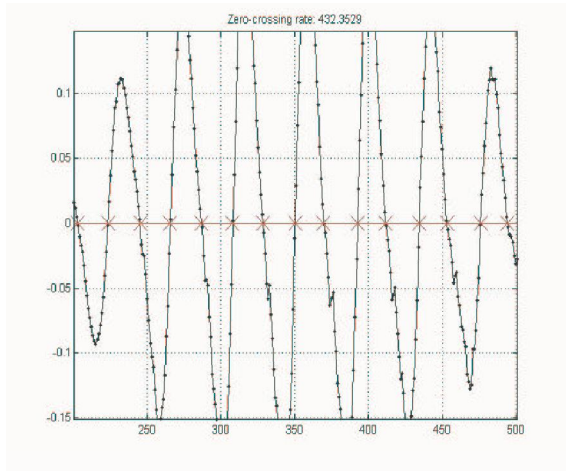


Figure II.1.a Son voisé (TPZ faible)

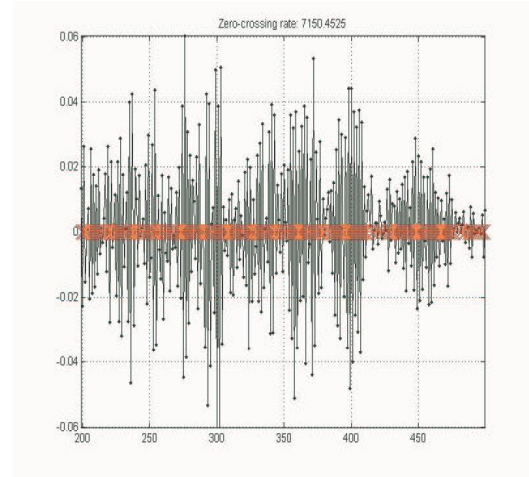


Figure II.1.b Son non-voisé (TPZ élevé)

II.1.2 L'autocorrélation

L'autocorrélation est la convolution du signal avec lui-même. C'est un outil qui permet de détecter des régularités et des profils répétés sur une fenêtre du signal. Elle est couramment utilisée pour déterminer la fréquence fondamentale. Dans le cas de signaux musicaux, l'autocorrélation est utilisée dans différents algorithmes pour détecter le rythme et le pitch [6,8]. Elle est définie par la relation suivante :

$$R_{xx}(k) = \frac{1}{x(0)^2} \sum_{n=0}^{N-k-1} x[n] \cdot x[n+k] \quad (\text{II.2})$$

II.1.3 Centroïde temporel

C'est le barycentre de l'énergie du signal. Il permet de distinguer les sons percutants des sons soutenus et est défini comme suit :

$$TC = \frac{\sum_{n=1}^N E[n] \cdot n}{\sum_{n=1}^N E[n]} \quad (\text{II.3})$$

Où $E[n]$ est l'énergie du $n^{\text{ième}}$ échantillon et N le nombre total d'échantillons.

II.1.4 Durée effective (Effective duration)

C'est la durée pendant laquelle l'énergie dépasse 40% de son amplitude maximale, ce qui nous permet de localiser les sons percutants.

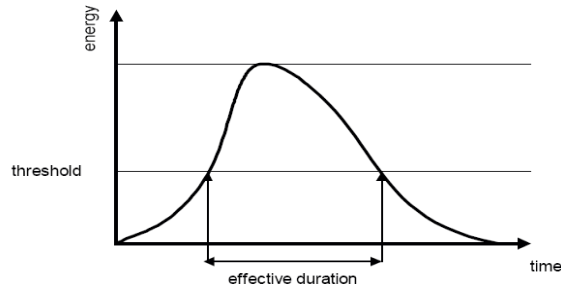


Figure II.2 Durée effective

II.1.5 ADSR (Attack, Decay, Sustain and Release)

L'ADSR est très utilisé pour représenter l'évolution de l'énergie d'un signal musical au cours du temps, en définissant les paramètres temps d'attaque (Attack), de déclin (Decay), de maintien (Sustain) et de relâchement (Release). Ils sont définis comme suit :

- Attack Time (*Attaque*)

C'est la durée que met l'énergie du signal pour passer d'un premier seuil à un autre ; Ce temps est calculé entre 20% et 90% de la valeur maximale de l'énergie.

- Decay (*Déclin*)

Le *déclin* indique le temps, au delà du pic, pour atteindre l'instant, où le son entre en phase de maintien.

- Sustain (*maintien*) :

Le *maintien* est le temps pour lequel le niveau sonore est conservé tant que l'excitation est présente. Lorsque celle-ci disparaît, le son entre directement en *relâchement*.

- Release (*relâchement*):

Le relâchement indique la rapidité avec laquelle le son diminue pour s'éteindre complètement.

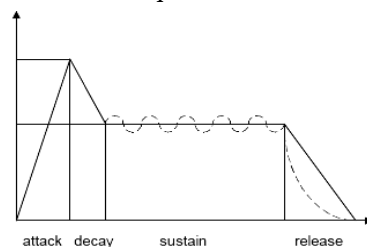


Figure II.3 Modélisation d'une enveloppe ADSR

Le calcul de l'ADSR nécessite le calcul des quatre paramètres suscités (ce qui peut se révéler parfois très ardu).

II.2. DESCRIPTEURS ENERGETIQUES

Ce sont des descripteurs dont le calcul inclut l'énergie du signal.

II.2.1 Energie Moyenne Totale

C'est l'énergie moyenne totale sur une fenêtre de durée N du signal.

$$E = \frac{1}{N} \sum_{n=1}^{N-1} x[n]^2 \quad (\text{II.4})$$

II.2.2 Taux de trames de basses énergies (Low energy rate)

Dans des applications de discrimination parole/musique, le taux de trames de basses énergies est utilisé car les distributions de probabilités de ce descripteur sont bien différentes : loi normale centrée sur 0.3 pour la parole et loi normale centrée sur une valeur négative (en fait, seules les valeurs positives nous intéressent, ce qui implique une courbe décroissante pour la partie positive) pour la musique. C'est le pourcentage de fenêtres qui ont une énergie inférieure à la moyenne des énergies (cette moyenne est calculée sur la totalité des fenêtres).

II.2.3 Energie de la partie harmonique (harmonic part energy) :

C'est l'énergie de la partie harmonique du signal (signal sans bruit) pour chaque fenêtre du signal.

II.2.4 Energie du bruit (noise part energy) :

C'est l'énergie du bruit (la partie non harmonique) pour chaque fenêtre du signal.

II.3. DESCRIPTEURS SPECTRAUX [18]

Les descripteurs spectraux sont basés sur le spectre du signal. Ils décrivent donc l'évolution de l'enveloppe spectrale de celui-ci.

II.3.1 Centroïde Spectral

Le centroïde spectral est le centre de gravité du spectre d'amplitude de la Transformée de Fourier à Court Terme du signal (ou Short Time Fourier Transform en anglais). Il mesure la **brillance spectrale** des sons. L'équation (II.5) définit mathématiquement le centroïde spectral :

$$C_t = \frac{\sum_{n=1}^N M_t(n) \cdot n}{\sum_{n=1}^N M_t(n)} \quad (\text{II.5})$$

Où : $M_t(n)$ est la transformée de Fourier du signal à l'instant t et à la fréquence associée à l'indice n.

II.3.2 Etallement Spectral (Spectral Spread)

L'étallement spectral est une mesure servant à caractériser la dispersion des fréquences autour de leur moyenne (le centroïde). Il s'agit en fait d'une variance que l'on calculera grâce à la relation suivante :

$$ES = \frac{\sum_{n=1}^N (n - C_t)^2 \cdot M_t(n)}{\sum_{n=1}^N M_t(n)} = m_2 \quad (II.6)$$

II.3.3 Aplatissement Spectral (Spectral skewness)

L'aplatissement spectral caractérise l'asymétrie du spectre par rapport à sa moyenne. Il est nul pour un spectre symétrique, négatif s'il est plus concentré à droite de la moyenne et inversement, comme illustré sur les figures II.4.a, II.4.b et II.4.c

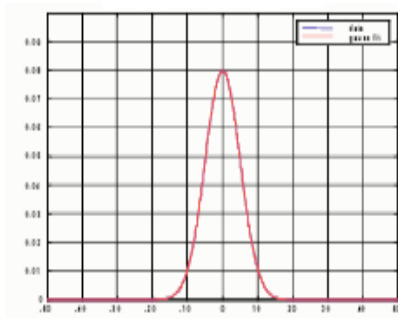


Figure II.4.a Skewness = 0

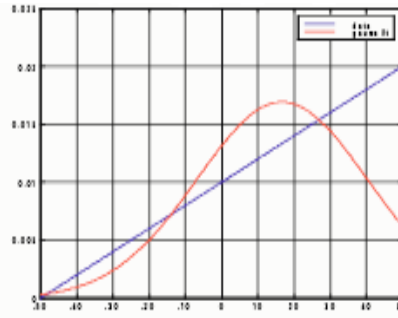


Figure II.4.b Skewness < 0

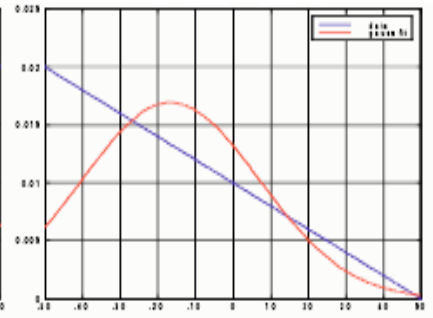


Figure II.4.c Skewness > 0

Il est donc défini comme :

$$AS = \frac{m_3}{\sigma^3} \quad (II.7)$$

Avec:
$$m_3 = \frac{\sum_{n=1}^N (n - C_t)^3 \cdot M_t(n)}{\sum_{n=1}^N M_t(n)} \quad \text{et} \quad \sigma^3 = (m_2)^{3/2}$$

II.3.4 Kurtosis Spectral

Il indique la forme du spectre (aplati ou en pointe). Il est égal à 3 pour une distribution normale, inférieur à 3 pour une distribution aplatie et supérieur à 3 pour une distribution en pointe, comme le montrent les Figures II.5.a, II.5.b et II.5.c.

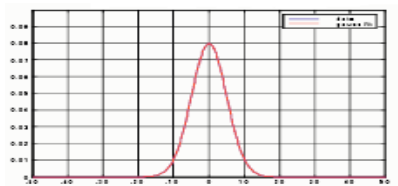


Figure II.5.a Kurtosis = 3

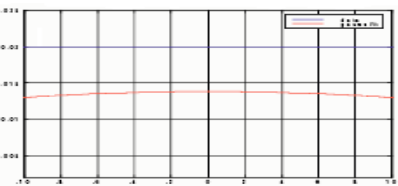


Figure II.5.b Kurtosis = 1.8

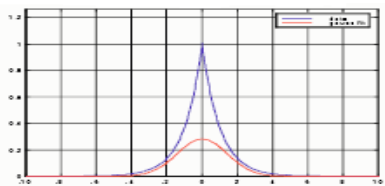


Figure II.5.c Kurtosis = 6

Il se calcule grâce à la relation : $Sk = \frac{m_4}{\sigma^4}$ (II.8)

Tel que $m_4 = \frac{\sum_n^N (n - C_t)^4 \cdot M_t(n)}{\sum_n^N M_t(n)}$ et $\sigma^4 = (m_2)^2$

II.3.5 Flux Spectral

Le flux spectral est le module au carré de la différence entre deux vecteurs spectraux consécutifs et traduit les variations locales du spectre du signal.

$$Flux = 1 - \frac{\sum_n M_{t-1}(n) \cdot M_t(n)}{\sqrt{\sum_n M_{t-1}(n)^2} \cdot \sqrt{\sum_n M_t(n)^2}} \quad (II.9)$$

Où : $M_t(n)$ et $M_{t-1}(n)$ sont les amplitudes du spectre aux instants t et t-1 du signal.

Lorsque le flux est égal à 0 les deux fenêtres sont similaires et dans le cas où le flux est égal à 1 elles sont complètement différentes.

II.3.6 Roulement Spectral (Spectral Roll-Off)

C'est la fréquence au dessous de laquelle les 95% de l'énergie sont concentrés.

$$\sum_{n=1}^R M_t[n]^2 = 0,95 \cdot \sum_{n=1}^N M_t[n]^2 \quad (II.10)$$

Où **R** est le spectral roll-off. Ce dernier est plus élevé pour les sons non voisés que pour les sons voisés.

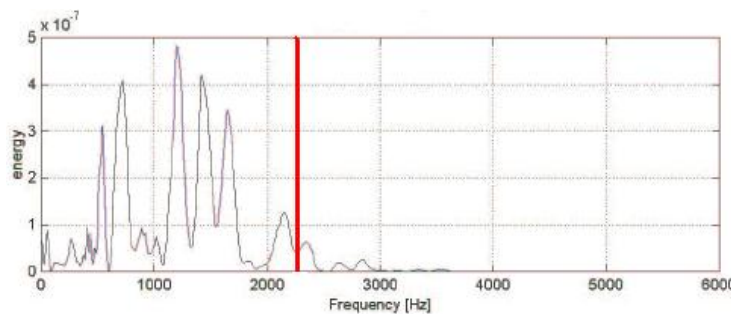


Figure II.6 Energie du spectre en fonction de la fréquence (Roulement Spectral en rouge).

II.3.7 Décroissance spectrale (Spectral Decrease)

Comme son nom l'indique, elle représente la décroissance du spectre. Bien que plusieurs formulations existent, nous avons choisi la relation suivante car elle provient de considérations perceptuelles bien adaptées au format de l'audition humaine.

$$DS = \frac{1}{\sum_{n=2}^N M_t[n]} \sum_{n=2}^N \frac{M_t[n] - M_t[1]}{n-1} \quad (II.11)$$

II.4. DESCRIPTEURS HARMONIQUES [26]

Les descripteurs musicaux se basent sur une analyse des propriétés fines du son, notamment ses propriétés harmoniques. Le signal est alors modélisé comme somme de trois composantes : sinusoïdes, transitoires et résidu (ou bruit).

II.4.1 Fréquence fondamentale et harmoniques

C'est l'harmonique de premier rang d'un son, elle détermine sa hauteur. Les harmoniques d'un son sont des multiples entiers de la fréquence fondamentale. Plusieurs méthodes ont été dédiées au calcul de la fréquence fondamentale (ou pitch) . Mais une difficulté supplémentaire s'ajoute dans le cas des signaux musicaux, à savoir la présence de pitch multiples (signaux polyphoniques), générant chacun une multitude d'harmoniques.

II.4.2 Déviation Spectre-Harmoniques (Harmonic Deviation)

C'est la déviation du spectre en amplitude du signal brut par rapport à celui de sa partie harmonique.

$$HDEV = \frac{1}{N} \sum_n (a(n) - M(n)) \quad (II.12)$$

N est le nombre d'harmoniques, $a(n)$ est l'amplitude de la $n^{\text{ième}}$ harmonique et $M(n)$ est l'amplitude du spectre évaluée à la fréquence associée à l'indice n .

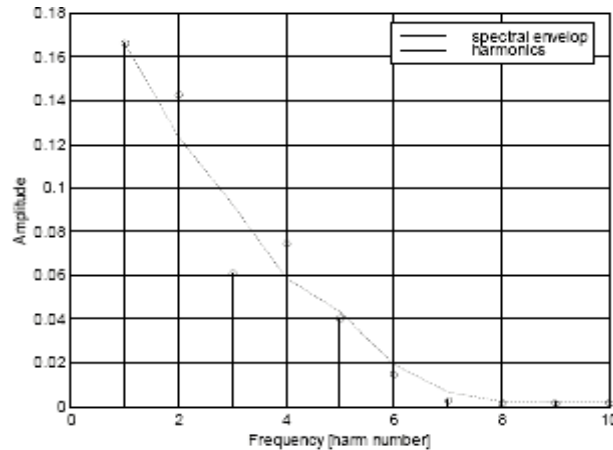


Figure II.7 Spectre du signal brut et Spectre de la partie harmonique.

II.4.3 Non-Harmonicité (Inharmonicity)

Elle représente la divergence en fréquences du spectre du signal total par rapport à celui de sa partie harmonique et vérifie donc si le signal total possède les mêmes fréquences que sa partie harmonique.

$$Inharmo = \frac{2}{f_0} \frac{\sum_n |f(n) - n \cdot f_0| \cdot M_t^2(n)}{\sum_n M_t^2(n)} \quad (II.13)$$

Ses valeurs sont comprises entre **0** et **1**. Elle prend la valeur **0** pour un signal purement harmonique et la valeur **1** pour un signal complètement bruité. f_0 est la fréquence fondamentale.

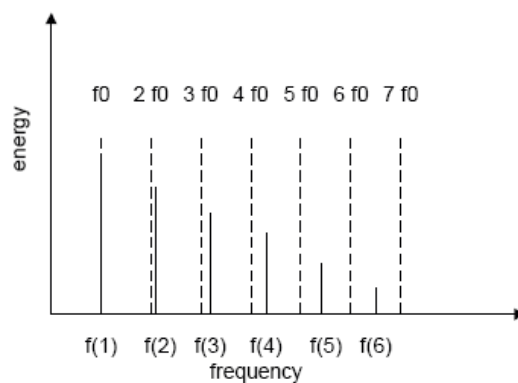


Figure II.8 Représentation des pics d'un signal (en lignes continues) et de ses harmoniques (en pointillés).

Ils sont calculés à partir d'un modèle harmonique du signal, ce qui semble bien adapté aux signaux musicaux en particulier.

II.5. DESCRIPTEURS PERCEPTUELS [25]

Les descripteurs perceptifs ou perceptuels sont issus de la psychoacoustique. Or, cette dernière est une branche d'une science plus générale : la psychophysique. La psychophysique étudie les relations entre les stimuli physiques et les sensations qu'ils induisent, autrement dit les relations entre un phénomène acoustique et la perception qu'il engendre. Ainsi, Mc Adams [] définit-il la psychoacoustique comme l'étude des relations quantifiables entre paramètres sonores (fréquence, amplitude, durée, etc.) et paramètres perceptuels (hauteur tonale, sonie, timbre, durée subjective). Une bonne connaissance du modèle auditif humain est donc nécessaire. Elle répondrait en partie à la question : comment percevons-nous les sons ? Nous décrirons, dans ce paragraphe, les descripteurs perceptuels qui sont la Sonie (spécifique et totale), l'acuité et enfin les coefficients MFCC (Mel Frequency Cepstral Coefficients).

II.5.1 Sonie (Loudness) [25]

Le système auditif humain est sensible à la pression acoustique i.e. aux variations de la pression atmosphérique provoquées par une onde acoustique. En acoustique, pour définir le niveau de pression, on utilise le décibel en niveau de pression acoustique ou dB Sound Pressure Level (dB SPL). Il est calculé à partir des valeurs de références en pression P_0 , ou en intensité I_0 , telles que : $P_0 = 20\mu \text{ Pa}$ et $I_0 = 10^{-12} \text{ W/m}^2$. La pression acoustique de référence P_0 est choisie comme la plus petite valeur de pression acoustique audible (I_0 est l'intensité correspondante). Le niveau de pression acoustique en décibels L_{dB} est alors défini en fonction de l'intensité acoustique I , ou de la pression efficace P par la relation :

$$L_{dB} = 10 \cdot \log\left(\frac{I}{I_0}\right) = 20 \cdot \log\left(\frac{P}{P_0}\right) \quad (\text{II.14})$$

Le niveau de pression acoustique est une mesure physique à laquelle correspond une sensation auditive connue sous le nom de **Sonie** (ou **Loudness** en anglais). La sonie correspond au niveau acoustique perçu. Elle permet une meilleure prédiction que le niveau acoustique car son calcul tient compte aussi bien de la répartition de l'énergie en fonction des bandes critiques et des effets de masquage. Par ailleurs, elle représente un attribut dominant de la qualité acoustique. Stevens proposa en 1955 [I.8], la relation suivante entre sonie et intensité acoustique :

$$S = k \cdot I^n \quad \text{avec } n = 0.3 \text{ et } k \text{ constante.} \quad (\text{II.15})$$

Où l'unité utilisée pour mesurer la sonie est le sone, de sorte que si l'intensité est multipliée par un facteur de 10, la sonie est elle multipliée par 2. Une autre échelle de mesure de l'intensité sonore subjective est le phone. Par définition, le niveau sonore en phones d'un son est la valeur du niveau sonore en dB SPL d'un son pur de 1 kHz évoquant la même sensation d'intensité. L'échelle des phones coïncide avec l'échelle des dB SPL pour 1 KHz. La sonie S en sonexprime alors en fonction du niveau Ph en phones grâce à la relation suivante :

$$S = 2^{(Ph-40)/10} \quad (\text{II.16})$$

Il est possible de voir, d'après la relation précédente, que la valeur de 1 sone est la sonie d'un son de 40 phones. Au dessus de 40 phones, la sonie double pour toute augmentation de 10 dB du niveau en phones. En deçà de la valeur suscitée, la sonie croît plus rapidement.

- Calcul de la Sonie

Zwicker a proposé un modèle de calcul de la sonie, qui a été érigé en norme (DIN 45 631 et ISO 532B). Trois étapes de calcul sont essentielles. La première consiste à filtrer le signal sonore pour tenir compte de la transmission acoustique à travers les oreilles externe et moyenne. L'étape suivante consiste à calculer les patterns d'excitation du son mesuré sur l'échelle usuelle des fréquences (Hz). Ces patterns sont ensuite recalculés sur l'échelle Bark. Finalement, les sonies spécifiques N_0 pour chacune des bandes critiques (en Bark) sont calculées en fonction de l'excitation E exprimée en unités de puissance. La sonie globale N est alors la somme des sonies spécifiques. Une approximation de la sonie spécifique est alors donnée par la relation suivante :

$$N' \cong \left(\frac{E}{E_0} \right)^{0.23} \quad (\text{II.17})$$

Avec E_0 excitation de référence (puissance) correspondant à un niveau acoustique 0 dB SPL.

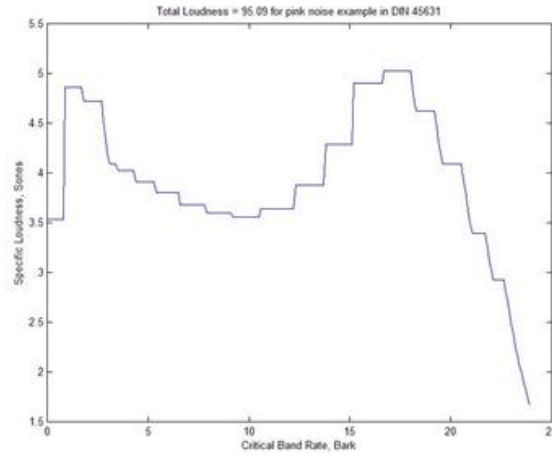


Figure II.9. Sonie spécifique

La sonie globale est définie par la relation suivante :

$$N = \int_0^{24 \text{Bark}} N' dz \quad (\text{II.18})$$

N est exprimée en sonos.

Le modèle de Zwicker produit une bonne estimation de la sonie pour des sons purs et des bandes de bruit, mais aussi pour des sons naturels stationnaires. Notons que la sonie joue un rôle majeur pour l'évaluation de la gêne ou du désagrément (bruits environnementaux, bruits synthétiques).

La notion de sonie pondérée (ou weighted loudness) est aussi souvent employée et elle est définie par la relation :

$$N_w = \int_0^{24 \text{Bark}} g(z).N'.dz \quad (\text{II.19})$$

où $g(z)$ est une fonction de pondération qui permet de rehausser certaines plages de fréquences.

II.5.2 Acuité (Sharpness)

L'acuité est calculée à partir des sonies spécifiques et totales. C'est l'équivalent du centroïde spectral mais calculé avec les sonies spécifiques au lieu des spectres d'amplitude. Elle est donc calculée comme suit :

$$A = 0.11 \cdot \frac{\int_0^{24 \text{Bark}} z.g(z).N'.dz}{N} \quad (\text{II.20})$$

Avec :
$$g(z) = \begin{cases} 1 & \text{pour } z < 15 \\ 0.066 \cdot e^{(0.171z)} & \text{pour } z \geq 15 \end{cases}$$

II.5.3 Mel Frequency Cepstral Coefficients (MFCC) [24]

Historiquement les MFCC ont été introduits en premier lieu dans le traitement de la parole, en reconnaissance de la parole et en identification du locuteur, les résultats obtenus montrent leur grande efficacité et leur parfaite adaptation au signal de parole. En effet, et de par leur construction, ils déconvoluent (séparent) les contributions de la source d'excitation et du conduit vocal. Récemment de nombreux travaux ont utilisé les coefficients MFCC dans le domaine des signaux musicaux. Leur principale motivation est que dans le cas de la musique, et plus spécifiquement dans le cas des instruments à cordes, le signal musical est alors considéré comme produit de convolution entre une excitation (pincement des cordes) et un filtre (cavité résonnante de l'instrument). La Figure (II.10) illustre les principales étapes pour l'extraction des coefficients MFCC. Nous détaillerons ensuite chaque étape à part.

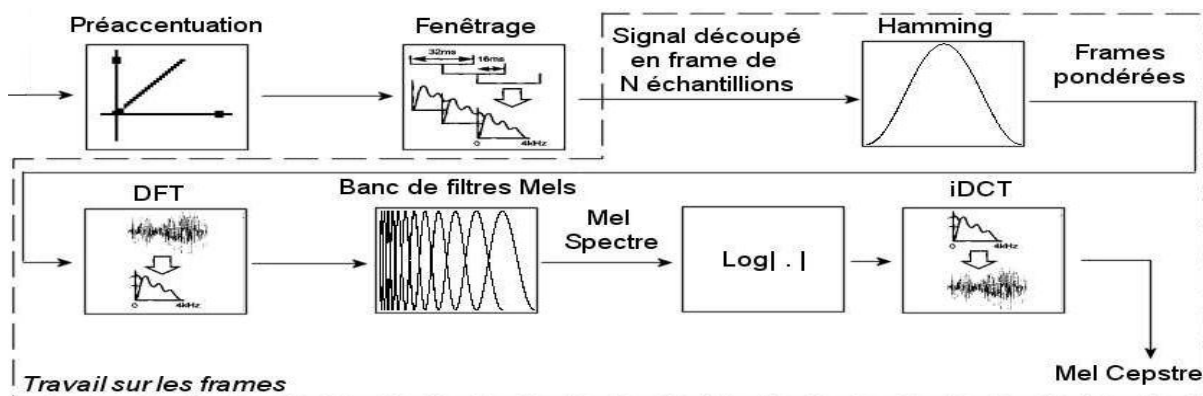


Figure II.10. Etapes de calcul des coefficients MFCC.

A- Préaccentuation

Une accentuation des aigus est nécessaire car les composantes fréquentielles aiguës sont toujours plus faibles que les graves. Un filtrage de type passe-haut est réalisé (sur tout le signal) grâce à la fonction de transfert $H(z)$ telle que :

$$H(z) = 1 - \check{U}z^{-1} \quad (II.21)$$

$\check{U}$ est le coefficient de pré accentuation ; il doit être compris dans l'intervalle $[0,1]$. En fait, $\check{U}$ est souvent choisis de 0,9 à 1. Dans notre cas, nous avons choisi $\check{U} = 0.97$.

B- Fenêtrage

Afin d'analyser un signal sur toute sa longueur, on le subdivise en plusieurs parties quasi

stationnaires (20 à 30ms), L'utilisation d'une fenêtre de Hamming (sur une trame acoustique), avec un recouvrement de moitié, permet d'éviter la formation d'artefacts liés aux effets de bord.

C- Transformée de Fourier

Le spectre du signal sur chaque trame est calculé par le biais de la transformée de Fourier rapide (FFT) : nous passons ainsi du domaine temporel au domaine fréquentiel. En sortie de ce module, nous disposons de la densité spectrale à court terme du signal puisque nous calculons le module carré de la transformée de Fourier.

D- Echelle MEL et Filtres Triangulaires

Si nous admettons que l'oreille humaine est le meilleur outil pour suivre un signal audio, il faut utiliser ce que l'oreille entend réellement comme signal. Pour simuler l'oreille humaine, la densité spectrale, obtenue lors de l'étape précédente, est filtrée par un banc de filtres triangulaires. Les filtres utilisés sont espacés de telle façon que leur évolution correspond à l'échelle Mel. Cette échelle tente de mettre en relation la tonie et les hauteurs perçues. Elle n'établit pas une correspondance entre les fréquences et les Mels mais tente d'établir un lien entre les sensations de hauteurs perçues. Elle met principalement une chose en valeur, c'est que sur le plan de la perception, une octave entre deux sons graves, 250 et 500Hz par exemple (250 Mels), paraît plus petite qu'une octave entre deux sons aigus, 1000 et 2000 Hz (1800 Mels). Il est possible de limiter les bornes du banc de filtre afin de calculer les MFCC sur une bande de fréquence particulière. Le nombre de filtres utilisés est aussi variable suivant les applications. Dans la littérature le nombre de 24 filtres est souvent retenu. Le spectre calculé à l'étape précédente est convertie à l'échelle Mel puis traité par le banc de filtres décrit précédemment. Les coefficients ainsi obtenus sont ensuite décorrélés grâce à la Transformée en Cosinus Discrète (DCT pour Discrete Cosine Transform). Le nombre de coefficients MEL est donc réduit : 13 pour la parole et à 5 pour la musique.

Remarques

- Le premier coefficient MEL représente l'énergie moyenne du signal et il est souvent ignoré dans le vecteur de paramètres.
- Certains auteurs préconisent d'utiliser également les dérivées premières et secondes des MFCC dans le vecteur acoustique des paramètres. Ces paramètres sont les dérivées temporelles des MFCC. Ils permettent de prendre en compte la variabilité temporelle du signal ce qui est une de ses caractéristiques importantes.

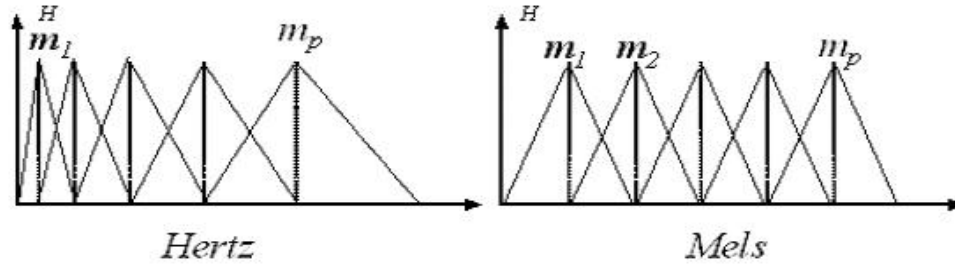


Figure II.11. Répartition des filtres triangulaires sur une échelle linéaire (Hz) et sur une échelle Mel.

II.6. DESCRIPTEURS ONDELETTES

II.6.1 Analyse en ondelettes [16,22]

L'analyse en ondelettes réalise une décomposition des signaux grâce à une famille de fonctions localisées en temps. Ces fonctions, auxquelles on réserve classiquement le terme « ondelettes » sont issues d'une même fonction dite ondelette mère; La notion de fréquence est remplacée par la notion d'échelle et pour tenir compte des hautes ou basses fréquences, l'ondelette mère sera contractée ou dilatée. Nous avons donc deux paramètres en jeu : le facteur d'échelle (ou de dilatation) relié à la notion de fréquence et le décalage relié à la notion de position temporelle. Une ondelette est une fonction $\psi \in L^2(\mathfrak{R})$ de moyenne nulle :

$$\int_{-\infty}^{+\infty} \psi(t) dt = 0 \quad (\text{II.22})$$

Où $L^2(\mathfrak{R})$ désigne l'espace des signaux d'énergie finie c.a.d tel que $\int |\psi(t)|^2 dt < +\infty$. Elle est normalisée à 1 : $\|\psi(t)\| = 1$. Nous désignons par $\psi_{a,b}(t)$ les ondelettes obtenues par dilatation d'un facteur a ($a \in \mathfrak{R}_+$) et décalage d'une position b ($b \in \mathfrak{R}$) de l'ondelette mère $\psi(t)$ conformément à l'équation suivante :

$$\psi_{a,b}(t) = \frac{1}{\sqrt{a}} \psi\left(\frac{t-b}{a}\right) \quad (\text{II.23})$$

Ces ondelettes restent de norme 1: $\|\psi_{a,b}\| = 1$, et on appelle transformée en ondelette, la transformation intégrale qui à toute fonction f appartenant à $L^2(\mathfrak{R})$ (Espace des fonctions réelles d'énergie finie) fait correspondre la fonction $T^{ond} f(a,b)$ définie par :

$$T^{ondelette} f(a,b) = \frac{1}{\sqrt{a}} \int_{-\infty}^{+\infty} f(t) \psi\left(\frac{t-b}{a}\right) dt = \langle f, \psi_{a,b} \rangle \quad (\text{II.24})$$

La variable a joue le rôle de l'inverse de la fréquence : plus a est petit, moins l'ondelette (la fonction analysante) est étendue temporellement, donc plus la fréquence centrale de son spectre est élevée. Par conséquent les grandes valeurs de a seront logiquement associées aux

hautes fréquences. Le facteur d'échelle comporte un aspect translation dans le temps amené par le terme b , mais aussi un aspect dilatation à la fois en temps et en amplitude apporté par les termes a et $\sqrt{a}$. La dilatation en amplitude permet de conserver une norme constante pour tous les éléments de la base (énergie de l'ondelette).

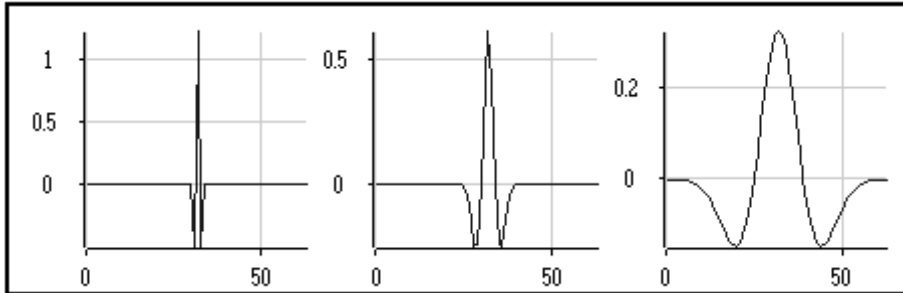


Figure II.12. : dilatation d'une ondelette.

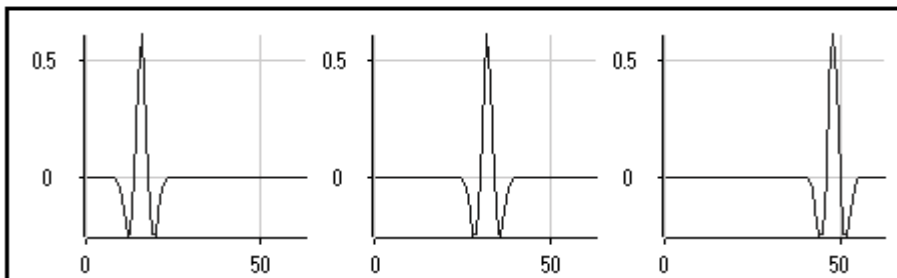


Figure II.13. : Translation d'une ondelette.

II.6.2 Types de transformées en ondelettes

A- Transformée en ondelette continue

$T^{\text{ondelette}} f(a,b)$ est la transformée en ondelettes continue de f par l'ondelette mère ψ , car les paramètres a et b varient continûment (et non parce que $T^{\text{ondelette}} f(a,b)$ est nécessairement une fonction continue). On notera la transformée en ondelette continue d'une fonction f par CWT $f(a,b)$. La reconstruction de f à partir de sa transformée en ondelettes n'est possible que si ψ vérifie la condition d'admissibilité. Auquel cas on a :

$$f(t) = \frac{1}{C_\psi} \int_0^{+\infty} \frac{da}{a^2} \int_{-\infty}^{+\infty} \text{CWT}f(a,b) \psi_{a,b}(t) db \quad (\text{II.25})$$

B- Transformée en Ondelettes Discrète

La transformée continue en ondelettes décompose un signal selon un ensemble d'ondelettes déduites d'une ondelette de référence par des opérations de dilatation et de translation, selon un facteur d'échelle a et un facteur de translation b variant continûment dans $\mathbf{R}$. Cependant, il est possible d'augmenter l'efficacité de la décomposition, c'est-à-dire de limiter le nombre de facteurs d'échelle à utiliser tout en conservant la même précision dans la décomposition. Pour cela, les facteurs a et b sont discrétisés, et dans ce cas là, la transformée obtenue est dite discrète. Ainsi la transformée en ondelettes discrète nécessite de trouver une sous-famille dénombrable $(\psi_{2^j 2^k t - k})_{j,k \in \mathbb{Z}}$ de la famille $(\psi_{a,b})_{a,b \in \mathbf{R}}$ et formant une base de $L^2(\mathbf{R})$. Deux

choix sont possibles et conduisent à deux algorithmes rapides de calcul.

➤ 1^{er} choix : Prendre $a_j = 2^j$, $b_k = 2^j k$, avec $j, k \in \mathbb{Z}$.

Ce type de choix est dit échelle dyadique. Les familles d'ondelettes obtenues sont orthogonales. L'algorithme pyramidal rapide de Mallat [1] permet un calcul efficace de la transformée en ondelette discrète (DWT).

➤ 2^{ème} choix : Prendre $a_j = 2^j$, $b_k = k$, avec $j, k \in \mathbb{Z}$.

Les familles d'ondelettes sont non orthogonales et dans ce cas, la transformée est redondante. L'algorithme à trous permet le calcul de la transformée dans ce cas.

Dans les deux cas, la famille d'ondelettes discrétisées est définie par :

$$\psi_{j,k}(t) = 2^{\frac{j}{2}} \psi(2^j t - k) \quad (\text{II.26})$$

Et la transformée en ondelettes discrète (DWT) par :

$$\text{DWT}f(j,k) = \langle \psi_{j,k}, f \rangle = \int_{-\infty}^{+\infty} f(t) \psi(2^j t - k) dt \quad (\text{II.27})$$

C- Algorithme de Mallat pour le calcul de la DWT

Nous avons choisi d'utiliser l'algorithme de Mallat [16] pour calculer la DWT ; En effet, puisque nous utiliserons les coefficients de la DWT pour construire les descripteurs ondelettes, la DWT fournie par cet algorithme est orthogonal et évite toute redondance. L'algorithme de Mallat est fondée sur l'analyse mutirésolution [18]. C'est un algorithme pyramidal efficace de calcul de la DWT. Le principe de cet algorithme est de séparer le signal en deux composantes, l'une représentant l'allure générale du signal, l'autre représentant ses détails. L'allure générale d'une fonction est représentée par ses basses fréquences, les détails par ses hautes fréquences. Pour séparer les deux, nous avons donc besoin d'une paire de filtres : un filtre passe-bas H pour obtenir l'allure générale (aussi appelée **approximation**), et un filtre

passe-haut G pour estimer ses **détails**, c'est-à-dire les éléments qui varient rapidement. Pour ne pas perdre d'information, ces deux filtres doivent bien sûr être complémentaires : les fréquences coupées par l'un doivent être conservées par l'autre. Les filtres H et G sont appelés des filtres miroirs conjugués. La Figure II.14 illustre l'analyse et la synthèse du signal par l'algorithme de Mallat. La DWT sera formée de l'ensemble des signaux de détail et de la dernière approximation.

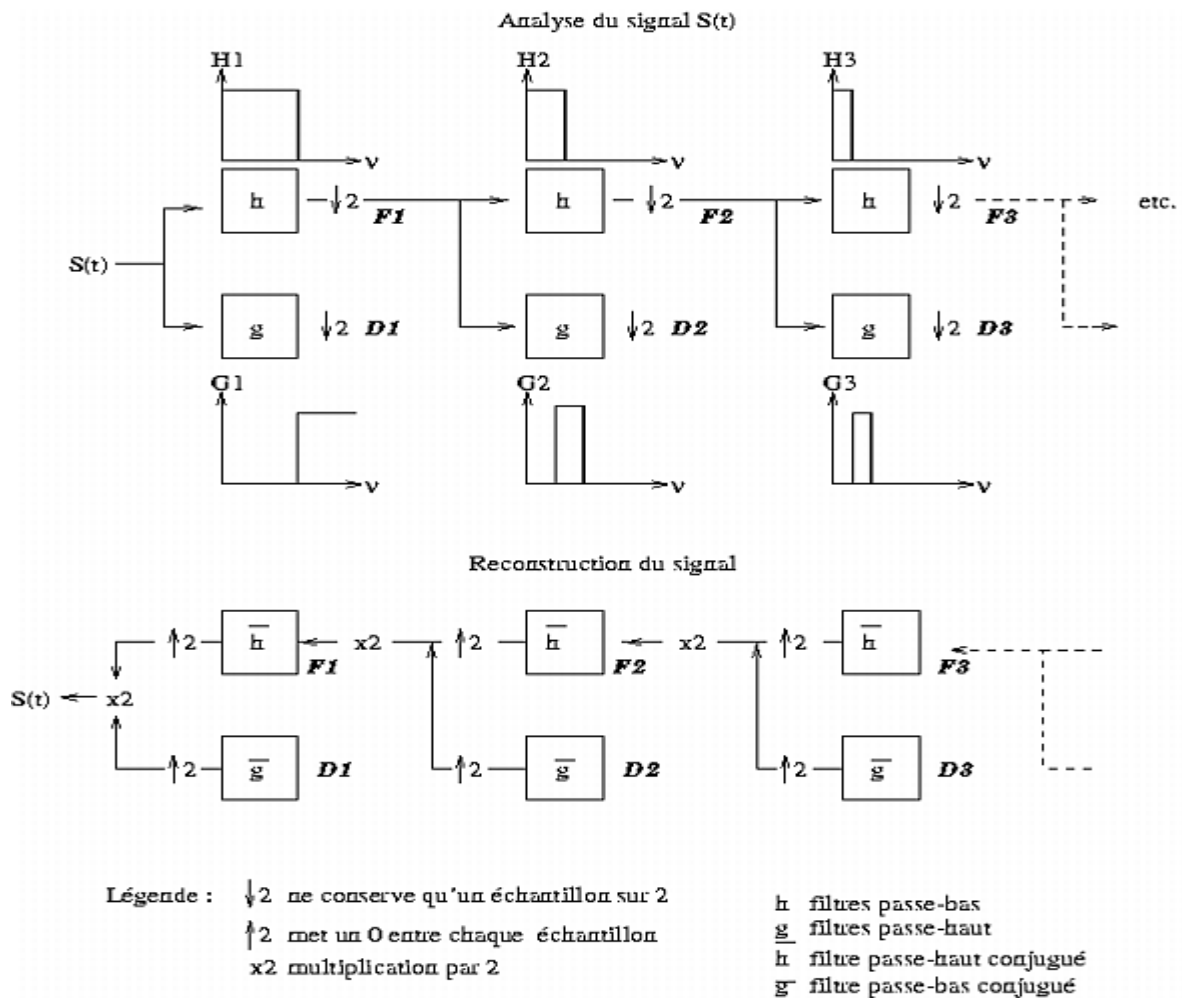


Figure II.14. Analyse-Synthèse d'un signal (Algorithme de Mallat).

D- Choix des filtres ondelettes pour les signaux musicaux [17]

La qualité de la décomposition en ondelettes dépend en grande partie de la capacité des ondelettes choisies (i.e. filtres) à approximer le signal à analyser. Lorsque la forme de l'ondelette ne ressemble pas suffisamment à celle du signal à analyser, les coefficients en ondelettes obtenus ne concentrent pas les principales caractéristiques du signal ; Le nombre de

coefficients non nuls est donc élevé. L'analyse est d'autant meilleure qu'il y a un nombre peu élevé de coefficients qui seront cependant significatifs.

Les signaux musicaux sont constitués d'ondes assez lisses et harmoniques. Le choix d'une ondelette d'allure lisse semble donc judicieux : elle doit donc être la plus régulière possible. La bande de transition des filtres passe-bas et passe-haut est aussi un facteur important. En effet, une bande de transition trop large (soit donc une décroissance lente de filtre) cause un recouvrement des bandes passantes. Les bandes passantes en sortie du banc de filtres ne sont pas correctement séparées et un effet de repliement détériore la qualité des coefficients. Les bandes des filtres passe bas et passe haut doivent donc être bien séparées. De plus, Les réponses des filtres utilisées doivent être à phase linéaire (ou à phase la plus linéaire possible), pour éviter toute distorsion dans le domaine fréquentiel. Une réponse en phase linéaire est assurée si les filtres sont symétriques. Enfin, les ondelettes ont des localisations temporelles différentes ; Les ondelettes dont le support est le plus court et dont la décroissance est la plus rapide assurent la meilleure localisation temporelle. Notons enfin que plus les filtres sont courts plus l'implémentation sera rapide ; Cependant cette condition est en général contradictoire avec la régularité de l'ondelette et des bandes de transition étroites pour les filtres. Un compromis raisonnable est donc nécessaire. Nous allons citer et définir ci-dessous quelques filtres ondelettes les plus utilisés.

- *Ondelette de Haar*

L'ondelette de Haar est spéciale. Elle produit un filtre à 2 coefficients (implémentation rapide). C'est la seule ondelette qui soit symétrique et orthogonale. Cependant de forme carrée, elle ne s'adapte pas à l'analyse de signaux musicaux.

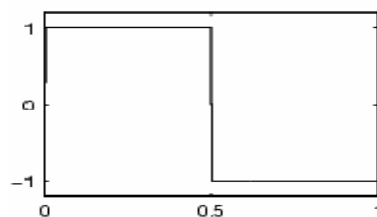


Figure II.15. Ondelette de Haar

- *Ondelette de Morlet*

Cette ondelette, inspirée du signal élémentaire de Gabor, est obtenue par modulation d'une gaussienne :

$$\psi(t) = (\pi t_0)^{-\frac{1}{4}} \exp \left[-\frac{1}{2} (2\pi t_0 j \nu_0)^2 \right] \quad (\text{II.28})$$

Une telle ondelette a l'avantage d'être explicite. Cependant, elle n'est pas admissible car sa transformée de Fourier ne s'annule pas à l'origine.

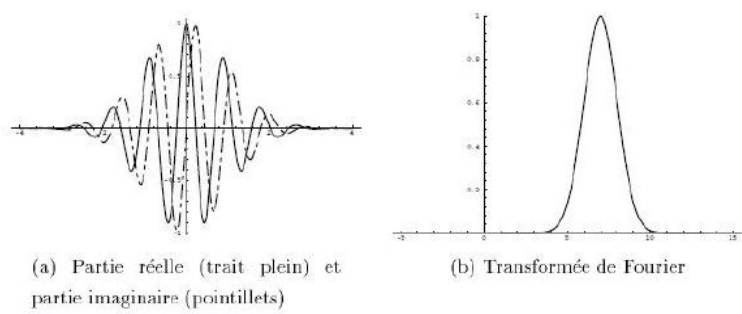


Figure II.16. Ondelette complexe de Morlet

L'ondelette de Morlet et sa Transformée de Fourier ont la propriété d'appartenir aux espaces des fonctions infiniment dérivables et à décroissance rapide. Elle est bien localisée en temps et en fréquence, et numériquement elle est considérée à support compact. De plus, elle est complexe, ce qui permet une analyse en module et phase. Sa régularité permet de l'utiliser pour l'étude des régularités des fonctions.

- *Chapeau Mexicain*

Le chapeau mexicain est une ondelette réelle qui doit son nom à sa forme. Elle est construite à partir de la dérivée seconde de la gaussienne :

$$\psi_{CM}(t) = (1 - t^2) \exp\left(-\frac{1}{2}t^2\right) \quad (\text{I.29})$$

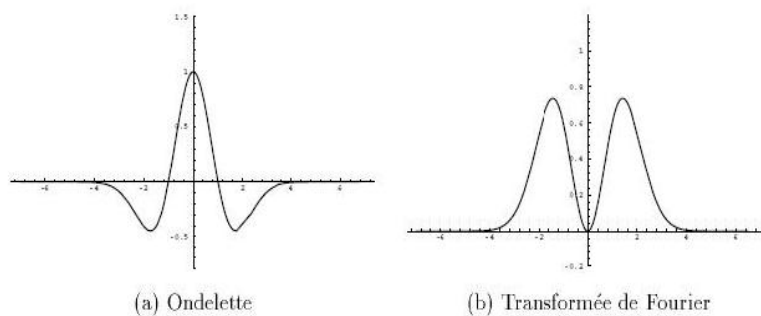


Figure II.17. Ondelette chapeau mexicain

Cette ondelette est symétrique, ce qui permet de ne pas introduire de déphasages dans la transformée en ondelettes. Elle est particulièrement adaptée à la détection des discontinuités

- *Ondelette de Meyer*

Cette classe d'ondelettes orthogonales est construite dans le domaine de Fourier grâce à la relation :

$$\hat{\psi}(\xi) = \begin{cases} (2\pi)^{-\frac{1}{2}} e^{i\frac{\xi}{2}} \sin \left[\frac{\pi}{2} v \left(\frac{3}{2\pi} |\xi| - 1 \right) \right], & \frac{2\pi}{3} \leq |\xi| \leq \frac{4\pi}{3} \\ (2\pi)^{-\frac{1}{2}} e^{i\frac{\xi}{2}} \sin \left[\frac{\pi}{2} v \left(\frac{3}{4\pi} |\xi| - 1 \right) \right], & \frac{4\pi}{3} \leq |\xi| \leq \frac{8\pi}{3} \\ 0, & \text{sinon} \end{cases}$$

avec v , une fonction de classe C^k (k arbitraire pouvant être égale à l'infini) telle que :

$$v(x) = \begin{cases} 0, & \text{si } x \leq 0 \\ 1, & \text{si } x \geq 1 \end{cases}$$

Avec $v(x) + v(1 - x) = 1$.

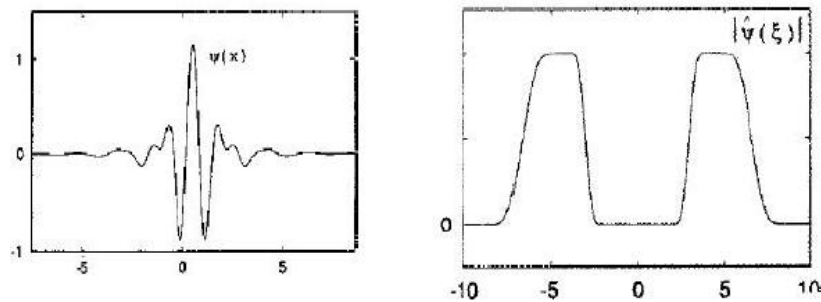


Figure II.18. Ondelette de Meyer et sa transformée de Fourier

- *Ondelettes de Daubechies* [21]

Les ondelettes à support compact et orthogonales de Daubechies ont été créées par Ingrid Daubechies dans les années 1980. Elles ont permis le réel essor de la transformée en ondelettes discrète. Les ondelettes de Daubechies donnent la possibilité de choisir le degré de régularité voulue en imposant un certain nombre de moments nuls p : le nombre de coefficients des filtres est alors égal à $2p$. Les ondelettes ou filtres qui en résultent sont appelées Daubechies p ou simplement D_p . Pour $p=1$, on retrouve l'ondelette de Haar. Le support compact assure une bonne résolution temporelle. Cependant elles ne sont pas symétriques, ce qui peut causer des problèmes dans le cas de signaux audio. La Figure II. 12 représente différentes ondelettes de Daubechies ainsi que les réponses en fréquences des filtres associés pour différentes valeurs de p .

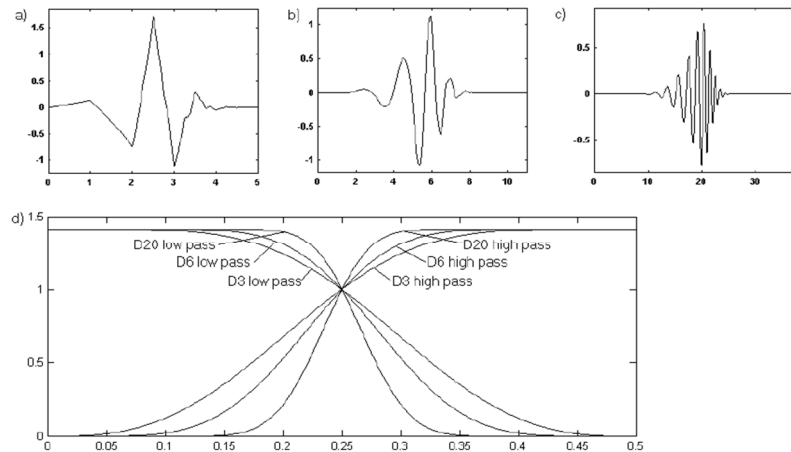


Figure II.19. Ondelettes de Daubechies et réponses en fréquences pour différentes valeurs de p

- *Symlets*

Daubechies a également construit une série de d'ondelettes à support compact, orthogonales et possédant p moments nuls. Elle a ajouté des contraintes supplémentaires de phase linéaire et de symétrie. Ces ondelettes ont été baptisées symlets. Les symlets ne sont pas parfaitement symétriques mais sont les ondelettes orthogonales « les moins asymétriques ».

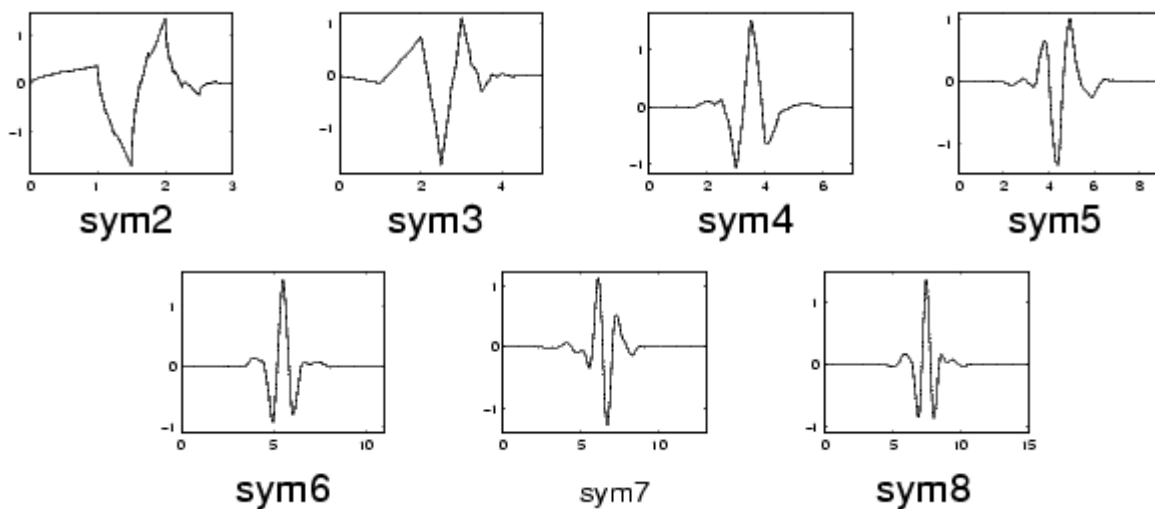


Figure II.20 : Ondelettes Symlets pour différentes valeurs de p

- *Coiflets*

A la demande de R.Coifman, I. Daubechies créa les coiflets. En plus de propriétés identiques à celles des ondelettes de Daubechies et des symlets (support compact, orthogonalité et p moments nuls), la contrainte supplémentaire voulue par Coifman consiste à imposer p moments nuls à la fonction d'échelle. Cette contrainte fait que les coefficients d'approximation sont dans ce cas identiques aux échantillons du signal.

Seulement, la longueur du filtre Coiflet est supérieure à celles des autres filtres de Daubechies ($6p$ au lieu de $2p$) ; La figure II.21 donne une représentation des ondelettes de Coifman.

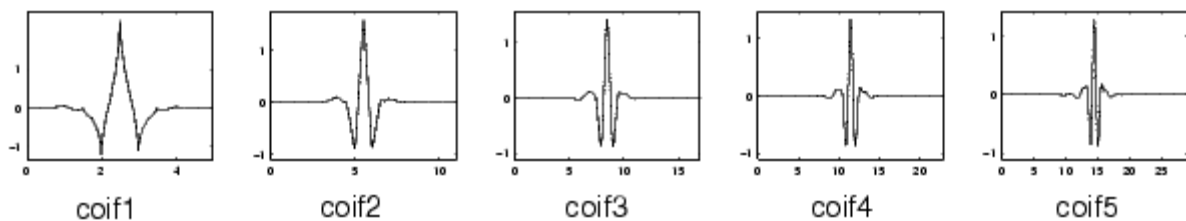


Figure II.21. Ondelettes Coiflets pour différentes valeurs de p

- *Ondelettes biorthogonales*

L'orthogonalité est un cas particulier de la biorthogonalité. La notion de biorthogonalité a été introduite par Tchamitchian en 1967 et celles des ondelettes biorthogonales par Cohen et al. Dans la pratique, les ondelettes orthogonales présentent un grand intérêt car elles facilitent l'inversion de la transformée ; La famille d'ondelettes orthogonales et à support compact est la plus intéressante, les filtres étant à réponse impulsionnelle finie ce qui facilite leur implémentation. L'orthogonalité permet d'avoir une bonne qualité de reconstruction du signal, cependant les ondelettes orthogonales n'étant pas symétriques, elles introduisent des distorsions. Les ondelettes biorthogonales ont été introduites pour pallier à cet effet. La base permettant la décomposition (analyse) n'est plus la même que celle permettant la reconstruction (synthèse). L'ondelette et la fonction d'échelle sont symétriques et les filtres ont impérativement un nombre impair de coefficients. Battle et Lemarié ont introduit des ondelettes biorthogonales basées sur les splines polynomiaux. Pour des splines d'ordre m , l'ondelette possède $m+1$ moments nuls. Elles ne sont pas à support compact : les filtres ne sont plus à réponse impulsionnelle finie ; Cette dernière ne peut être estimée que par troncature. Si m est impair les ondelettes sont symétriques. Et grâce aux propriétés des splines, elles sont $m-1$ dérivables et donc parfaitement régulières. Les binlets sont également des ondelettes biorthogonales à coefficients réels faciles à implémenter.

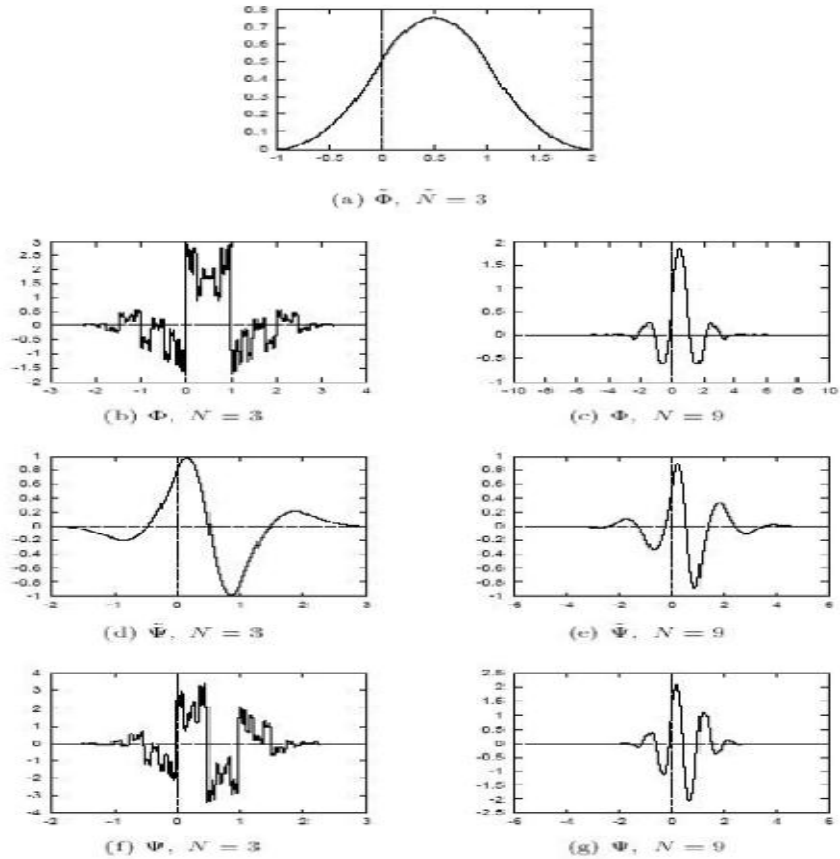


Figure II.22. Ondelettes splines et fonctions d'écaille associées

II.6.2 Descripteurs ondelettes

Une fois les coefficients de la transformée en ondelettes discrète obtenus, il reste cependant à construire le vecteur acoustique à associer à chaque fichier musical. Il est impossible de garder tous les coefficients car même si la transformée en ondelettes est orthogonale le nombre de coefficients d'ondelettes est égal aux échantillons du signal. Nous avons recensé dans la littérature, les descripteurs suivants issus des coefficients en ondelettes [9, 11].

A- Moyenne de la valeur absolue des coefficients dans chaque sous-bande

Ce descripteur donne des informations sur la distribution en fréquence du signal. Il est calculé sur chaque échelle puisqu'une échelle correspond à une sous-bande.

B- Ecart-Type des coefficients dans chaque sous-bande

Ce descripteur donne des informations sur les changements qui se produisent dans la distribution en fréquences du signal.

C- Rapport des moyennes des coefficients entre deux sous bandes adjacentes

Ce descripteur contient des informations sur le contenu fréquentiel du signal mais également sur son harmonicité puisqu'il compare le contenu fréquentiel sur deux bandes adjacentes (échelles consécutives).

D- Trois premiers moments

Une autre série de descripteurs acoustiques est construite à partir des histogrammes des coefficients. Pour caractériser la distribution de la forme d'onde les trois premiers moments de l'histogramme sont utilisés qui sont : la moyenne, la variance et le skewness de chaque sous bande (échelle).

II.6. COMBINAISON DE DESCRIPTEURS ET REDUCTION DE DIMENSIONNALITE

Plusieurs travaux de l'état de l'art [5,6,23] montrent que la combinaison de plusieurs descripteurs hétérogènes permet d'améliorer les performances des applications. L'idée de base soutenue dans [15] est que la construction d'un vecteur acoustique étendu composé de la concaténation de plusieurs descripteurs hétérogènes permet une complémentarité entre les différents paramètres dans le sens que l'information manquée par un descripteur est capturée par un ou plusieurs autres. Ainsi, cette combinaison doit être menée dans le sens à regrouper des informations temporelles, spectrales et perceptuelles ou autres. Cette façon de procéder a des conséquences sur la redondance d'informations, c'est pourquoi nous avons exploité des méthodes de décorrélation de descripteurs afin de réduire la dimensionnalité des données traitées. Aussi, nous proposons dans la partie résultats de rendre la combinaison de descripteurs indépendants par analyse en Composantes Indépendantes. Les expériences menées sur une base de signaux audio réels montrent l'efficacité et la pertinence d'une telle approche.

Conclusions

Le choix d'un espace de représentation de signaux audio est crucial et conditionne les performances de la chaîne complète d'indexation et de classification. Un espace de représentation idéal doit être capable de caractériser la totalité des classes sonores tout en conservant l'information temporelle. Celle-ci est primordiale afin de développer des systèmes performants de classification sonore. Ainsi, nous avons relaté dans ce chapitre, les descripteurs d'ordre général permettant de caractériser les propriétés temporelles et spectrales des signaux audio ainsi que des descripteurs plus spécifiques aux signaux musicaux. Des descripteurs perceptuels basés sur une analogie avec le système auditif humain ont également été étudiés. Nous avons également introduits les descripteurs ondelettes construits à partir de la DWT. Nous proposons de concaténer différentes combinaisons de descripteurs temporels, spectraux et perceptuels pour former les vecteurs acoustiques qui seront associés à chaque extrait musical. Cette concaténation peut entraîner une redondance d'information préjudiciable. Nous avons alors pensé à rendre ces coefficients statistiquement indépendants pour éviter cet inconvénient. Les descripteurs ondelettes nous ont semblé intéressants car la DWT est aussi bien adapté aux signaux stationnaires qu'aux signaux musicaux par sa nature même. Les performances de tous ces descripteurs seront évaluées et comparées dans le chapitre résultats.

- [1] Liu, Z. Wang, Y. and Chen, T. "Audio Feature Extraction and Analysis for Scene Segmentation and Classification," *Journal of VLSI Signal Processing Systems* 20, 61-49, June 1998.
- [2] Lu, L., Zhang, H.-J., and Jiang, H. "Content Analysis for Audio Classification and Segmentation," *IEEE Trans. Speech Audio Processing*, 10(7), 504-516, 2002.
- [3] Lu, L., Zhang, H.-J., and Li, S., "Content-based Audio Classification and Segmentation by Using Support Vector Machines," *ACM Multimedia Systems Journal*, 8(6), 482-492, 2003
- [4] Cai, R., Lu, L., Zhang, H.-J., and Cai, L.-H. "Improve Audio Representation by Using Feature Structure Patterns," *Proc. 29th IEEE Int'l Conf. on Acoustics, Speech, and Signal Processing*, vol. 4, 345-348, 2004.
- [5] Saunders, J. "Real-time Discrimination of Broadcast Speech/Music," in *Proc. Int'l Conf. on Acoustics, Speech, and Signal Processing* 96, Vol. II, 993-996, 1996.
- [6] Scheirer, E., and Slaney, M. "Construction and Evaluation of a Robust Multifeature Music/Speech Discriminator," in *Proc. Int'l Conf. on Acoustics, Speech, and Signal Processing* 97, Vol II, 1331-1334, 1997.
- [7] Zhang, T., and Kuo, C.-C. J. "Video Content Parsing based on Combined Audio and Visual Information," in *Multimedia Storage and Archiving Systems IV, Proc. of SPIE 3846*, Vol. IV, 78-89, 1999.
- [8] Tzanetakis, G., and Cook, P. "MARSYAS: a Framework for Audio Analysis," *Organised Sound*, 4: 169-175, Cambridge University Press, 2000.
- [9] Srinivasan, S., Petkovic, D., and Ponceleon, D. "Towards Robust Features for Classifying Audio in the Cue Video System," in *Proc. the 7th ACM Multimedia*, 393 - 400, 1999.
- [10] Foote, J. "Content-based Retrieval of Music and Audio," in *Kuo, C. C. J., et al. eds., Multimedia Storage and Archiving Systems II, Proc. SPIE*, Vol. 3229, 138-147, 1997.
- [11] Li, G., and Khokhar, A.A. "Content-based Indexing and Retrieval of Audio Data using Wavelets," in *Proc. IEEE International Conference on Multimedia and Expo*, Vol.2, 885-888, 2000.
- [12] Peltonen, V., Tuomi, J., Klapuri, A.P., Huopaniemi, J., and Sorsa, T. "Computational Auditory Scene Recognition," in *Proc. of IEEE International Conference on Acoustic, Speech and Signal Processing*, Vol.2, 1941-1944, 2002.
- [13] Christensen H., "Speech recognition using heterogeneous information extracting in multi-stream based systems," PHD Thesis, Aalborg University, Denmark, 2002.
- [14] Hegde R.M., Murthy H.A., Gadde V.R.R., "Significance of Joint features derived from the modified group delay functions in speech processing" in *EURASIP Journal on Audio, Speech and Music Processing*, 2007.

- [15] Morgan Nelson and al., "Pushing the envelope aside" in IEEE Signal Processing Magazine, Septembre 2005
- [16] S.G. Mallat, "A theory for multiresolution signal decomposition: the wavelet representation", IEEE Trans. on pattern analysis and machine intelligence, Vol.11, No.7, July 1989.
- [17] M. Vetterli and C. Herley, "Wavelets and Filter banks: theory and design", IEEE Trans. on signal processing, Vol.40, No.9, September 1992.
- [18] G. Tzanetakis and P. Cook, "Music genre classification of audio signals", IEEE Trans. on Speech and audio processing, Vol.10, No.5, July 2002.
- [19] J. Pinquier, "Indexation sonore: recherche de composantes primaires pour une structuration audiovisuelle", PHD, Université de Toulouse III, Décembre 2004.
- [20] A. Zils, « Extraction de descripteurs musicaux : une approche évolutionniste », PHD, Université Paris 6, Septembre 2004.
- [21] I. Daubechies, « Orthonormal bases of compactly supported wavelets », Communications on pure and applied mathematics, 41 :9.9-996, 1988.
- [22] S. G. Mallat, A wavelet tour of signal processing, Academic Press, 1999.
- [23] S. Rossignol, X. Rodet & all, « Features extraction and temporal segmentation of musical signal », In Proc. Int. Computer Music Conf. ICMC, pages 199-202. ICMA, 1998.
- [24] B. Logan, « Mel frequency cepstral coefficients for music modeling », In Proc. Int. Symposium on music information retrieval (ISMIR), 2000.
- [25] E. Skovenberg, S.H. Nielsen, « Evaluation of different loudness models with music and speech material », 117th convention of audio engineering society (AES), San Francisco, October 2004.

Segmentation de Signaux Audio

INTRODUCTION

Le développement de systèmes d'indexation de bases de données audio est un domaine de recherches toujours en évolution et focalise davantage l'intérêt de la communauté scientifique. Ceci est motivé par l'accroissement et la diversité de sources multimédia. La mise en œuvre de ces systèmes nécessite souvent une étape préalable et cruciale de discrimination de classes acoustiques. La segmentation audio consiste à discriminer des classes sonores et à les poursuivre sur un segment temporel donné.

La bande sonore d'un document audiovisuel regroupe plusieurs types de signaux : parole, musique, parole+musique, jingles, bruits, etc. Selon le type d'application ciblé différentes segmentations sont envisageables :

- Segmentation musique /non musique pour la classification en genre ou par type d'instruments de musique [20, 6],
- Séparation parole / fond musicale des segments parole+musique pour des applications de séparation de sources ou de mixage audio [14],
- Segmentation parole /non parole pour la transcription orthographique et de l'indexation audio [11, 16, 14].

Les systèmes de segmentation de l'état de l'art sont généralement basés sur une modélisation des données et font appel aux modèles de Markov cachés (HMM) [17, 15], les modèles de Mélanges de Gaussiennes (GMM), les k plus proches voisins (KNN), les réseaux de neurones et plus récemment les Machines à Vecteurs de Support (SVM) [16, 6, 17]. Dans un contexte de traitement de la parole, une modélisation GMM est généralement adoptée. Bien que ces techniques aient prouvé leur efficacité et donnent de bonnes performances, elles sont néanmoins coûteuses en temps de traitement et en charges de calculs pour des signaux audio issues de bases de données professionnelles et sont souvent tributaires d'une adéquation avec un modèle mathématique difficilement conciliable.

III.1 LA SEGMENTATION AUDIO : ETAT DE L'ART

Le flux audio est généralement l'enregistrement d'un signal acoustique provenant de plusieurs sources sonores. Le document ainsi obtenu est constitué de plusieurs composantes dont les plus fréquentes sont la parole, la musique et le bruit. Cette dernière composante regroupe en fait toutes les composantes non identifiables à de la parole ou de la musique. Ces documents audio sont généralement issues d'émissions radiophoniques ou télévisuelles et constituent le cadre général d'un système d'indexation audio. Cet environnement est caractérisé par une forte variabilité acoustique, on peut ainsi rencontrer de larges segments de parole en bande élargie ou en bande téléphonique, de très brefs segments de jingles et des segments de quelques minutes de musique instrumentale ou de la voix chantée.

Les méthodes de l'état de l'art décomposent ce problème en plusieurs classes : classes binaires parole /non parole, musique / non musique, classe bruit et classe silence.

Un système d'indexation se décompose généralement en deux étapes : une étape d'extraction de descripteurs et une étape de modélisation statistique de ces descripteurs. Dans une perspective de décomposition parole/non parole les descripteurs MFCC (Mel Frequency Cepstral Coefficients) sont généralement adoptés et donnent des résultats satisfaisants. Les méthodes de classification de l'état de l'art sont généralement basées sur modélisation des données telles que les méthodes par mélange de gaussiennes (GMM) ou les chaînes de Markov Cachées (HMM) [1], des méthodes géométriques comme les k plus proches voisins, les histogrammes [19] ou les méthodes neuronales (RNN) et récemment les Méthodes par Vecteurs de Supports (SVM) représentant les méthodes de classification discriminatives [4, 11, 10, 3].

Les méthodes génératives telles que les GMM ont largement fait leur preuve avec succès dans beaucoup d'applications pratiques [18, 2, 21], bien qu'elles soient aujourd'hui surpassées par les méthodes discriminatives en terme de performances de classification. Ceci est principalement due aux caractéristiques des données modernes : la grande dimension et le nombre d'exemples limité. En effet les méthodes génératives sont particulièrement sensibles à ces deux facteurs, alors que les méthodes discriminatives ne le sont généralement pas car opérant dans un espace de grande dimension. Actuellement, les méthodes discriminatives sont donc particulièrement utilisées pour résoudre les problèmes modernes car elles fournissent des résultats quantitativement bons mais au prix d'une faible adaptabilité et de charges de calculs assez lourds [21, 8]. En particulier, les méthodes discriminatives à noyaux, de type SVM, sont très utilisées pour le traitement du signal, de l'image et du traitement de la parole [17, 8] car elles fournissent d'excellents résultats nonobstant les inconvénients cités en amont.

III.2 Une méthode hybride : Analyse en Composantes Principales et le critère d'alignement

Cet algorithme est basé sur une analyse du signal audio en composantes principales afin d'estimer la direction de variance maximale des données et d'en extraire les structures principales. Le critère d'alignement permet de choisir un hyperplan optimal permettant la discrimination des classes sonores. Avant de décrire plus explicitement cet algorithme, nous pensons qu'il est plus pédagogique de résumer les notions essentielles du critère d'alignement développé et exploité essentiellement dans les méthodes à noyaux [5, 9].

III.2.1 Le critère d'alignement

Soit $D = \{(x_1, y_1), (x_2, y_2), \dots, (x_m, y_m)\}$ un ensemble de m données d'apprentissage et leurs étiquettes associés. L'apprentissage statistique inductif permet d'exploiter la relation existante entre la distribution géométrique des données et celle des étiquettes de classes. Dans ce cadre, il est aussi démontré que la matrice de Gram K ou matrice noyau, symétrique semi-définie positive permet de caractériser les ressemblances (ou dissemblances) entre les paires de données d'entrée en rapport avec la métrique induite par une fonction noyau [5, 9, 22]. Chaque élément k_{ij} est le résultat de l'application d'une fonction noyau à la paire de donnée (x_i, x_j) . D'autre part, si nous considérons que les étiquettes de classes notées (± 1) sont contenues dans un vecteur colonne y , nous pouvons définir la matrice $Y = y \times y^T$. Y est une matrice symétrique de même dimension que la matrice K telle que :
 $y_{ij} = +1$ si $y_i = y_j$ et $y_{ij} = -1$ si $y_i \neq y_j$. Ainsi Y résume toute l'information véhiculée par les données d'apprentissage représentés par leurs étiquettes. Puisque K représente la similarité géométrique des données et Y représente la similarité des étiquettes de ces données, il est raisonnable de penser que ces deux matrices présentent aussi des ressemblances entre elles. Cette propriété est alors mesurée et quantifiée par le critère d'alignement. Celui-ci n'est en fait qu'une généralisation aux matrices, du produit scalaire normalisé de deux vecteurs (x, y) :

$$\cos(\varphi) = \frac{x^T y}{\sqrt{x^T x * y^T y}} \quad (\text{III.1})$$

Dans le cas de deux matrices A et B cette généralisation traduit le critère d'alignement qui s'écrit :

$$A(A, B) = \frac{\langle A, B \rangle_F}{\sqrt{\langle A, A \rangle_F * \langle B, B \rangle_F}} \quad (\text{III.2})$$

Le terme $\langle A, B \rangle_F$ signifie le produit scalaire généralisé aux matrices et l'indice F traduit la norme de Frobenius de la matrice A . Ainsi la formule précédente de l'alignement se réécrit comme :

$$A(A, B) = \frac{\langle A, B \rangle_F}{(\|A\|_F * \|B\|_F)^{\frac{1}{2}}} \quad (\text{III.3})$$

De ce fait pour mesurer l'alignement entre la matrice des données K et la matrice Y des données d'apprentissage étiquetés, nous évaluons la quantité $A_{k,d}(Y, K)$ telle que définie par l'expression (III.2) ou (III.3).

$$A_{k,d}(Y, K) = \frac{\langle Y, K \rangle_F}{(\|Y\|_F * \|K\|_F)^{\frac{1}{2}}} \quad (\text{III.4})$$

III.2.2 L'algorithme de segmentation ACP_Alignement

Etant donné $D = \{(x_1, y_1), (x_2, y_2), \dots, (x_m, y_m)\}$ un ensemble constitué de m données d'apprentissage étiquetées (exemple d'un segment de signal constitué de deux classes sonores : parole + musique). Il s'agit d'évaluer la matrice de covariance des données puis d'en extraire les valeurs propres et vecteurs propres. On considère alors seulement le premier vecteur propre correspondant à la plus grande valeur propre. On détermine par la suite l'hyperplan optimal perpendiculaire à la direction principale donnée par le premier vecteur propre en ayant recours au critère d'alignement entre la matrice des étiquettes de données d'apprentissage et celle des étiquettes estimées par la fonction de décision. La maximisation de cet alignement permet d'évaluer le paramètre optimal de l'hyperplan puis d'en déduire ainsi la valeur du vecteur étiquette des données test et par conséquent la classe recherchée de données inconnues analysés. Suite à l'étape de paramétrisation, nous obtenons le signal des descripteurs suivant : $X = \{x_1^d, x_2^d, \dots, x_m^d\}$ l'ensemble de données exemples de d paramètres (MFCC) avec $d \ll m \ll n$ taille du signal test.

Etape 1 : Calculer la matrice de Covariance des données

$$M_X = \text{Cov}(X) \quad (\text{III.5})$$

Etape 2 : Evaluer les valeurs propres et vecteurs propres associés : λ_i et V_i $i \in \{1, \dots, d\}$

Etape 3 : Déterminer le premier vecteur propre correspondant à la plus grande valeur propre : $V_1 = \max(V_i)$ (III.6)

Etape 4 : Déterminer Hyperplan Optimal H

$$x \in H \text{ si } x.V_1 - \alpha_{opt} = 0 \quad (\text{III.7})$$

$$f(x) = y_{est} = \text{sign}(x.V_1 - \alpha) \quad (\text{III.8})$$

Faire varier le paramètre α revient à glisser l'hyperplan H le long de la direction de variance maximale des données (Figure III.1). Soient y_{app} le vecteur étiquette de données d'apprentissage et y_{test} le vecteur étiquette estimée de la donnée d'apprentissage :

Pour $\alpha = \alpha_1 \rightarrow \alpha_2$

Calculer $y_{est} = \text{sign}(x.V_1 - \alpha) \forall x \in D$

Evaluer, $Y_{est} = y_{est} * y_{est}^t$ et $Y = y_{app} * y_{app}^t$

$$\text{Calculer } A_\alpha(Y_{est}, Y) = \frac{\langle Y_{est}, Y \rangle_F}{\sqrt{(\|Y_{est}\|_F * \|Y\|_F)}} \quad (\text{III.9})$$

Fin Pour

Evaluer $A_{\max}(\alpha) \rightarrow \alpha_{\text{opt}}$ (le max. de A en fct. de α)

Soit x_{test} une donnée appartenant à l'ensemble de données test (autre fichier que celui ayant servi pour l'entraînement) on détermine alors son étiquette y_{test} en ayant recours simplement à la fonction de décision $y_{\text{test}} = \text{sign}(x_{\text{test}} \cdot V_1 - \alpha_{\text{opt}})$ (III.10)

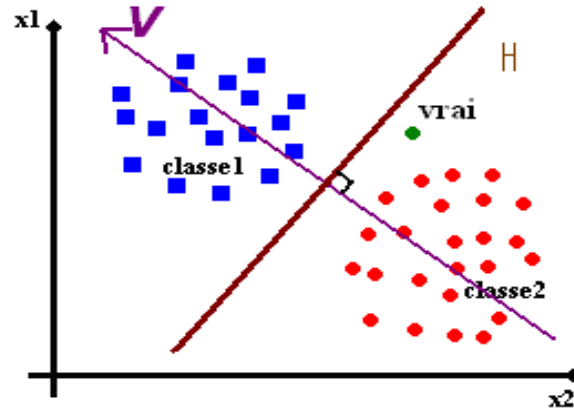


Figure III.1. Exemple de données linéairement séparables

III.2.3 Cas de données linéairement séparables

Dans ce cas (Figure III.1), nous avons une marge où l'alignement est maximum, et égale à la valeur maximale qui vaut 1. C'est-à-dire que le pourcentage des données classées faux est nul. La largeur de cette marge égale à $(\alpha_2 - \alpha_1)$. Elle représente l'équivalent de la marge maximale recherchée par les SVMs et $(H_1) : (x \cdot V - \alpha_1)$, $(H_2) : (x \cdot V - \alpha_2)$ les deux hyperplans canoniques respectivement. α_{opt} correspond au milieu de la marge de séparation $\alpha_{\text{opt}} = (\alpha_2 + \alpha_1)/2$, d'où $(H_{\text{opt}}) = (x \cdot V - \alpha_{\text{opt}})$ est l'hyperplan optimal recherché (Figure III.2).

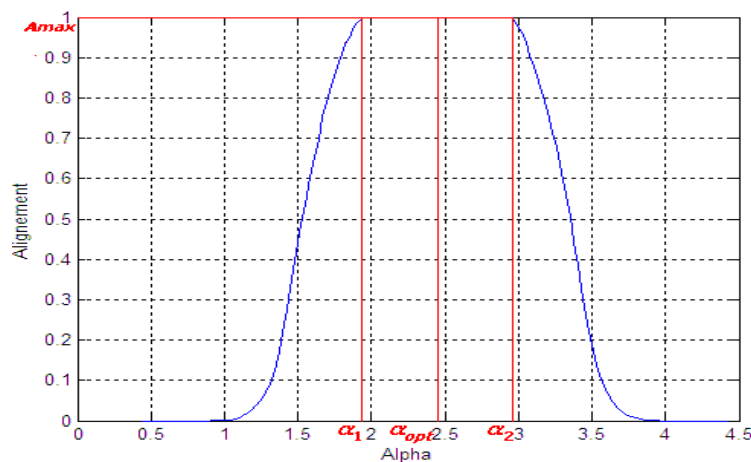


Figure III.2. Le paramètre α en fonction de l'alignement dans un cas séparable

III.2.4 Discussions sur le choix de α_1 et α_2

Soient X_1 et X_2 deux sous ensembles de X correspondants (par exemple) respectivement à la classe parole et musique. Soient $\alpha_1 = V_1.M_1$ et $\alpha_2 = V_1.M_2$ avec M_1 et M_2 resp. les centres d'inertie des nuages de points X_1 et X_2 . Ainsi, le principe de notre algorithme ressemble à la philosophie des SVM dans le sens où on utilise un ensemble restreint de données exemples $d \ll m \ll n$ (la matrice de covariance ne dépend que la taille des descripteurs et non de leurs nombre) afin d'estimer l'hyperplan optimal qui constitue la frontière de décision adéquate permettant de classer tout nouvel élément appartenant à l'une ou l'autre des classes discriminées. Cependant les charges de calcul de notre méthode sont nettement moindres que ceux des méthodes SVM et de par sa simplicité notre algorithme est beaucoup plus rapide permettant son exploitation dans un processus de décision en ligne.

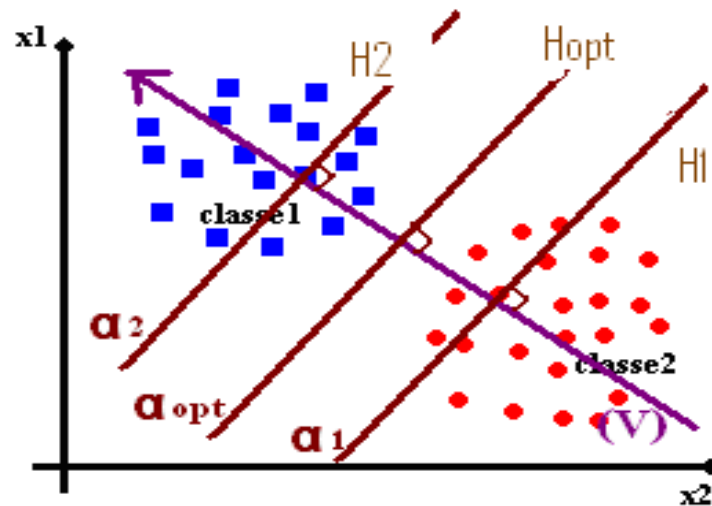


Figure III.3. L'influence du paramètre α sur l'hyperplan séparateur.

III.2.5 Cas de données linéairement non séparables

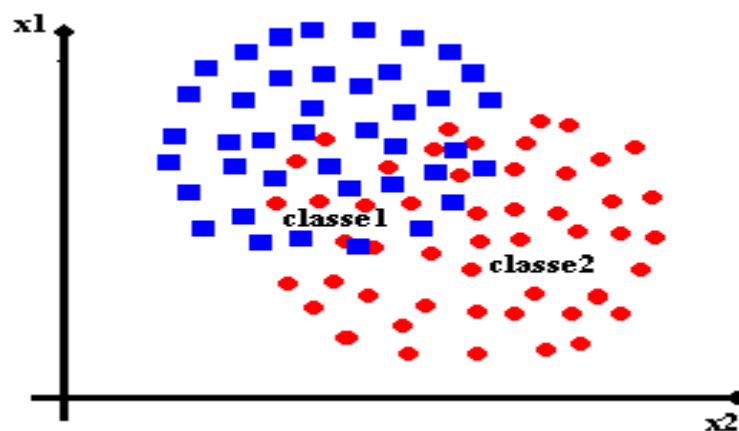


Figure III.4. Exemple de données linéairement non séparables

Dans le cas où le nuage de points n'est pas séparable et où les données de deux classes distinctes sont équitablement distribuées (Figure III.4), il y aura recouvrement entre les deux nuages ce qui entraîne des erreurs de classification. Dans ce cas, nous n'avons qu'une valeur où l'alignement est maximum, et égale à une valeur inférieure à 1, c'est-à-dire qu'il y a des données qui ont été mal classées, et le taux de recouvrement entre les deux classes égal à $1 - A_{\max}$ (dans l'exemple précédent, il est de 32%). Cette valeur α_{opt} est optimale, d'où (H_{opt}) est l'hyperplan de classification optimal telle que $f(x) = \text{sign}(x.V_1 - \alpha_{opt})$ (Figure III.5).

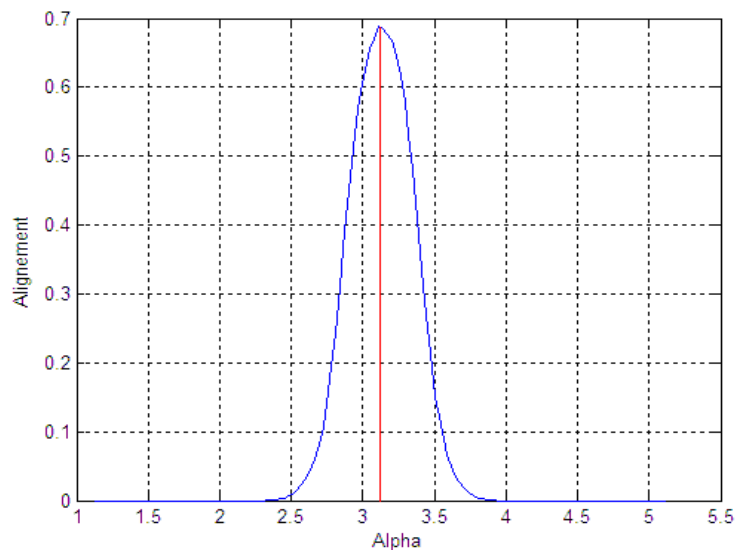


Figure III.5. Le paramètre α en fonction de l'alignement dans un cas non séparable.

Conclusions

Nous venons de présenter dans ce chapitre une méthode originale de segmentation sonore permettant une indexation audio en classes multiples (parole, musique, bruit et silence). Cette méthode de classification semi supervisée fait appel à une méthode combinant l'analyse en composantes principales et le critère d'alignement de noyaux introduit et souvent exploité dans l'apprentissage statistique [5,22]. Cette contribution constitue une des rares applications de ce critère en traitement de signaux audio [7].

Nous montrons en chapitre IV la faisabilité de cette méthode et nous évaluerons ces performances sur une base de signaux sonores issues d'enregistrements radiodiffusés dont nous avons varié les durées et le contenu. Son principal avantage par rapport aux méthodes génératives et discriminatives (GMM, HMM et SVM par exemple) est son indépendance vis-à-vis de la modélisation des classes et sa faible charge de calculs qui traduit par conséquent une rapidité de mise en œuvre puis de décision, ce qui permet d'envisager son implantation sur des processeurs DSP ou FPGA et par conséquent son exploitation dans un processus de décision en ligne. Un travail en perspective concerne l'adaptation de la méthode avec les descripteurs acoustiques autres que les MFCC ainsi que sa robustesse par rapport au recouvrement temporel de classes sonores

Références Bibliographiques du Chapitre

- [1] Duda, R., Hart, P. E., et Stork, *Pattern Classification (2nd edition)*. John Wiley & Sons, 2000
- [2] Ajmera J. , *Robust Audio Segmentation* in PhD Thesis, Ecole Polytechnique Fédérale de Lausanne (EPFL), 2004
- [3] Bishop M. C., *Pattern Recognition and Machine Learning* Springer, Jordan M., Kleinberg J., Schölkopf B. Eds. 2006
- [4] Burges C., *“A tutorial on Support Vector Machines for pattern recognition”* Data Mining and Knowledge Discovery 2 (2). 1998
- [5] Cristiannini N., Shaw-Taylor J., Elisseeff A. and Kandola J. , “ *On kernel -target alignment* ” Advances in Neural Informations Processing Systems, Vol.14, pp. 367-373., 2002
- [6] Essid S., *“ Classification automatique de signaux audio-fréquence : reconnaissance des instruments de musique ”*, PhD Thesis, Université Pierre et Marie Curie , 2005
- [7] **Fergani L.**, Fergani B., Houacine A., *“Segmentation de signaux audio : Une nouvelle approche utilisant le critère d’alignement”* in Canadian Acoustics, Vol 38 N°2, pp. 3-10, 2010
- [8] Herbrich R., *“Learning Kernel Classifiers: Theory and Algorithms”* MIT Press, Cambridge USA, 2002
- [9] Kandola J., Shaw-Taylor J. and Cristianini N., (2002) *“On the extensions of kernel alignment”*, Dept. of Computer Science, University of London, Tech. Rep. 120, 2002
- [10] Llorente Godino J. I. and Co, *“Discriminative methods for the detection of voice disorders”* NoLisp, April 2005.
- [11] Lu L, Li S. and Zhang H.-J., *“Content-based audio segmentation using Support Vector Machines”* ACM Multimedia Conference, Canada, 2001
- [12] Lu L. and Zhang H.-J., *“Content analysis for audio classification and segmentation”* IEEE Trans. SAP 10(7):504-516, 2002
- [13] Mauclair Julie, *“Mesures de confiance en traitement automatique de la parole et applications”*, PhD Thesis Université du Maine, Laboratoire d’Informatique de l’Université du Maine, 2006
- [14] Meignier S., Moraru D., Fredouille C., Besacier L., Bonastre J.F. (2004) *“ Benefits of prior acoustic segmentation for automatic speaker segmentation”* IEEE ICASSP 2004, Montreal Canada, 2004
- [15] O’Shaughenssy D., *“Automatic Speech Recognition: History, Methods and Challenges”* Pattern Recognition , Elsevier, 2008
- [16] Pinquier J., “ *Indexation Sonore: recherche de composantes primaires pour une structuration audiovisuelle* ”, PhD Thesis, Université Paul Sabatier , Toulouse III, 2004
- [17] Rabiner L., *“A tutorial on hidden markov models and selected applications in speech recognition”* Proc. IEEE 77(2): 257-286, 1989

- [18] Seck M., Magrin-Chagnolleau I. and Bimbot F., *“Experiments on speech tracking in audio documents using GMM”* IEEE ICASSP USA, 2001
- [19] Theodoridis S. and Koutroumbas K. , *“Pattern Recognition, 2nd Ed.”*, Elsevier, 2003
- [20] Tzanetakis G. and Cook P. , *“Musical Genre Classification of Audio Signals”* IEEE Transactions on Speech and Audio Processing, 10(5):293-302, 2002
- [21] Ulusay I., Bishop M. C., *“Comparison of Generative and Discriminative Techniques for Object Detection and Classification”* LNCS4170 Springer-Verlag, 2006
- [22] Vert J.P., Tsuda K., Schölkopf B. , *“ A primer on kernel methods”* Kernel Methods in Computational Biology, Schölkopf B, Tsuda K and Vert J.P. Eds. Cambridge MA: the MIT Press, pp.35-70, 2004
- [23] Young S. and all , *“The HTK Book”* Ver.3.2., 2002

Séparation de Sources Sonores

INTRODUCTION

La majorité des signaux rencontrés dans les diverses applications envisagées par le traitement du signal sont en fait des mélanges auxquels contribuent plusieurs sources. Dans le cas des signaux audio, l'oreille humaine effectue une écoute sélective et retrouve chacune des sources originales. Différentes méthodes mathématiques et numériques visent à faire imiter ce type de comportement par un système artificiel. Elles sont regroupées sous l'appellation Séparation Aveugle de Sources (S.A.S.) : le terme aveugle indique qu'aucune information sur les sources n'est disponible [1,2]. Les domaines des télécommunications et du traitement de signaux audio ou biomédicaux regorgent d'applications potentielles à ces méthodes. Nous proposons dans le présent chapitre de séparer des sources audio constituées de mélanges linéaires de musique et de parole, par deux méthodes différentes : la Décomposition Empirique Modale (DEM) et l'Analyse en Composantes Indépendantes (ACI).

Dans les années 90, de nombreux algorithmes ont été développés pour résoudre le problème de Séparation Aveugle de Sources par Analyse en Composantes Indépendantes [3]. Même si elles reposent toutes sur l'hypothèse d'indépendance statistique des sources, deux approches concurrentes donnèrent lieu à une multitude d'algorithmes : l'approche algébrique et l'approche neuronale. Les algorithmes proposés par chacune des deux approches se scindent toujours en deux entités : fonction de contraste d'une part et algorithme d'optimisation d'autre part. Leurs performances diffèrent par leur stabilité, leur vitesse de convergence et leur besoin en mémoire. Citons parmi les plus connus l'algorithme de Jutten et Herault [21], l'algorithme JADE de Cardoso [12], l'algorithme SOBI de Belouchrani [11] et enfin l'algorithme FASTICA de Hyvarinen [6] que nous utiliserons.

Une nouvelle approche de Séparation Aveugle de Sources a été proposée par Aissa El Bey & all [9,10]; Elle se base sur une méthode introduite par Huang [13] et popularisée par P. Flandrin [8]: la Décomposition Empirique Modale (E.M.D pour Empirical Mode Decomposition)). L'E.M.D. repose sur une décomposition adaptée qui décrit le signal localement comme une succession de contributions d'oscillations rapides (hautes fréquences) et d'oscillations plus lentes (basses fréquences). L'E.M.D. permet donc la décomposition du signal en contributions de base, appelées Modes Empiriques ou IMF (Intrinsic Mode Function). Les oscillations locales considérées ne sont pas nécessairement harmoniques au sens de Fourier. L'E.M.D. est bien adaptée à l'étude des signaux non-stationnaires et/ou générés par des systèmes non linéaires tels que les mélanges de signaux musicaux et de parole. La séparation de Sources se fait en deux étapes : Une première étape d'Analyse est réalisée grâce à l'application de l'E.M.D. à chaque mélange pris séparément. Une deuxième étape, complémentaire de la première, permet de regrouper toutes les composantes modales qui correspondent au même signal source afin de le reconstituer. Cette tâche sera accomplie par une méthode de classification non supervisée tels que les K-Moyennes.

Nous proposons de comparer les deux méthodes de séparation des sources citées ci-dessus pour la séparation de mélanges audio bruités. Alors que souvent, les qualités de séparation sont évaluées de façon subjective en observant les signaux restitués, nous

proposons d'évaluer les qualités de séparation et de restitution de chacune des méthodes par des critères objectifs. Il s'agit des critères DSR (Distorsion to Signal Ratio) et ISR (Interference to Signal Ratio). Le critère DSR représente la distorsion relative totale entre le signal source et le signal restitué alors que le critère ISR représente la distorsion relative introduite par les interférences dues aux autres sources sur le signal restitué.

Ce chapitre sera organisé comme suit : En premier lieu, la Séparation de Sources par Analyse en Composantes Indépendantes sera détaillée. Nous montrerons ensuite comment réaliser cette même séparation de sources à partir de la Décomposition Empirique Modale. Enfin nous expliciterons les critères d'évaluation proposés : DSR et ISR. Une conclusion clôturera notre exposé.

IV.1 ANALYSE EN COMPOSANTES INDEPENDANTES (A.C.I.)

L'Analyse en Composantes Indépendantes consiste à rechercher une décomposition d'un vecteur aléatoire en composantes statistiquement indépendantes. Elle permet donc la séparation de sources sous l'hypothèse d'indépendance statistique de ces dernières. Il existe un certain nombre d'algorithmes d'estimation du modèle ACI. Ces différentes versions varient selon la fonction de contraste choisie et le schéma adopté pour son optimisation. Leurs performances diffèrent par leur stabilité, leur vitesse de convergence et leur besoin en mémoire [3]. Citons parmi les plus populaires dans les applications, l'algorithme JADE, l'algorithme FAsTICA, ainsi que l'algorithme SOBI [6, 7, 11,12].

VI.1.1 Formulation Générale

L'analyse en composantes indépendantes consiste à rechercher une décomposition d'un vecteur aléatoire en composantes statistiquement indépendantes. Souvent, à titre d'illustration, le problème de la « cocktail party » est cité comme exemple. Deux personnes ou plus parlent simultanément dans une pièce, et leurs propos sont recueillis par deux microphones situés à deux endroits différents. Chacun des deux signaux enregistrés est un mélange des deux signaux originaux émis par les deux personnes. Le problème de la « cocktail party » consiste à estimer ces deux signaux originaux à partir des signaux enregistrés. Dans ce contexte, les signaux observés, $\mathbf{x}(\mathbf{t})$, sont considérés comme un mélange inconnu de signaux dits sources, $\mathbf{s}(\mathbf{t})$, supposés indépendants et non observés. Le cas le plus simple et le plus souvent traité définit le mélange comme une combinaison linéaire et considère un mélange spatial instantané, c'est-à-dire pour lequel la dimension temporelle est ignorée. Le mélange est décrit par une matrice $\mathbf{A}$ appelée matrice de mélange. L'objectif est d'estimer conjointement la matrice $\mathbf{A}$ et le vecteur $\mathbf{s}$, soit de séparer les sources en déterminant une matrice de séparation, $\mathbf{W}$ (voir Figure IV.1).

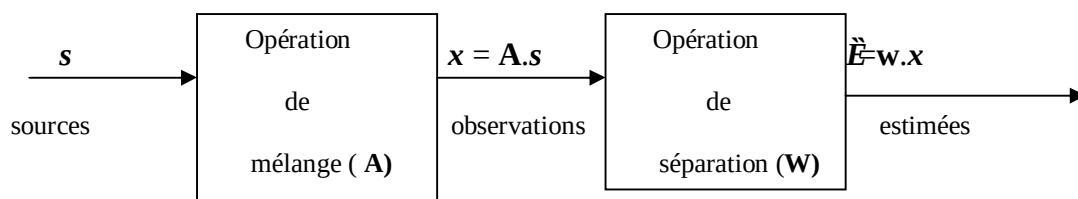


Figure IV.1: Modèle général d'un système de séparation des sources.

Du point de vue statistique, la formulation générale s'écrit de la façon suivante [20, 21, 17]:

L'analyse en composantes indépendantes d'un vecteur m -dimensionnel $\mathbf{x} = (x_1, \dots, x_m)^T$ consiste à estimer le modèle :

$$\mathbf{X} = \mathbf{A}\mathbf{s} \quad (\text{IV.1})$$

Où les variables s_i du vecteur $\mathbf{s} = (s_1, \dots, s_n)^T$ sont supposées indépendantes et où $\mathbf{A}$ est une matrice $m \times n$ de mélange.

IV.1.2 Conditions d'estimation et indéterminations

Le modèle (IV.1) ne peut être estimé que sous certaines conditions et à certaines indéterminations près [17]. Une restriction fondamentale à l'estimation du modèle est qu'au plus une composante soit gaussienne. De plus, la matrice $\mathbf{A}$ doit être de rang plein, et le nombre d'observations supérieur au nombre des sources à estimer, i.e. $m > n$ (en pratique m et n sont souvent pris égaux). Les matrices $\mathbf{A}$ et $\mathbf{W}$ sont dans ce cas inverse l'une de l'autre. Lorsque $\mathbf{s}$ et $\mathbf{A}$ sont parfaitement inconnus, leur identification n'est possible qu'à un facteur multiplicatif près ainsi qu'à une permutation près. Les composantes indépendantes ne seront donc pas ordonnées.

IV.1.3) Mesures d'indépendance

A- Indépendance statistique

L'hypothèse d'indépendance statistique mutuelle des sources est une hypothèse statistiquement forte. Elle stipule en effet, que la densité de probabilité conjointe des sources, p , est le produit des densités de probabilité marginales p_i :

$$p(s_1, \dots, s_n) = \prod_{i=1}^n p_i(s_i) \quad (\text{IV.2})$$

L'analyse ne garantit pas la détermination de sources indépendantes, mais plutôt des sources aussi indépendantes que possible. Le vecteur aléatoire des sources estimées sera noté $\mathbf{y}$. On a : $\mathbf{Y} = \mathbf{W}\mathbf{x}$ et les composantes y_i aussi indépendantes que possible au sens de la maximisation d'une fonction $\bar{u}(y_1, \dots, y_n)$ mesurant l'indépendance. Il est donc nécessaire de définir cette fonction au sens de laquelle l'analyse deviendra optimale. Diverses mesures, encore appelées fonctions de contrastes, ont été proposées [3]. Parmi les familles de fonctions contraste, l'une s'appuie sur une notion essentielle en ACI : la non-gaussianité.

B- Non-gaussianité

Le théorème central-limite stipule que « la somme de variables aléatoires indépendantes de même loi tend vers une variable aléatoire gaussienne ». Intuitivement, cela sous-entend que mélanger des variables les « gaussianise », d'où l'idée de maximiser une mesure de non-gaussianité pour les séparer. La non-gaussianité $\bar{u}^G(\mathbf{y})$ du vecteur aléatoire $\mathbf{y}$ est l'écart à la gaussianité de ce vecteur. Elle est exprimée par la divergence de Kullback-Leibler de P_y à la loi normale de mêmes moments d'ordre 1 et 2 que $\mathbf{y}$:

$$\bar{\mathbf{u}}^G(\mathbf{y}) = K(P_{\mathbf{y}} \ln N(E[\mathbf{y}], \text{Cov}[\mathbf{y}])) \quad (\text{IV.3})$$

Cette quantité est désignée sous le terme de « néguentropie », également exprimée par :

$$J(\mathbf{y}) = H(\mathbf{y}_{\text{gauss}}) - H(\mathbf{y}) \quad (\text{IV.4})$$

Où H désigne l'entropie différentielle. Pour un vecteur aléatoire $\mathbf{y}$ de densité $P_{\mathbf{y}}$, elle est donnée par :

$$H(\mathbf{y}) = - \int P_{\mathbf{y}}(\mathbf{y}) \log(P_{\mathbf{y}}(\mathbf{y})) d\mathbf{y} \quad (\text{IV.5})$$

L'entropie exprime le degré d'information portée par une variable. Plus cette variable est aléatoire, plus son entropie est élevée. Une variable gaussienne a l'entropie la plus élevée parmi toutes les variables de même variance. La néguentropie est donc une quantité positive, nulle si et seulement si $\mathbf{y}$ suit une loi gaussienne. La néguentropie est une mesure adéquate de non-gaussianité mais sa maximisation est difficile. Elle n'est pas optimisée de façon directe mais approchée par des fonctions mettant en jeu des statistiques d'ordre supérieur. Une estimation de la néguentropie classiquement utilisée est le kurtosis d'ordre 4 :

$$\text{Kurt}(\mathbf{y}) = E[\mathbf{y}^4] - 3E^2[\mathbf{y}^2] \quad (\text{IV.6})$$

Ou plus généralement des fonctions de la forme :

$$J_G(\mathbf{y}) = (E[G(\mathbf{y})])^2 - E[G(\mathbf{z})]^2 \quad (\text{IV.7})$$

Où G est une fonction non quadratique deux fois différentiable, et $\mathbf{z}$ une variable aléatoire gaussienne centrée réduite. Le choix de G conduit à des estimateurs plus ou moins robustes [7].

Il existe un certain nombre d'algorithmes d'estimation du modèle ACI. Ces algorithmes varient selon la fonction de contraste choisie et le schéma adopté pour son optimisation. Leurs performances diffèrent par leur stabilité, leur vitesse de convergence et leur besoin en mémoire. Citons parmi les plus populaires, dans les applications, l'algorithme « JADE » [13] et l'algorithme FastICA [5, 6]. Nous exposerons dans ce qui suit l'algorithme FastICA.

A- de Blanchiment des données

Un prétraitement est généralement appliqué avant d'estimer effectivement les composantes indépendantes. Il consiste à centrer et blanchir les données telle sorte que leur moyenne soit nulle et leur covariance l'identité. Ceci est classiquement réalisé au moyen d'une analyse en composantes principales. Plutôt que d'avoir à estimer une matrice de mélange $\mathbf{A}$ quelconque, ce prétraitement réduit le problème à l'estimation d'une matrice orthogonale. La complexité s'en trouve diminuée et la convergence accélérée. Le blanchiment des données revient à annuler les statistiques d'ordre deux (variance). Il consiste à transformer linéairement le vecteur $\mathbf{x}$ de façon à décorréliser ses composantes et à leur assigner une variance unité. Le nouveau vecteur $\mathbf{x}$ sera alors dit blanc et sa matrice de covariance sera la matrice unité $\mathbf{I}$.

$$\text{Soit :} \quad E[\mathbf{x}\mathbf{x}^T] = \mathbf{I} \quad (\text{IV.8})$$

Il faut donc déterminer une matrice $\mathbf{B}$ dite de blanchiment, tel que $\mathbf{x} = \mathbf{B}\mathbf{x}$ vérifie la relation précédente. A partir de la décomposition en valeurs propres de la matrice de covariance de $\mathbf{x}$, nous pouvons déterminer la matrice de blanchiment $\mathbf{B}$.

Soit alors :

$$\mathbf{\Sigma}_{xx} = \mathbf{U} \cdot \mathbf{D} \cdot \mathbf{U}^T$$

Avec : $\mathbf{D}$ la matrice diagonale contenant les valeurs propres de $\mathbf{\Sigma}_{xx}$.

et $\mathbf{U}$ est la matrice orthogonale des valeurs propres associées.

La matrice de blanchiment $\mathbf{B}$ est définie alors par : $\mathbf{B} = \mathbf{D}^{-1/2} \cdot \mathbf{U}^T$ (IV.9)

Dorénavant, nous considérerons les données centrées blanchies.

B- Algorithme FastICA

Une première version de l'algorithme FastICA consiste à estimer les composantes indépendantes une à une. Chaque composante $y_i = \mathbf{w}^t \mathbf{x}$ où $\mathbf{w}^t$ est un vecteur ligne de la matrice de séparation $\mathbf{W}$ et $i = 1, \dots, n$, est obtenue en maximisant un critère de non-gaussianité. Le critère utilisée est la néguentropie, $J(\mathbf{w}^t \mathbf{x})$, mesurée par l'approximation donnée en (IV.7). La procédure d'optimisation s'appuie sur la méthode du point fixe et produit l'algorithme décrit ci-dessous :

- 1- Choisir un $\mathbf{w}$ initial , aléatoirement, de norme unité
- 2- $\mathbf{w} \leftarrow E\{ \mathbf{x}g(\mathbf{w}^t \mathbf{x}) \} / g'(\mathbf{w}^t \mathbf{x})\mathbf{w}$ où g et g' désignent les dérivées première et seconde de la fonction non-quadratique G de l'équation (IV.7).
- 3- $\mathbf{w} \leftarrow \mathbf{w} / \|\mathbf{w}\|$
- 4- Reprendre en 2 jusqu'à convergence.

L'estimation de plusieurs composantes à la fois peut être réalisée de façon séquentielle ou parallèle.

- Approche par déflation : dans le premier cas, l'algorithme décrit ci-dessus est exécuté autant de fois que l'on veut estimer de composantes. Pour éviter d'estimer deux fois la même composante, les sorties $y_i, i = 1, \dots, n$, sont décorrélées à l'issue de chaque itération. Pour des données blanchies $\mathbf{x}$, cette décorrélation équivaut à une orthogonalisation réalisée, par exemple, par la méthode de Gram-Schmidt.
- Approche symétrique : dans le second cas, les composantes sont estimées parallèlement en s'appuyant sur un schéma d'orthogonalisation symétrique. Si m est le nombre de composantes à estimer, chacune des m composantes est estimée par l'algorithme estimant une composante. Une matrice $\mathbf{W}$ est obtenue puis orthogonalisée. Le processus est itéré jusqu'à converger. Cette version a l'avantage de ne privilégier aucune composante sur les autres.

IV.2 DECOMPOSITION EMPIRIQUE MODALE (D.E.M.)

La DEM est une décomposition adaptée qui décrit le signal localement comme une succession de contributions d'oscillations rapides (hautes fréquences) et d'oscillations plus lentes (basses fréquences). Conceptuellement, la mise en place de la méthode permettant de décrire cette décomposition est assez simple. Il suffit de considérer le signal à l'échelle de ses oscillations locales, de soustraire l'oscillation la plus rapide du signal et de réitérer le processus sur le résidu, de manière à écrire le signal $x(t)$ comme une combinaison finie d'oscillations [7]. Cette technique permet donc la décomposition du signal en différents modes en supposant qu'il est localement périodique.

IV.2.1 Principe de la méthode

L'É.M.D. décrit le signal $x(t)$ entre deux extrema consécutifs (par exemple, deux minima situés aux temps t_1 et $t_1 + \Delta t$). Il est alors défini une contribution « hautes fréquences » locale $\{d_1[x](t), t_1 < t < t_1 + \Delta t\}$ ou détail local, ainsi qu'une contribution « basses fréquences » locale correspondante $m_1[x](t)$ ou tendance locale. Le signal $x(t)$ peut alors s'écrire :

$$x(t) = m_k[x](t) + d_1[x](t) \quad \text{pour} \quad t_1 \leq t \leq t_1 + \Delta t \quad (\text{IV.10})$$

Ce processus est réitéré pour l'ensemble des oscillations constituant le signal, la procédure est alors appliquée au résidu $m_1[x](t)$ formé par l'ensemble des tendances locales et considéré comme un nouveau signal, conduisant à un nouveau détail $d_2[x](t)$ et à un nouveau résidu $m_2[x](t)$. La décomposition obtenue est composée de différents modes (ou IMF, pour « Intrinsic Mode Functions ») $d_k[x](t)$. Ces derniers sont extraits itérativement et produisent une représentation de type :

$$x(t) = m_k[x](t) + \sum_{k=1}^K d_k[x](t) \quad (\text{IV.11})$$

Où K est la profondeur de décomposition.

Le signal $x(t)$ peut également s'écrire :

$$x(t) = \sum_{j=0}^K \text{IMF}_j(t) + r(t) \quad \text{avec} \quad K \in \mathbb{N}^* \quad (\text{IV.12})$$

Où : $\text{IMF}_j(t)$ est le $j^{\text{ème}}$ mode, $r(t)$ le résidu de la décomposition et K le nombre d'IMF (ou modes). La méthode est par construction, une décomposition complète, K étant fini. En effet, l'addition de toutes les IMF et du résidu permet une reconstruction linéaire du signal original, sans pertes ou déformation de l'information initiale (du moins en théorie). Un IMF est une fonction qui doit remplir les conditions suivantes :

- Avoir une moyenne nulle.
- Être Oscillatoire.
- Entre un minimum et un maximum successif, un IMF passe par zéro.

IV.2.2 Algorithme DEM

L'analyse du signal $x(t)$ par DEM suit les étapes suivantes :

- 1- Identification de tous les extrema de $x(t)$.
- 2- Interpolation entre les minima (resp. maxima du signal), conduisant à une enveloppe $e_{\min}(t)$ (resp. $e_{\max}(t)$).
- 3- Calcul de la moyenne $m(t) = (e_{\min}(t) + e_{\max}(t))/2$.
- 4- Extraction du détail $d(t) = x(t) - m(t)$.
- 5- Itération sur le résidu $m(t)$ jusqu'à l'obtention d'un résidu d'énergie quasi nulle.

En pratique, la procédure précédente doit être raffinée par un processus de tamisage : les étapes de 1 à 4 sont répétées sur le signal $d(t)$ jusqu'à ce qu'il puisse être considéré comme de moyenne nulle suivant un critère d'arrêt à définir. A ce moment, le détail est dit Mode Intrinsic (IMF). Le résidu correspondant est calculé et l'étape 5 est mise en œuvre.

Par construction, le nombre d'extrema diminue d'un résidu au suivant ; La décomposition se termine donc avec un nombre fini de modes. Les modes et résidus ainsi que les notions de « hautes » ou « basses » fréquences sont introduits mais n'induisent qu'une interprétation locale ; Le processus doit être considéré comme un filtrage adaptatif dépendant du signal et non pas comme un filtrage en sous bande du type transformée en ondelettes par exemple.

IV.2.3 Interpolation, Echantillonnage et effets de bord

La version de base est conceptuellement simple, mais un certain nombre de degrés de liberté (nature des interpolations, gestion des effets de bord, critère d'arrêt) sont laissés à l'appréciation de l'utilisateur. Cependant, un certain nombre d'auteurs a proposé plusieurs variations permettant de mieux contrôler l'algorithme [Flandrin]. L'interpolation des minima et maxima est une chose très importante dans la mesure où elle conditionne la forme des IMF. Dans la littérature l'interpolation par des splines cubiques est la plus usitée. La raison de leur succès est principalement liée au compromis entre la qualité de l'interpolation et la simplicité de calcul. Mais d'autres fonctions d'interpolation pourraient être utilisées. Les extrema doivent être correctement localisés, ceci nécessite un sur-échantillonnage du signal. Une solution aux effets de bord, dus à la durée finie du signal serait de les périodiser par effet miroir.

IV.2.4 Critère d'arrêt

Comme mentionné précédemment, l'extraction d'un mode est conditionnée par l'arrêt du processus de tamisage. Deux conditions doivent alors être respectées : en premier lieu, le nombre d'extrema et le nombre de passages par zéro doivent différer au moins d'une unité ; Et en second lieu les enveloppes supérieures et inférieures doivent tendre vers zéro suivant un critère à préciser. En pratique, il faudrait comparer entre l'amplitude de la moyenne et le

mode correspondant. Un seuil trop bas conduirait à une sur-décomposition. Le critère suivant est celui que nous avons retenu car il est le plus utilisé et le plus simple à mettre en œuvre. Il est basé sur deux seuils θ_1 et θ_2 . Si nous définissons $a(t) = (e_{\max}(t) - e_{\min}(t))/2$ comme amplitude du mode et la fonction d'évaluation $\sigma(t) = |m(t)/a(t)|$, alors le processus de tamisage est répété jusqu'à ce que $\sigma(t) < \theta_1$ pour une durée de $(1-\alpha)$ de la durée totale, alors que $\sigma(t) < \theta_2$ pour le reste de la durée. Typiquement $\alpha \approx 0.05$, $\theta_1 \approx 0.05$ and $\theta_2 \approx 10\theta_1$.

IV.2.5 Variations sur la mise en œuvre

L'algorithme initial présenté à la section IV.2.2, opère le tamisage sur l'ensemble du signal. Il est poursuivi tant qu'il existe une zone locale pour laquelle la moyenne des enveloppes n'est pas suffisamment petite. Cependant cette sur-itération présente l'inconvénient d'étaler les différentes composantes du signal sur les modes adjacents et aussi l'inconvénient d'uniformiser l'amplitude. De plus, la nature non linéaire de la DEM assure en aucun cas que la DEM de l'ensemble du signal soit la concaténation des DEM individuelles. Deux solutions ont été proposées dans la littérature [Flandrin] : la « DEM locale » et la « DEM en ligne ».

A- DEM locale [Flandrin]

Cette variation de l'algorithme initial introduit une étape intermédiaire dans le processus de tamisage. Les zones locales où l'erreur reste importante sont localisées et isolées. Les itérations supplémentaires seront appliquées uniquement à ces zones. Ceci est réalisé grâce à une fonction de pondération $w(t)$ telle que $w(t)=1$ sur les segments pour lesquels $\sigma(t) > \theta_1$ et 0 en dehors. L'étape 4 de l'algorithme initial est modifiée comme suit : $d(t)=x(t)-w(t).m(t)$.

B- DEM en ligne [Flandrin]

La seconde variation de l'algorithme utilise le fait que l'opération de tamisage est basée sur l'interpolation entre les extrema ; Or cette dernière ne nécessite que 5 minima et 5 maxima pour le spline cubique par exemple. Il est donc possible de calculer un mode sans connaître tout le signal (ou même le résidu précédent). Cet état de fait suggère la possibilité d'un algorithme en ligne opérant par blocs sur des flux de données. Un pré-requis à l'extraction par blocs est l'utilisation du même nombre d'opérations de tamisage afin d'éviter les discontinuités. Ce nombre est donc fixé a priori à moins de 10 et typiquement 4. La DEM en ligne est réalisée grâce à une fenêtre glissante. Sur chaque fenêtre, la DEM locale est appliquée. La fenêtre avance au fur et à mesure que les données arrivent.

IV.3 SEPARATION DE SOURCES PAR L'APPROCHE DEM

La séparation de sources utilisant l'approche DEM est constituée de deux étapes: une étape analyse et une étape synthèse que nous allons décrire dans ce qui suit.

IV.3.1 Etape d'analyse

Dans cette étape, la DEM est appliquée à chaque mélange séparément. Le but est d'extraire toutes les composantes (IMFs) contenues dans chaque mélange. Cette étape a largement été décrite dans la section précédente.

IV.3.2 Etape de synthèse

Cette étape est complémentaire de l'étape précédente. Elle permet de regrouper toutes les composantes modales (IMFs) qui correspondent au même signal source afin de reconstituer le signal d'origine. Cette tâche sera accomplie par une méthode de classification non paramétrique : les K-moyennes [8]. Après l'étape d'analyse, les signaux sources sont supposés décomposables en somme de composantes modales c_i^j , nous pouvons alors écrire que :

$$s_i(t) = \sum_{j=1}^{I_i} c_i^j(t) \quad \text{pour} \quad t=0, \dots, T-1 \quad (\text{IV.13})$$

L'hypothèse d'indépendance des sources est remplacée ici par l'hypothèse d'orthogonalité des composantes modales, i.e. :

$$\frac{\int_{-T}^T s_i(t) s_j(t) dt}{\int_{-T}^T s_i^2(t) dt \int_{-T}^T s_j^2(t) dt} \approx 0 \quad \text{pour} \quad (i, j) \neq (i, i)$$

Où :

$$\int_{-T}^T s_i(t) s_j(t) dt = \sum_{j=1}^{I_i} \int_{-T}^T c_i^j(t) c_j^k(t) dt \quad \text{et} \quad \int_{-T}^T s_i^2(t) dt = \sum_{j=1}^{I_i} \int_{-T}^T c_i^j(t)^2 dt$$

L'équation (IV.13) appliquée aux observations $x_i(t)$ nous permet d'écrire:

$$x_i(t) = \sum_{j=1}^N a_{ij} s_j(t) = \sum_{j=1}^N \sum_{k=1}^{I_j} a_{ij} c_j^k(t) \quad (\text{IV.14})$$

Les $\hat{c}_i^j(t)$ sont des estimées des composantes $c_i^j(t)$. L'hypothèse de quasi orthogonalité permet d'écrire :

$$\frac{\langle x | c_i^j \rangle}{\|c_i^j\|^2} = \frac{1}{\|c_i^j\|^2} \begin{bmatrix} \langle x_1 | c_i^j \rangle \\ \vdots \\ \langle x_M | c_i^j \rangle \end{bmatrix} \approx a_i \quad (\text{IV.15})$$

Où a_i représente la $i^{\text{ème}}$ colonne de la matrice $\mathbf{A}$ de mélange. Chaque composante $\tilde{c}_j^k$ est associée à une colonne de la matrice $\mathbf{A}$ et sera estimée grâce à la relation :

$$\hat{a}_j^k = \frac{\langle x | \tilde{c}_j^k \rangle}{\|\tilde{c}_j^k\|^2} \quad (\text{IV.16})$$

Deux composantes du même signal source sont associées au même vecteur colonne de $\mathbf{A}$, il faut donc regrouper ces dernières par classification en N classes. Enfin, en additionnant les différentes composantes d'une même classe, on reconstruit les sources initiales à une constante près. Un classificateur non supervisé tel que les k-moyennes s'avère efficace pour une telle tâche. L'organigramme de la figure IV.2 résume les étapes décrites ci-dessus.

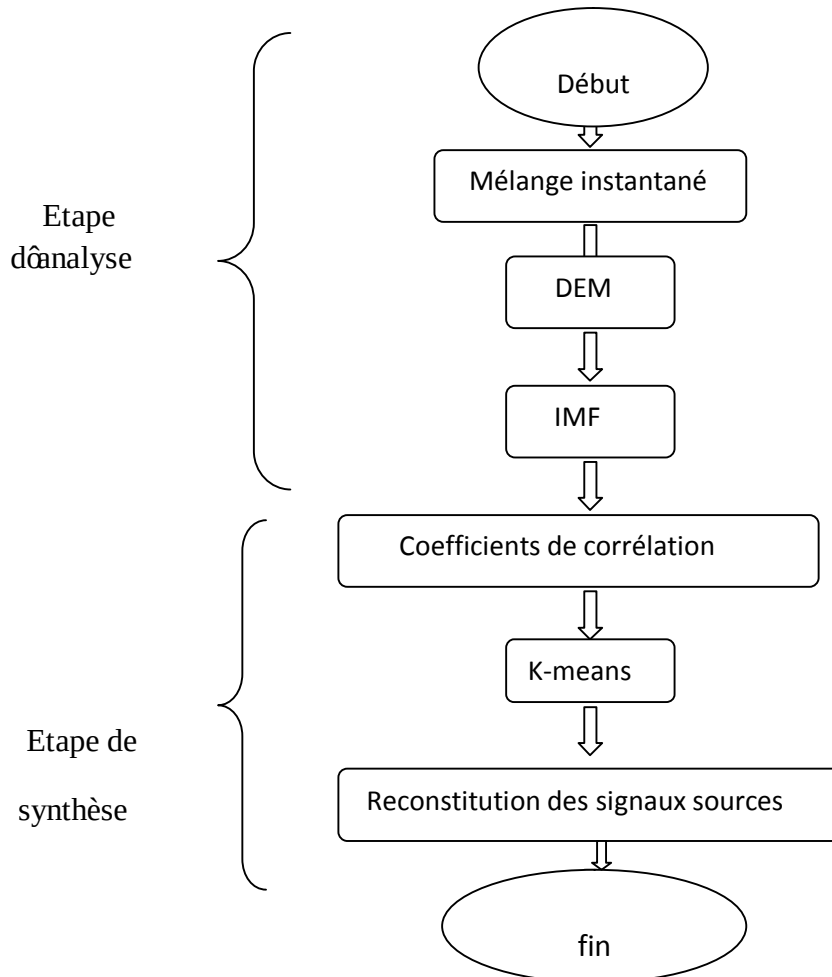


Figure IV.2 : Organigramme SAS par approche EMD.

IV.4 CRITERES D'ÉVALUATION

Au delà de critères visuels ou auditifs qui restent subjectifs et ne permettent en aucun cas l'automatisation du processus d'évaluation de la qualité de séparation, il existe dans littérature divers critères objectifs d'évaluation [2, 3]. Ces critères ont été utilisés, dans la littérature, pour mesurer les qualités de séparation et de restitution des méthodes de séparation ainsi que les distorsions qu'elles peuvent induire. Elles permettent la mesure de l'erreur totale sur l'estimation d'une source mais aussi la mesure des contributions des différentes causes d'erreur possible (interférences, bruit, artefacts). Nous proposons d'adopter les deux critères objectifs suivants : le critère DSR (Distorsion to Signal Ratio) et le critère ISR (Interference to Signal Ratio).

IV.4.1 Le Critère DSR

Le critère Distorsion to Signal Ratio (DSR) est une mesure de la distorsion relative totale. Il est défini comme le rapport des énergies des deux termes de l'équation (IV.17) :

$$\check{\mathbf{S}}_k = \left\langle \check{\mathbf{S}}_k, \frac{\mathbf{s}_k}{\|\mathbf{s}_k\|} \right\rangle \frac{\mathbf{s}_k}{\|\mathbf{s}_k\|} + \mathbf{e}_{total} \quad (IV.17)$$

Le terme $\left\langle \check{\mathbf{S}}_k, \frac{\mathbf{s}_k}{\|\mathbf{s}_k\|} \right\rangle \frac{\mathbf{s}_k}{\|\mathbf{s}_k\|}$ représente la projection de la source estimée $\check{\mathbf{S}}_k$ sur la source originale $\mathbf{s}_k$. Le DSR est alors défini comme :

$$DSR = \frac{\|\mathbf{e}_{total}\|^2}{\left\| \left\langle \check{\mathbf{S}}_k, \frac{\mathbf{s}_k}{\|\mathbf{s}_k\|} \right\rangle \frac{\mathbf{s}_k}{\|\mathbf{s}_k\|} \right\|^2} \quad (IV.18)$$

Le terme $\mathbf{e}_{total}$ est par construction orthogonal à la contribution de la source originale $\mathbf{s}_k$. Avec le théorème de Pythagore, nous avons donc :

$$\|\mathbf{e}_{total}\|^2 = \|\check{\mathbf{S}}_k\|^2 - \left\| \left\langle \check{\mathbf{S}}_k, \frac{\mathbf{s}_k}{\|\mathbf{s}_k\|} \right\rangle \frac{\mathbf{s}_k}{\|\mathbf{s}_k\|} \right\|^2 \quad (IV.19)$$

Si nous remplaçons (IV.19) dans (IV.18) nous obtenons une nouvelle équation pour le DSR :

$$DSR = \frac{\left\| \check{\mathbf{S}}_k \right\|^2 - \left| \left\langle \check{\mathbf{S}}_k, \frac{\mathbf{s}_k}{\|\mathbf{s}_k\|} \right\rangle \right|^2}{\left| \left\langle \check{\mathbf{S}}_k, \frac{\mathbf{s}_k}{\|\mathbf{s}_k\|} \right\rangle \right|^2} \quad (IV.20)$$

Si l'estimation $\check{\mathbf{S}}_k$ de la source $\mathbf{s}_k$ est parfaite alors le $DSR = 0$; et lorsque cette estimation est très éloignée de la source, alors le DSR tend vers l'infini.

IV.4.2 Le Critère ISR

Le terme e_{total} inclut les contributions des autres sources (interférences), du bruit n , ainsi que des éventuels « artefacts » dus aux algorithmes de séparation. Dans certains problèmes de la SAS, la nature de l'erreur d'estimation est très importante et il convient d'évaluer les différentes contributions des différents types d'erreur. Par exemple, en séparation de sources audio, l'erreur due aux artefacts de l'algorithme (ajout de zéros dans la transformée de Fourier à court terme) est plus gênante à l'écoute qu'un résidu de bruit ou d'interférences. e_{total} peut alors décomposée en trois termes : un terme relatif aux interférences, un terme relatif au bruit, et un terme relatif aux artefacts. L'équation (IV.17) peut alors se réécrire :

$$\check{\mathbf{S}}_k = \left\langle \check{\mathbf{S}}_k, \frac{\mathbf{s}_k}{\|\mathbf{s}_k\|} \right\rangle \frac{\mathbf{s}_k}{\|\mathbf{s}_k\|} + e_{int\ erf} + e_{bruit} + e_{artef}$$

Les algorithmes fournissant une estimation des sources par application d'une matrice de séparation aux observations ne génèrent pas d'artefacts [2]. L'erreur due aux artefacts e_{artef} est nulle dans notre cas puisque l'ICA fait partie de cette classe d'algorithme. Nous nous intéressons plus particulièrement aux erreurs introduites par les interférences dues aux autres sources e_{interf} définie par l'équation :

$$e_{int\ erf} = \sum_{l \neq k} \left\langle \check{\mathbf{S}}_k, \frac{\mathbf{s}_l}{\|\mathbf{s}_l\|} \right\rangle \frac{\mathbf{s}_l}{\|\mathbf{s}_l\|} \quad (IV.21)$$

Le critère ISR mesurant la distorsion relative due aux interférences est défini par l'équation :

$$ISR = \frac{\left\| e_{int\ erf} \right\|^2}{\left\| \left\langle \check{\mathbf{S}}_k, \frac{\mathbf{s}_k}{\|\mathbf{s}_k\|} \right\rangle \frac{\mathbf{s}_k}{\|\mathbf{s}_k\|} \right\|^2} \quad (IV.22)$$

La qualité de séparation et de restitution est meilleure lorsque le critère de distorsion DSR et le critère de restitution ISR tendent vers zéro.

Conclusions

Dans ce chapitre, nous avons présenté deux méthodes concurrentes de séparation de sources : l'Analyse en Composantes Indépendantes (ACI) et la Décomposition Empirique Modale (DEM).

L'Analyse en Composantes Indépendantes (ACI) est une des solutions élégantes et efficaces de Séparation Aveugle de Sources. C'est une méthode statistique linéaire, qui vient en prolongement direct de l'Analyse en Composantes Principales (ACP) en remplaçant la notion de décorrélation par celle d'indépendance. Le modèle ACI peut être résolu par divers algorithmes ; Nous avons détaillé l'un des plus rapides et des plus utilisés : l'algorithme FastICA. Il introduit la notion de non-gaussianité dans le processus de séparation. La mesure de non-gaussianité utilisée par FastICA est la néguentropie (non negative entropy) et l'optimisation repose sur l'algorithme du point fixe. Deux approches coexistent : l'approche dite par déflation et l'approche symétrique.

La Décomposition Empirique Modale (EMD) repose sur l'identification itérative et adaptative d'un nombre fini d'IMF en séparant localement une contribution rapide d'une tendance plus lente. Ces IMF représentent les oscillations naturelles et intrinsèques présentées dans le signal considéré. De manière plus formelle, cette décomposition repose sur un algorithme itératif, le tamisage et non pas sur une formulation analytique contrairement à l'algorithme FastICA. Cette approche présente des propriétés remarquables par rapport à d'autres méthodes du traitement de signal: c'est une méthode qui est définie algorithmiquement. Elle analyse localement le signal et cette propriété fait de l'EMD un outil bien adapté pour l'analyse de signaux non stationnaires, comme les signaux audio. L'EMD se rapproche d'une analyse multi résolution dans la mesure où elle explore le signal depuis les plus hautes fréquences vers les plus basses, du plus petit détail au plus grand. Enfin les oscillations des IMFs ne sont pas supposées harmoniques, ce qui traduit une grande liberté d'analyse.

Pour comparer ces deux méthodes, nous avons introduits deux critères objectifs, le critère DSR et le critère ISR. Ils représentent respectivement la distorsion relative totale et la distorsion relative totale due aux interférences. Nous avons montré que lorsque ces deux critères tendaient vers 0, la qualité de séparation était meilleure.

- [1] Emmanuel Vincent, « modèles d'instruments pour la séparation des sources et la transcription d'enregistrement musicaux », thèse de doctorat, université Paris VI –Pierre et Marie curie, Décembre 2003.
- [2] Cédric Févotte, “*Approche temps fréquence pour la séparation aveugle de source non stationnaires*”, PhD thesis, Ecole Centrale de Nantes et Université de Nantes, Octobre 2003.
- [3] E. Vincent, C. Févotte, and R. Gribonval et al., « A tentative typology of audio source separation tasks, *In Proc. 4th Symposium on Independent Component Analysis and Blind Source Separation (ICA'03)*, Nara, Japan, Avril 2003.
- [4] Hyvärinen A., “*Survey on independent component analysis*”, *Neural Computing Surveys*, 2 : 94-128, 1999.
- [5] A. Hyvärinen, E. Oja, « A fast fixed-point algorithm for independent component analysis », *Neural Computation*, 9(7):1483-1492, 1997.
- [6] A. Hyvärinen, “A fast and robust fixed-point algorithm for independent component analysis”, *IEEE Transactions on Neural Networks*, 10(3):626-634, 1999.
- [7] A. Hyvärinen, E. Oja, “Independent component analysis: algorithms and applications”, *Neural Networks*, 13:411-430, 2000.
- [8] Flandrin P., Rilling G. and Gonçalves P. «*Empirical Mode Decomposition as a filter bank* », in *IEEE Sig.Proc.Lett.*, Vol.11, N°2, pp. 112-114.
- [9] Aissa-El-Bey A., Abed-Meraim K. et Grenier Y., « *Séparation de sources aveugle sous-déterminée basée sur l'algorithme EMD* », 20ème Colloque GRETSI, Septembre 2005.
- [10] Aissa-El-Bey A., Abed-Meraim K. et Grenier Y., « *Blind Separation of audio sources using Modal Decomposition* », 8th ISSPA, Vol.2, pp.451-454, Sydney
- [11] A. Belouchrani, K. Abed-Meraim, J.F. Cardoso, and E. Moulines, “Blind source separation using second order statistics” *IEEE-SP*, vol. 45, no. 2, pp.434-444, Feb. 1997.
- [12] J.F. Cardoso, “Source separation using higher order moments”, *In Proc. ICASSP'89*, pp 2109-2112, 1989.
- [13] J-F. Cardoso, “Blind signal separation: statistical principles”, *Proceedings of the IEEE*, 9(10):2009-2025, October 1998.
- [14] J-F. Cardoso, “High-order contrasts for independent component analysis”, *Neural Computation*, 11(1):157-192, 1999.
- [15] J-F. Cardoso, “The three easy routes to independent component analysis: contrasts and geometry”, *In Proc. of the ICA 2001 workshop*, pp. 1-6, San Diego, December 2001.

- [16] P. Comon, "Analyse en composantes indépendantes et identification aveugle", *Traitement du signal*, Vol. 7, no. 5, pp.435-450, 1990.
- [17] P. Comon, "Independent component analysis, a new concept", *Signal Processing*, 36:287-314, 1994.
- [18] J. Héroult and B. Auzan, « Réseaux de neurones à synapses modifiables : décodage de messages sensoriels composites par un apprentissage non supervisé et permanent », *In C.R. Acad ; Sci.*, pp. 525-528, série III, Paris, 1984.
- [19] Y. Deville, « Panorama des applications biomédicales des méthodes de séparation aveugle de sources. » *In Actes 19^{ème} colloque GRETSI sur le Traitement du Signal et des Images*, PARIS, Sep. 2003.
- [20] C. Jutten, "Calcul neuromimétique et traitement du signal", PhD. Thesis, Université scientifique et médicale, Institut National Polytechnique de Grenoble, 1987.
- [21] C. Jutten, J. Héroult, « Blind separation of sources, part I : an adaptive algorithm based on neuromimetic architecture », *Signal Processing*, 24:1-10, 1991.

CHAPITRE 5

Classification en Genres de Signaux Musicaux
--

Introduction

L'exploitation des bases de données musicales, de tailles sans cesse croissantes, a provoqué l'émergence de systèmes de recherche d'information musicale. Les stations radio et les chaînes de télévisions musicales disposent de millions d'enregistrements ; L'indexation automatique de telles bases de données est alors nécessaire pour rechercher rapidement et efficacement une information. Le genre musical est probablement le descripteur musical le plus commun et le plus évident. Ainsi, un système qui reconnaîtrait le genre musical permettrait à une radio de sélectionner automatiquement les stations diffusant un genre favori de musique. Un genre musical peut être défini comme un label regroupant un ensemble de signaux musicaux qui partagent un ensemble de caractéristiques communes à ce genre, parmi lesquelles : le type d'instruments, le contenu harmonique et rythmique du morceau musical. Cependant, les frontières entre genres restent mal définies et le genre reste largement dépendant des contextes historiques et culturels. ?

Trois approches de la classification en genres de signaux musicaux coexistent [33]; L'approche manuelle utilise les connaissances d'experts sur les titres musicaux : elle est coûteuse en temps. C'est l'approche adoptée par le projet européen CUIDADO [34]. Les deux autres approches sont automatiques ; D'une part l'approche dite prescriptive tente d'extraire automatiquement l'information sur le genre musical à partir du signal audio en utilisant des méthodes de traitement du signal et des méthodes d'apprentissage. Tous les travaux proposés [12, 29, 31, 32] partent d'une taxonomie des genres imposée et tentent d'associer chaque morceau de musique de la base de données à un genre donné. Elles procèdent toutes en deux étapes : une première étape d'extraction des descripteurs acoustiques relatifs au timbre, au rythme ou au contenu harmonique, et une seconde étape de classification (apprentissage machine) pour laquelle un algorithme de classification, appliquée aux vecteurs acoustiques, et qui associe à chaque vecteur le label de la classe la plus probable : son genre. Il s'agit donc d'une classification supervisée puisque la modélisation des classes requiert une étape d'apprentissage préalable. Les travaux réalisés dans le cadre de cette approche reposent sur le choix des vecteurs acoustiques qui reste un facteur prépondérant dans la classification en genres. En effet un vecteur acoustique est formé d'un ensemble de descripteurs calculés à partir du signal audio. La plupart des descripteurs musicaux sont issus de domaine de la parole (Mel Frequency Cepstral Coefficients, Linear Prediction Coefficients, Taux de Passage par Zéro;) ou de l'acoustique (Loudness, Centre de Gravité Spectral, Flux spectral,é) ; Une corrélation certaine existe entre ces descripteurs dits de bas niveau et le descripteur de haut niveau timbre [12]. Certains auteurs préconisent de compléter le vecteur acoustique par des descripteurs relatifs au rythme, alors que d'autres proposent de le compléter par des descripteurs relatifs au pitch . Les machines d'apprentissage utilisées à ce jour pour classer les morceaux musicaux (ou plus exactement les vecteurs acoustiques) sont les mixtures de gaussiennes, les réseaux de neurones et enfin la quantification vectorielle . Une nouvelle machine d'apprentissage très en vogue actuellement est la machine à vecteurs support suscite l'intérêt des chercheurs bien que très peu de travaux aient été publiés sur la classification en genres de signaux musicaux par machines à vecteur support. D'autre part, une approche

émergente, à l'opposé de l'approche prescriptive décrite précédemment, regroupe en classes les signaux musicaux de la base de données grâce à une mesure de similarité : la classification en genres émerge alors de ce regroupement. Il s'agit en fait d'une classification non supervisée [33].

Dans ce chapitre, nous nous intéresserons à la classification en genres de signaux par l'approche prescriptive. Nous avons décrits dans le chapitre II la construction et les différentes combinaisons de descripteurs permettant la construction des vecteurs acoustiques associées aux signaux musicaux de la base de données que nous avons extraite de la base de données de la radio algérienne. Nous les avons regroupés sous les catégories : descripteurs temporels, descripteurs spectraux, descripteurs perceptuels et ondelettes. L'évaluation de la pertinence des différents vecteurs acoustiques peut se faire à travers plusieurs machines d'apprentissage ou classifieurs tels que : Mixture de gaussiennes, Réseaux de neurones, K plus proches voisins et Machines à vecteurs support. L'originalité étant de tester cette approche pour la classification en genres de signaux musicaux orientaux puisque toutes les études de ce style ne s'adressent qu'à des signaux de musique occidentale. Nous introduirons les Machines à support vecteur multilabels pour cette classification en genres de signaux musicaux. La notion de multilabels tire profit des connaissances à priori sur les signaux musicaux, à savoir que certains genres ne sont pas complètement disjoints : un signal pourra ainsi appartenir à une ou plusieurs classes. Nous commençons donc par décrire les différentes machines d'apprentissage .

V.1 ETAT DE L'ART SUR LA CLASSIFICATION EN GENRES

Dans la littérature on retrouve plusieurs tentatives de classification de signaux musicaux par des approches génératives ou discriminatives. Les premiers et pertinents travaux sont dûs à Tzanetakis et Cook [12]. Ils ont testé les GMM (avec 2 à 5 classes) et un classifieur kppv sur trois bases de données musicales. Chaque fichier musical est représenté par un vecteur acoustique comprenant le timbre et la fréquence fondamentale ainsi que des descripteurs temporels. Sur les trois bases de données, le classifieur GMM avec trois clusters donne les meilleurs résultats de classification, selon les auteurs. Les travaux de Lippens et al. [25] comparent les résultats fournis par un classifieur monogaussien avec ceux fournis par un classificateur humain (une personne) sur une base de 160 chansons réparties en 10 genres. La classification humaine a obtenu les meilleurs résultats avec un taux de 20% supérieur au classifieur monogaussien (soit 88% pour l'être humain contre 65% pour le classifieur monogaussien). On relève aussi dans Aucouturier et Pachet [26] des expériences sur les fichiers musicaux afin de trouver des paramètres optimaux d'un système de classification en genres basé sur une modélisation MFCC et un classifieur GMM. Les résultats aboutissent, selon les auteurs, à un vecteur acoustique de 20 MFCC combiné avec 50 gaussiennes constituant le modèle. Un échantillonnage de Monte Carlo est exploité pour produire la matrice de similarité audio. Celle-ci constitue l'entrée d'un algorithme kppv. Plusieurs chercheurs tel que Flexer et al. [27] ont exploité la piste des HMM pour la classification en genres musicaux. Les expériences ont été menées sur un système composé de vecteurs MFCC et 50 GMM. L'objectif attendu était d'utiliser les HMM pour modéliser les variations

temporelles des descripteurs MFCC. Les résultats obtenus ne sont pas meilleurs que ceux obtenus par Aucouturier et Pachet.

Il est intéressant de souligner les travaux de Mitri et al. reportés dans [28]. Trois algorithmes de classification ont été comparés sur des segments de 10 secondes de fichiers musicaux représentant 10 genres audio. Des descripteurs temporels et spectraux de dimension 58 ont été mis en œuvre. Mandel et Elis comparent dans [29] les performances d'une classification SVM avec celle d'un kppv sur une tâche de classification en genre et d'artistes. Les résultats montrent qu'une classification par SVM donne de meilleurs taux d'erreurs sur la base de descripteurs issues de l'analyse des signaux musicaux « isolés » comparativement à des descripteurs issues d'une analyse de segments incluant le morceau musical et la voix de l'artiste.

Dans la section suivante, nous présentons les méthodes de classification les plus utilisées dans la littérature pour la classification de signaux audio. Ces méthodes peuvent être subdivisées en deux approches distinctes :

- Les approches dites « génératives » telles que : l'Analyse discriminante linéaire, les Modèles de Mélanges de Gaussiennes, les Modèles de Markov Cachés et les Réseaux Bayésiens [5,11].
- Les approches dites « discriminantes », telles que : les K plus proches voisins, les réseaux de neurones et les Machines ou Méthodes à Vecteurs de Support (SVM) [5,11].

V.2 FORMALISATION PROBABILISTE DE LA CLASSIFICATION

Un Classifieur assigne à une observation x (un vecteur acoustique) une étiquette $l \in \{1, \dots, n_c\}$ correspondant à une classe. (n_c désigne le nombre de classes). Selon la théorie des probabilités, chaque observation est générée par une variable aléatoire, dont la distribution peut se décomposer selon la loi de Bayes :

$$p(x, l) = p(l | x)p(x) = p(x | l)p(l) \quad (V.1)$$

Le but d'un classifieur est de minimiser l'espérance du coût des erreurs ou risque global. On montre dans [1] que le classifieur de Bayes est celui qui minimise ce risque global. En pratique les véritables densités sont inconnues et l'on dispose d'observations d'apprentissage $\{a_i, l_i\}$ à partir desquelles on instancie des modèles paramétriques.

V.2.1 Approches génératives

Les approches génératives regroupent des méthodes qui utilisent les données d'apprentissage pour modéliser les densités de probabilité $p(x | l)$ de chaque classe par une famille de fonctions paramétriques. Le terme « génératif » désigne le fait que la règle de

décision, déduite d'après les relations de Bayes , soit basée sur une modélisation de la probabilité $p(x|l)$ qui génère les observations x pour une classe donnée.

V.2.2 Approches discriminantes

Les approches discriminantes regroupent une variété de méthodes statistiques qui utilisent les données d'apprentissage pour construire directement une correspondance entre les entrées X et les sorties Y . Il est souvent dit que ces méthodes modélisent directement la probabilité a posteriori $p(l|x)$. La plupart du temps, l'accès à la probabilité a posteriori n'est pas trivial et nécessite des hypothèses supplémentaires à celles faites lors de l'apprentissage du classifieur discriminatif. Globalement, nous désignons par approches discriminantes les méthodes qui ne s'intéressent qu'aux relations d'entrée-sortie lors de l'apprentissage.

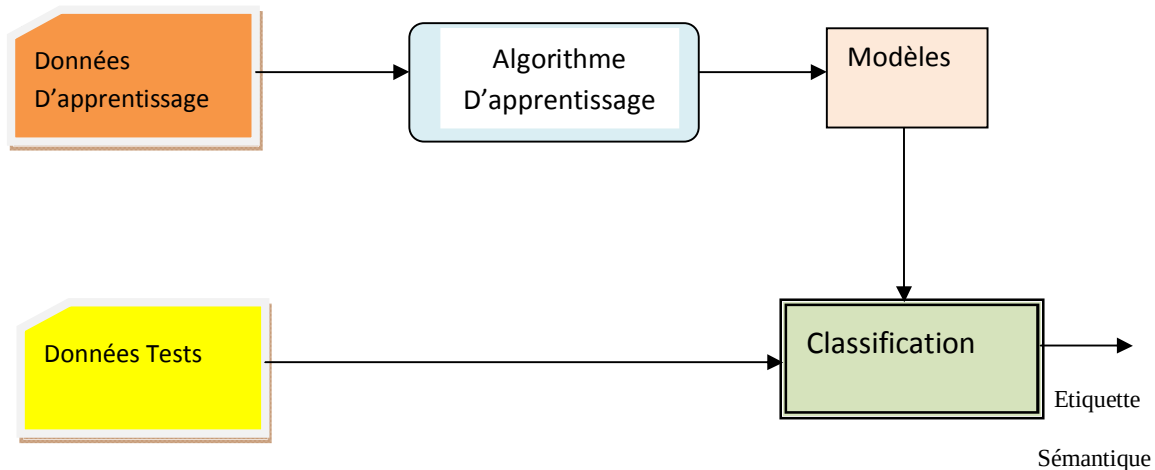


Figure V.1 : Schéma Général de la Classification

V.3 FORMALISATION DE LA CLASSIFICATION EN TERMES DE RECONNAISSANCE DE FORMES

Lors de l'apprentissage, les méthodes génératives optimisent plusieurs fonctions objectives, chacune faisant intervenir un sous-ensemble de vecteurs d'apprentissage correspondant à une seule et même classe. A l'opposé, les méthodes discriminantes optimisent une ou plusieurs fonction(s) objective(s) faisant chacune intervenir plusieurs classes. En d'autres mots, les méthodes génératives exploitent l'information intra-classe et les méthodes discriminantes utilisent l'information inter-classes. Dans le cas particulier d'une classification binaire, un apprentissage discriminatif optimise une seule fonction objective qui fait intervenir les deux classes, alors qu'un apprentissage génératif optimise séparément deux densités de probabilité conditionnelles. Ainsi :

- les approches génératives optimisent les densités de probabilité de chacune des classes et induisent une frontière de décision dont l'expression analytique fait intervenir des modèles avec les paramètres choisis.
- Les approches discriminantes optimisent directement une frontière de décision.

V.3.1 Cas Multi-Classe

Les approches discriminantes sont naturellement dédiées aux bi-classes. Certaines méthodes comme les kppv se déclinent au cas multi-classes d'une façon naturelle, d'autres sont basées sur des fonctions objectives d'apprentissage bi-classes. On distingue alors deux approches pour généraliser les méthodes discriminantes au cas multi-classes :

- Construire $nc(nc-1)/2$ classifieurs binaires visant à séparer chaque paire de classes, et construire un arbre de décision à partir de ces classifieurs. C'est la stratégie *un contre un* (one against one).
- Construire nc classifieurs binaires visant à séparer chaque classe de toutes les autres, et combiner les sorties des différents classifieurs (typiquement, la classe choisie est celle renvoyant le plus grand score). C'est la stratégie *un contre tous* (one against all).

V.3.2 Combinaison d'approches

Afin de tirer profit des deux types d'approches, des recherches récentes ont été conduites afin de construire des méthodes hybrides (génératives/discriminatives) : Citons ces exemples d'applications en bio-informatique, vision et audio [2, 3]. Il s'agit en fait de tirer parti des avantages de chaque méthode : pour les modèles génératifs, la possibilité de traiter naturellement des séquences de taille variable (avec une robustesse aux observations aberrantes) et pour les modèles discriminatifs, l'adéquation du critère d'apprentissage au problème de classification (minimisation d'un terme erreur adéquat, avec garanti de généralisation). Il y a plusieurs manières de combiner une modélisation générative (G) et une modélisation discriminante (D) [3]:

- Deux classifieurs G et D agissent en parallèle, et leurs résultats sont combinés.
- G est imbriqué dans D.
- D est imbriqué dans G.

V.4 EXEMPLES D'APPROCHES GENERATIVES

V.4.1 MELANGES DES LOIS GAUSSIENNES

Les Mélanges des Lois Gaussiennes, ou Gaussien Mixture Models (GMM) en anglais, sont une famille de classificateurs paramétriques où on suppose que la distribution des données pour chaque classe est une combinaison de plusieurs distributions gaussiennes dans l'espace de

représentation. Etant donné les paramètres de chaque Gaussienne, on peut calculer aisément la probabilité d'appartenance d'un point à une classe. L'apprentissage des GMM se fait généralement par l'algorithme Expectation Maximisation (EM) qui permet un calcul itératif des paramètres de chaque Gaussienne, c'est à dire le vecteur moyen μ et la matrice de covariances, ainsi que leur poids ω

$$\forall x \in \mathbb{R}^d, p(x/\theta) = \sum_{g=1}^G \omega_g N(x/\mu_g, \Sigma_g) \quad (V.2)$$

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

Rappelons que la densité de probabilité de la loi gaussienne s'écrit :

$$N(x/\mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^d \det(\Sigma)}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)} \quad (V.3)$$

Si l'on note d la dimension des vecteurs d'entrée, le nombre de paramètres libres d'un GMM à G Gaussiennes est $M = G(d^2 + d + 1)$. Mais en pratique, les matrices de covariances sont souvent contraintes à être diagonales. Le nombre de paramètres libres est alors réduit à $M = G(2d + 1)$. Cela a l'avantage, de réduire la complexité de la modélisation pour éviter le sur-apprentissage. Aussi la complexité de la modélisation GMM se règle principalement avec le nombre G de Gaussiennes, par validation (croisée ou simple).

V.4.2 MODELES DE MARKOV CACHES (HMM)

Les HMM sont des modèles probabilistes dynamiques, capables de capturer les dépendances temporelles entre les différentes variables aléatoires. Ils permettent donc de manipuler les séquences sans faire d'hypothèse restrictive comme l'indépendance entre les observations qui forment chaque séquence. Les HMM ont une structure composée de variables aléatoires continus qui dépendent de variables aléatoires discrètes cachées, appelées « états ». Il est supposé que chaque état à un instant $t+1$ ne dépend que de l'état à l'instant t précédent. Un HMM est paramétré par les probabilités initiales de chaque état, la matrice des probabilités de transition entre les états et enfin les fonctions de densité modélisant les probabilités d'émission de la variable continue selon chaque état. Pour la reconnaissance de la parole [22], les variables observées sont les vecteurs acoustiques et les états cachés représentent les phonèmes à reconnaître. Leur probabilité d'émission selon chaque état est couramment modélisée par un GMM.

V.5 EXEMPLES D'APPROCHES DISCRIMINANTES

Dans cette partie, nous présentons trois exemples de modélisation discriminante dont les Machines à Vecteurs de Support. Ces méthodes d'optimisation n'exploitent pas la théorie des probabilité contrairement aux modèles génératifs.

V.5.1 *k-plus proches voisins*

Cette méthode discriminante appartient à la catégorie des algorithmes graphiques et ne comporte pas d'étape d'apprentissage à proprement parler. Elle consiste à garder en mémoire tous les vecteurs d'apprentissage étiquetés et elle repose sur une mesure de distance arbitraire entre les vecteurs. En phase de test, les distances entre le vecteur à classer et tous les vecteurs d'apprentissage sont estimées et rangées en ordre décroissant. Pour la décision, on procède par vote majoritaire parmi les k vecteurs d'apprentissage les plus proches.

Outre sa simplicité, l'avantage de cette méthode est qu'elle peut naturellement s'appliquer au cas multi-classes même avec un nombre élevé de classes [1,23]. Mais les inconvénients sont de taille :

1. Un volume important de données d'apprentissage implique une capacité des ressources mémoire nécessaires d'autant plus élevée, ainsi qu'une forte complexité calculatoire en phase de test.

2. Le renvoi d'une mesure de confiance de la décision (score) ne peut se faire que de manière arbitraire, par exemple en calculant une moyenne des distances au k -plus proches voisins à partir des distances calculées.

V.5.2 *Les Réseaux de Neurones*

Les Réseaux de Neurones (ANNs) sont une des premières méthodes discriminantes pour la classification et la régression à avoir connu un réel succès dans de nombreuses applications [1,3]. Ils peuvent servir à la classification comme à la régression non linéaire de fonctions. Construit sur un paradigme biologique et inspiré du système nerveux humain, l'ANN est un système de neurones formels inter-connectés dans un réseau, et qui coopèrent pour fournir un score comme représenté dans l'exemple de Fig.V.2. Les neurones sont des opérateurs qui appliquent une fonction aux valeurs d'entrée qu'il reçoit d'un autre neurone (ou de l'entrée du réseau) pour renvoyer une valeur vers un autre neurone ou une sortie du réseau. Les paramètres libres de chaque fonction de neurone sont appelés coefficients synaptiques. On distingue deux types de fonctions synaptiques :

- les fonctions de combinaison, qui sont des combinaisons linéaires des entrées pour les réseaux de type MLP, où des fonctions monotones de la distance entre le vecteur d'entrée et un vecteur de référence pour les réseaux de type RBF.

- les fonctions d'activation, qui sont typiquement des fonctions sigmoïdes (seuillage régularisé).

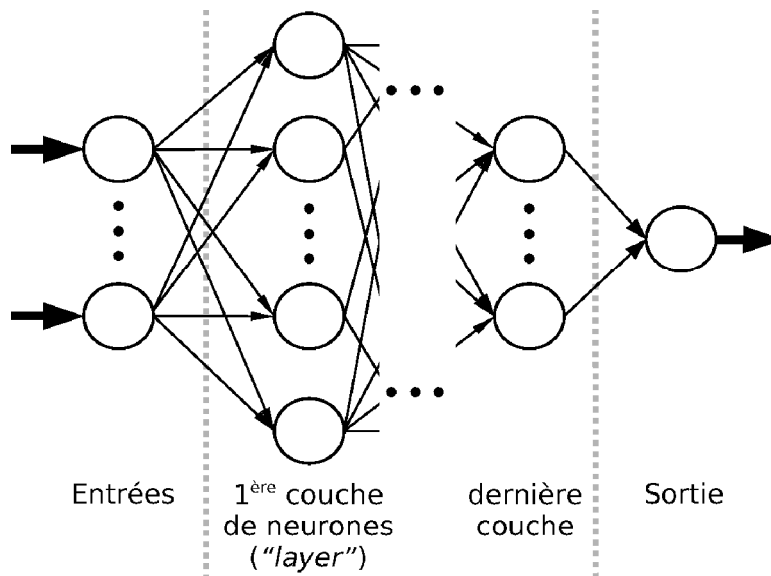


Figure V.2 : Structure d'un Réseau de Neurone

Les RNA donnent de bonnes performances pour de nombreux problèmes statistiques. Ils ont été appliqués avec succès à la vérification du locuteur par [24]. Peu de travaux ont été consacré cependant à l'audio et aux signaux musicaux.

V.5.3 Les Support Vector Machines (SVMs)

L'acronyme SVM est l'abréviation du terme anglophone « Support Vector Machine » ; ses adaptations françaises sont « Machine à Vecteur de Support » et « Séparateur à Vaste Marge ». Les SVMs sont une méthode de discrimination proposée par Vladimir Vapnik durant les années 90 [], ils visent à faire la distinction entre deux ou plusieurs ensembles de points par un hyperplan séparateur (frontière de décision) [6,3].

A- Principe de fonctionnement

L'objectif des SVMs est de trouver un classificateur qui sépare les données en classes en maximisant la marge entre elles.

En effet il existe une infinité d'hyperplans séparateurs ; mais l'objectif des SVMs est que l'hyperplan considéré doit être « **optimal** ».

L'hyperplan optimal doit non seulement passer entre les points des deux classes, mais il faut aussi que la distance minimale entre l'hyperplan et les points qui lui sont le plus proches (marge) soit la plus grande possible, d'où l'appellation « *séparateurs à vaste marge* ».

La figure suivante illustre deux exemples de classificateurs, celui de gauche a une faible marge (ce qui entraîne une erreur de classification), l'autre a une marge optimale.

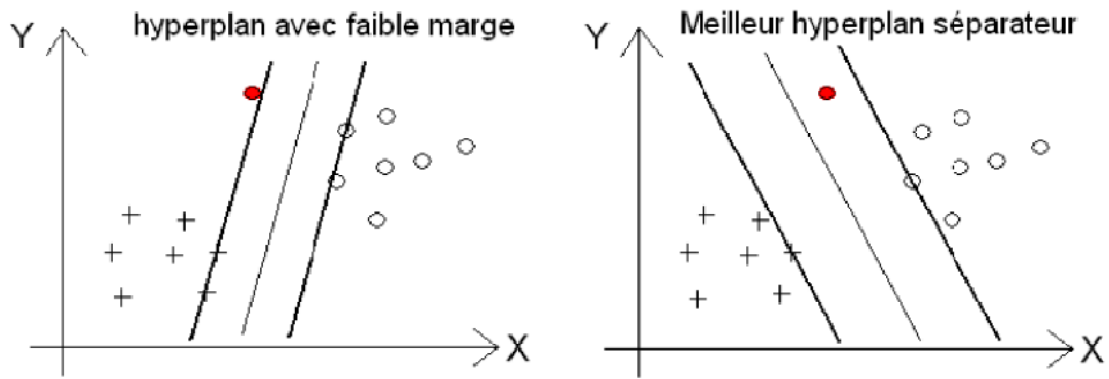


Figure V.3 : Exemple de séparation des classes avec différents hyperplans.

L'hyperplan séparateur correspond à la médiatrice du plus petit segment de droite reliant les enveloppes des deux catégories. Du moment qu'il existe deux plans parallèles à l'hyperplan séparateur correspondants à chaque classe, ces deux plans seront appelés « *hyperplans canoniques* ».

B- Les différentes formulations des SVM

Parmi les modèles de SVMs, on constate les cas linéairement séparables, les cas non linéairement séparables et les cas non séparables. Les premiers sont les plus simples, car ils permettent de trouver facilement un classificateur linéaire. Dans la plupart des problèmes réels il n'y a pas de séparation linéaire possible entre les données, ce sont des cas non linéairement séparables et des cas non séparables.

-Cas linéairement séparable (formulation à marge dure)

Etant donné un ensemble de points d'apprentissage $\{x_i, y_i\}_{i=1}^l$ où chaque x_i appartient à $\mathbb{R}^m$ et y_i à l'ensemble $\{-1, +1\}$. Les y_i définissent la classe d'un exemple donné, dans ce cas, l'hyperplan permet de séparer l'ensemble d'apprentissage de sorte que tous les points d'une même classe soient d'un même côté de l'hyperplan.

- Construction de l'hyperplan optimal

Tout d'abord considérons l'ensemble suivant : $D = \{(x_i, y_i) \in \mathbb{R}^m \times \{-1, 1\} \text{ pour } i = 1, \dots, m\}$

Soit $H : (w \cdot x) + b$ l'hyperplan qui peut satisfaire les conditions suivantes :

$$\begin{cases} w \cdot x_i + b \geq 1 & \text{si } y_i = 1 \\ w \cdot x_i + b \leq -1 & \text{si } y_i = -1 \end{cases} \quad (\text{V.4})$$

Ce qui est équivalent à :

$$y_i (w \cdot x_i + b) \geq 1 \quad \text{pour } i=1, \dots, m \quad (\text{V.5})$$

Maximiser la marge M équivaut à maximiser la distance entre les hyperplans canoniques. Ainsi la marge a l'expression suivante :

$$M = \min_{x_i/y_i=1} \frac{w \cdot x + b}{\|w\|} - \max_{x_i/y_i=-1} \frac{w \cdot x + b}{\|w\|}$$

Soit :
$$M = \frac{1}{\|w\|} - \frac{-1}{\|w\|} \quad \text{d'où} \quad M = \frac{2}{\|w\|} \quad (\text{V.6})$$

Trouver l'hyperplan optimal revient à maximiser la marge M , et donc à minimiser le rapport $\frac{\|w\|^2}{2}$ sous contrainte. Cette opération s'appelle « *le problème primal* ».

- **Principe de FERMAT**

Les points qui minimisent ou maximisent une fonction dérivable et annulent sa dérivée sont appelés « *points stationnaires* ».

- **Principe de LAGRANGE**

Résoudre un problème d'optimisation revient à rechercher un point stationnaire z_0 du Lagrangien $L(z, \check{U})$ de la fonction à optimiser.

$$L(z, \alpha) = g(z) + \sum_{i=1}^m \alpha_i C_i^g(z) \quad (\text{V.7})$$

Tels que $g(z)$ la fonction à optimiser, les coefficients C_i^g sont les contraintes et α_i sont les coefficients de Lagrange.

L'introduction des coefficients de Lagrange transforme le « *problème primal* » en « *problème dual* ».

- **Principe de KUHN-TUCKER**

Il est possible de trouver un « *point selle* » $(z_0, \check{U})$ qui vérifie.

$$\min_z L(z, \alpha^*) = L(z_0, \alpha^*) = \max_{\alpha \geq 0} L(z_0, \alpha) \quad (\text{V.8})$$

En appliquant ce principe, nous sommes appelés à trouver un « *point selle* » $(w_0, b_0, \check{U})$, le Lagrangien correspondant à ce problème s'écrit comme tel :

$$L(w, b, \alpha) = \frac{1}{2} w \cdot w - \sum_{i=1}^m \alpha_i \cdot [y_i \cdot [(x_i \cdot w) + b] - 1] \quad (\text{V.9})$$

Le Lagrangien doit être minimal par rapport à w et b et maximal par rapport à $\check{U}$ (avec $\alpha \geq 0$).

En vertu du principe de KUHN-TUCKER, la condition pour que $\alpha^0 = (\alpha_1^0, \dots, \alpha_m^0)$ soit une solution optimale est :

$$\alpha_i^0 \cdot [y_i \cdot [(w_0 \cdot x_i) + b_0] - 1] = 0 \quad \text{pour } i = 1, \dots, m$$

C'est-à-dire que soit $\alpha_i^0 = 0$ ou alors $y_i \cdot [(w_0 \cdot x_i) + b_0] = 1$.

Prenons le cas où : $y_i \cdot [(w_0 \cdot x_i) + b_0] = 1$ et $\alpha_i > 0$.

En se servant des formules utilisées lors de la résolution du « problème primal », la variable

$$b_0 = -\frac{\max_{y_i=-1}(w_0 \cdot x_i) + \max_{y_i=1}(w_0 \cdot x_i)}{2}$$

$$b_0 = -\frac{1}{2} \cdot [(w_0 \cdot x^*(1)) + (w_0 \cdot x^*(-1))] \quad (\text{V.10})$$

Ce cas est par définition l'interprétation des « Vecteurs de Support ». Ces derniers peuvent être traduits par l'acronyme « SV », ils sont considérés comme les représentants de leurs catégories respectives et appartiennent aux hyperplans canoniques, leur nombre est en général très petit devant celui des données d'apprentissage.

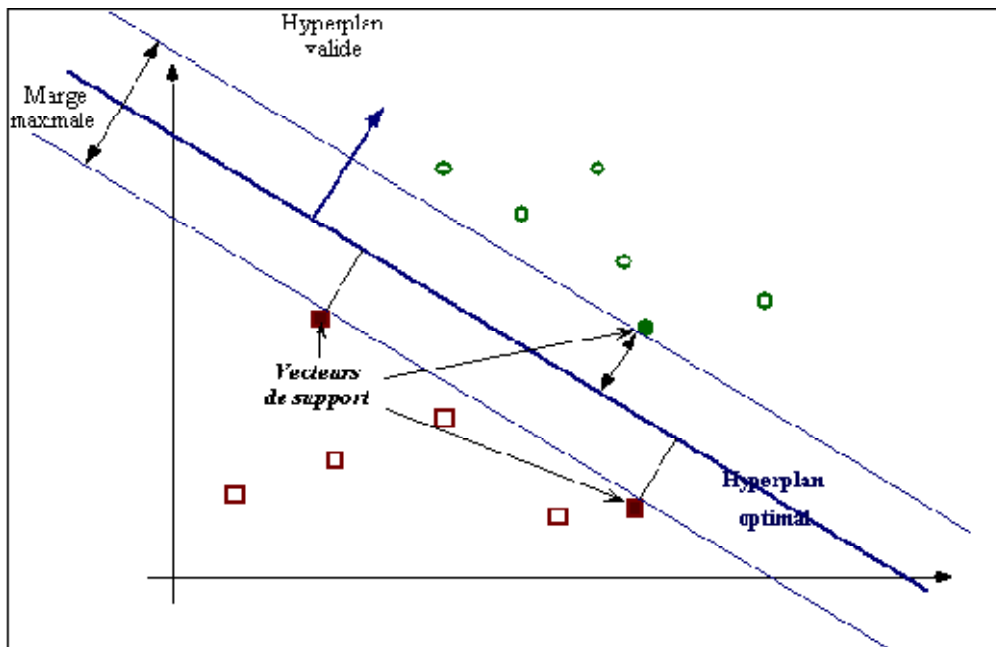


Figure V.4 : Hyperplan optimal et vecteurs de support.

La fonction de décision est déterminée alors par :

$$f(x) = \text{sign}[(w_0 \cdot x) + b_0] \quad (\text{V.11})$$

En remplaçant w_0 par l'équation ci-dessus, la fonction de décision devient.

$$f(x) = \text{sign} \left[\sum_{xi \in SV} \alpha_i^* \cdot y_i \cdot (x_i \cdot x) + b_0 \right] \quad (\text{V.12})$$

Si la fonction de décision est inférieure à 0 la classe de l'objet est « -1 », autrement il est de classe « +1 ».

Dans la réalité les problèmes de classifications ne sont pas toujours linéairement séparables.

- Cas non séparable (formulation à marge molle)

Dans le cas où le nuage de points n'est pas séparable, et où les données de deux classes distinctes sont raisonnablement distribuées (i.e : le problème de séparation réside uniquement dans certains points), le problème d'optimisation énoncé n'a pas de solutions, et donc les conditions KKT ne sont jamais satisfaites. L'idée consiste à accepter des erreurs d'apprentissage, tout en les pénalisant en introduisant la variable «C ». Des variables « ressort » (de relâchement) ξ_i sont introduites.

Le « problème primal » s'écrit dans ce cas :

$$\min \left[\frac{1}{2} \cdot \|w\|^2 + C \sum_{i=1}^m \xi_i \right] \quad (\text{V.13})$$

$$\text{Sous la contrainte} \quad y_i \cdot (w \cdot x_i + b) \geq 1 - \xi_i, \quad \forall i \quad (\text{V.14})$$

Le Lagrangien correspondant à ce problème d'optimisation s'écrit comme tel :

$$L(w, b, \zeta, \alpha, \mu) = \frac{1}{2} \|w\|^2 + C \sum_{i=1}^m \xi_i - \sum_{i=1}^m \alpha_i (y_i (w \cdot x_i + b) - 1 + \xi_i) - \sum_{i=1}^m \mu_i \xi_i \quad (\text{V.15})$$

Avec α et μ multiplicateurs de Lagrange.

où $\xi_i \geq 0$, $\alpha_i \geq 0$ et $\mu_i \geq 0$

Le Lagrangien doit alors être minimisé par rapport à w, b et ξ et maximisé par rapport à α et μ . Contrairement au cas précédent (cas linéairement séparable), la variable α_i a une borne supérieure égale à « C »,

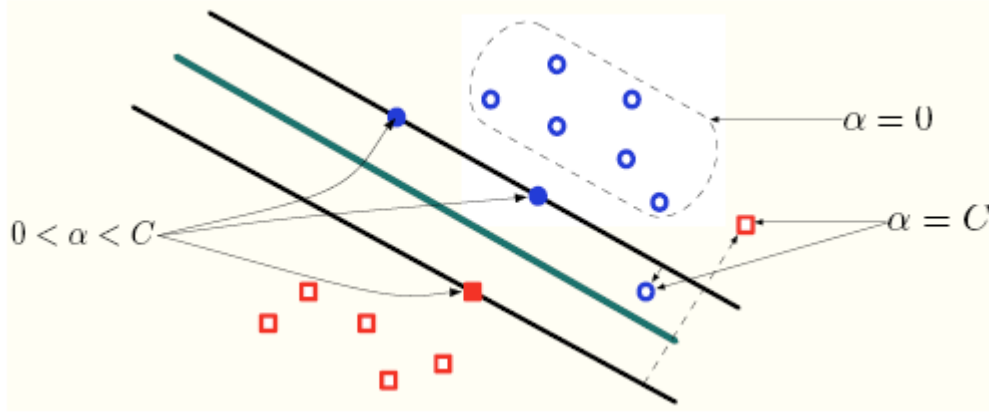


Figure V.5 : Les différentes valeurs de α .

-Cas non linéairement séparable

Dans ce cas, la séparation linéaire des données en classes n'est pas possible. La seule façon de contourner ce problème est de projeter les données dans un espace « déployé » H (qui est un espace de plus grande dimension) dans lequel les données deviennent linéairement séparables, en passant par une fonction $\Phi(x)$.

Telle que :

$$\Phi : R^d \rightarrow F$$

$$x \mapsto \Phi(x)$$

Le Lagrangien correspondant à ce problème d'optimisation est défini par :

$$L(w, b, \alpha) = \frac{1}{2} w \cdot w - \sum_{i=1}^m \alpha_i \cdot [y_i \cdot ((\Phi(x_i) \cdot w) + b) - 1] \quad (V.16)$$

Le Lagrangien doit être minimal par rapport à w et b et maximal par rapport à $\check{\alpha}$ (tel que $\alpha \geq 0$).

$$\text{La résolution de ce problème se fait par : } \begin{cases} \max \left[\sum_{i=1}^m \alpha_i - \frac{1}{2} \cdot \sum_{i,j=1}^m \alpha_i \alpha_j y_i y_j \langle \Phi(x_i) \cdot \Phi(x_j) \rangle \right] \\ \alpha_i \geq 0 \quad i = 1, \dots, m \\ \sum_{i=1}^m \alpha_i y_i = 0 \end{cases} \quad (V.17)$$

La fonction de décision se traduit donc par :

$$f(x) = \text{sign} \left[\sum_{xi \in SV} \alpha_i^* \cdot y_i \cdot (\Phi(x_i) \cdot \Phi(x)) + b \right] \quad (\text{V.18})$$

Nous avons exposé le cas où les deux classes sont linéairement séparables dans l'espace des descripteurs, alors qu'il existe un autre cas où les deux classes sont non séparables, où nous autorisons des erreurs d'apprentissage, tout en les pénalisant en introduisant la variable «C». La résolution de ce problème se fait alors par :

$$\left\{ \begin{array}{l} \max \left[\sum_{i=1}^m \alpha_i - \frac{1}{2} \cdot \sum_{i,j=1}^m \alpha_i \alpha_j y_i y_j \langle \Phi(x_i) \cdot \Phi(x_j) \rangle \right] \\ 0 \leq \alpha_i \leq C \quad i = 1, \dots, m \\ \sum_{i=1}^m \alpha_i y_i = 0 \end{array} \right. \quad (\text{V.19})$$

$$\text{Et la solution a la forme suivante : } f(x) = \text{sign} \left[\sum_{xi \in SV} \alpha_i^* \cdot y_i \cdot (\Phi(x_i) \cdot \Phi(x)) + b \right] \quad (\text{V.20})$$

La solution du problème ne dépend que du produit scalaire $(\Phi(x_i) \cdot \Phi(x))$, au lieu de choisir une transformation $\Phi : R^d \rightarrow F$. Nous choisissons une fonction $K : R^d \times R^d \rightarrow R$, appelée fonction noyau, cependant il faut que cette dernière satisfasse la condition de MERCER.

- **Condition de MERCER**

Pour toute fonction $K(x; z)$ symétrique et vérifiant l'intégrale $\int K(x, z) f(x) f(z) dx dz \geq 0$

Telle que f est une fonction non nulle L2-intégrable ($\int f(x)^2 dx$ finie).

Il existe une fonction Φ telle que : $K(x, z) = \Phi(x) \cdot \Phi(z) \quad \forall x, z$

- Les fonctions noyaux

Il existe dans la littérature des fonctions noyaux qui vérifient la condition de MERCER.

a-Noyau linéaire

La forme de ce noyau est :

$$K(x_1, x_2) = \langle x_1, x_2 \rangle \quad (\text{V.21})$$

b-Noyau polynomial

La forme générique de ce noyau est :

$$K(x_1, x_2) = (a \cdot \langle x_1, x_2 \rangle + b)^d \quad (\text{V.22})$$

Où a et b sont des constantes et d la puissance du produit scalaire.

c-Noyau RBF (Radial Basis Function)

Il est défini par :

$$K(x_1, x_2) = \exp\left(-\frac{\|x_1 - x_2\|^2}{2\sigma^2}\right) \quad (\text{V.23})$$

Où σ est appelé terme de régulation.

d-Noyau sigmoïde

Il est défini par :

$$K(x_1, x_2) = \tanh(a \cdot \langle x_1, x_2 \rangle + b) \quad (\text{V.24})$$

Où a et b sont des constantes.

e-Composition de Noyaux

Il est possible de composer de nouveaux noyaux en utilisant des noyaux existants, en prenant K_1 et K_2 des fonctions satisfaisant à la condition de Mercer.

$$K(x_i, x_j) = K_1(x_i, x_j) + K_2(x_i, x_j)$$

$$K(x_i, x_j) = \alpha \cdot K_1(x_i, x_j) \quad \text{avec } \alpha \in \mathbb{R}^+$$

$$K(x_i, x_j) = K_1(x_i, x_j) \cdot K_2(x_i, x_j)$$

V.5.4 SVM multiclass

À l'origine les SVMs ont été conçus pour résoudre des problèmes de classification binaire, néanmoins il existe des méthodes qui permettent de traiter des problèmes multiclass.

La première approche est l'approche « **un parmi K** » (one against all) qui comprend K sous-classifieurs, le i -ième sous-classifieur résout un problème de classification en séparant la i -ième classe de toutes les autres en la considérant comme étant la classe « $+1$ » et en

considérant les autres classes comme étant la classe « -1 ». L'inconvénient de cette approche est qu'elle peut trouver des régions de conflit (régions où deux classes sont choisies simultanément). Elle peut aussi trouver des régions d'ambiguïté (régions où aucune classe n'est choisie).

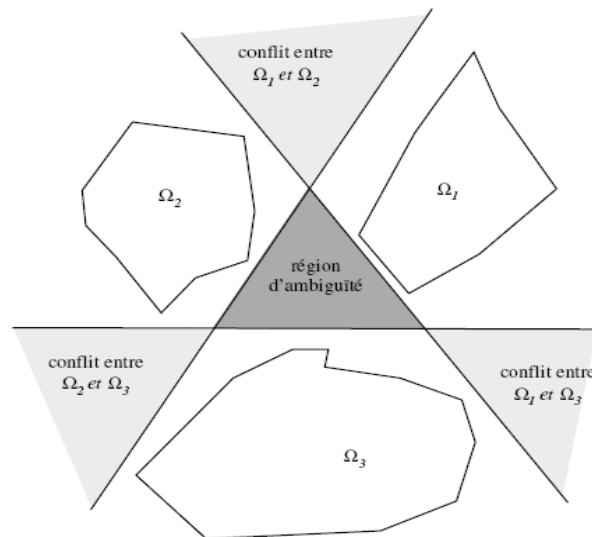


Figure V.6 : Classification « 1 parmi K ».

La seconde approche est l'approche « **un contre un** » (one against one), elle permet de tracer des séparations entre les classes prises deux à deux en construisant $\frac{K(K-1)}{2}$ sous-classifieurs, généralement la décision revient à faire une opération de « vote » entre les différents sous-classifieurs. Cette approche se calcule plus lentement que la première car elle comporte beaucoup plus de sous-classifieurs, cependant les zones de conflit sont moins importantes.

La troisième approche est celle des « **arbres de décision** », cette méthode peut se faire de deux façons différentes :

- **Top-down binary tree** : initialement un groupe contient toutes les classes, un processus récursif permet de réduire la taille de ce groupe en vérifiant l'appartenance des données à une classe à chaque itération, la classe de l'objet est connue une fois que toutes les itérations sont effectuées.
- **Bottom-up binary tree** : un algorithme récursif permet de comparer entre chaque paire de classes, seulement la classe qui vérifie le plus le test est retenue jusqu'à ce que toutes les itérations soient effectuées.

V.5.5 SVM Multi-Labels

Dans les problèmes de classification automatique classiques, les classes sont mutuellement exclusives par définition. Cependant, dans différentes applications (diagnostic médical, classification de documents écrits), il semble naturel de considérer qu'un même élément puisse appartenir simultanément à plusieurs classes. Il en va de même pour la classification en genres de signaux musicaux : ainsi un même morceau musical peut appartenir à plusieurs genres différents suivant la taxonomie utilisée et le contexte culturel [21].

De fait, les méthodes single-label utilisent une base d'apprentissage dont chaque exemple est associé à un seul label l appartenant à l'ensemble L des labels, $|L| > 1$. Si $|L| = 2$, la classification est dite binaire, alors que si $|L| > 2$, elle est dite multi-classe et la base d'exemples est associée à un ensemble de labels $Y \subseteq L$. Généralement, un problème de classification est formalisé comme suit : Etant donné un ensemble d'apprentissage sous la forme $\langle x_i, f(x_i) \rangle$, le problème de classification est ramené à un problème d'apprentissage de la fonction d'approximation $f(x)$. Pour une classification multi-label, la fonction d'approximation peut prendre différentes valeurs appartenant à l'ensemble « labels des classes ». Les classes ne sont pas mutuellement exclusives et peuvent se recouvrir par définition.

- Description et mesures d'évaluation

Il est alors possible de formaliser le problème comme suit : Soit alors l'ensemble d'apprentissage χ et l'ensemble des labels $Y = \{1, \dots, k\}$; Etant donné un ensemble d'exemples de la forme $\langle x_i, Y_i \rangle$, $x_i \in \chi$, $Y_i \in 2^{|Y|}$, l'enjeu est d'apprendre une fonction d'approximation $f(x)$ qui prend des valeurs uniques dans $2^{|Y|}$ avec une faible erreur. Il semble cependant difficile de définir la notion d'erreur pour une classification multi-label. Dans la plupart des cas l'approche multi-label induit un ordonnancement des labels possibles pour une instance donnée ; L'algorithme d'apprentissage pouvant alors être décrit par une fonction $f : \chi \times Y \rightarrow R$ qui permet de classer les labels suivant les valeurs de $f(x, \cdot)$. On définit alors $\text{rang}_f(x, l)$ comme le rang du label l pour l'instance x sous f . f étant une fonction bijective sur $\{1, \dots, k\}$. De plus si $f(x, l_1) \leq f(x, l_2)$ alors $\text{rang}_f(x, l_1) \leq \text{rang}_f(x, l_2)$. Soit alors $T = \langle (x_1, Y_1), \dots, (x_m, Y_m) \rangle$, l'ensemble de test.

- Mesures d'évaluation

La classification Multi-label nécessite des métriques d'évaluation différentes de celles des classifieurs classiques. Soit un ensemble D de données Multi-label constitué de $|D|$ exemples Multi-label tels que :

$$(x_i, Y_i), i=1 \dots |D|, Y_i \subseteq L.$$

Et soit H le classifieur Multi-label avec $Z_i = H(x_i)$ l'ensemble des labels prédits par H pour l'exemple x_i . Les principales métriques d'évaluation utilisées dans la littérature sont explicitées ci-dessous.

a- Hamming Loss

Cette mesure tient compte des erreurs de prédiction (un faux label a été prédit), ainsi que des erreurs d'omission (un label n'est pas prédit). Soit alors la fonction d'apprentissage $f: X \rightarrow 2^{Y_b}$, la mesure de Hamming de f sur l'ensemble de test T .

b- Coverage

Il est parfois souhaitable de tester les performances d'un système avec tous les labels possibles pour chaque échantillon. La mesure coverage est définie comme la distance moyenne qui permet d'atteindre (couvrir) tous les labels possibles assignés à l'échantillon x .

c- Average Precision

C'est une mesure qui permet d'évaluer les algorithmes de façon globale. Elle mesure la fraction moyenne de labels classés au dessus d'un label particulier $l \in Y_i$ qui appartient effectivement à Y_i .

d- Accuracy

Elle est directement reprise à partir des définitions utilisées en reconnaissance de forme classique. Si Y_i est l'ensemble des vrais labels associés à l'échantillon x_i et f_i celui des labels prédits.

• Modèles d'apprentissage Multi-label

Pour la classification multi label, la phase d'apprentissage est primordiale. Comment donc utiliser des exemples à labels multiples pour la phase d'apprentissage ? Plusieurs modèles coexistent ;

a- Single-label model (modèle-s)

Les données sont labelisées avec le label de la classe à laquelle ils appartiennent de façon la plus proche en suivant un critère subjectif.

b- Ignore-label model (modèle-i)

Toutes les données multi-label sont ignorées lors de la phase d'apprentissage.

c- New-class model (model-n)

Il est également possible de considérer les données multi-label comme une nouvelle classe et donc de leur associer un nouveau label. L'inconvénient de cette approche étant la disparité des données multi-label qui souvent rend difficile la construction de modèles cohérents.

d- Cross-training model (model-x)

Les données multi-label peuvent également être utilisées plusieurs fois lors de l'apprentissage. Elles sont utilisées pour chaque classe dont elles possèdent le label. L'apprentissage apparaît alors comme plus efficace que celui établi pour les autres modèles car il contient plus d'exemples pour chaque classe.

- *Critères de test Multi-label*

Nous traiterons dans ce qui suit de la sortie de la machine d'apprentissage ; Comment obtenir les labels multiples à partir de la sortie des modèles de base pour chaque classe. Puisque nous allons utiliser comme machine d'apprentissage les Machines à Vecteurs Support (SVM), nous traiterons ici uniquement du cas de ce classifieur. Le SVM bi-classe classique, labellise l'exemple test là où le score du SVM est positif. Dans l'approche « un contre tous », un classifieur est entraîné pour chaque classe et chaque classifieur attribue un score à l'exemple test. L'exemple est ensuite labellisé comme appartenant à une classe unique : celle qui correspond au SVM produisant le score maximum (même si plusieurs scores sont positifs).

Une généralisation de l'approche « un contre tous » pour la classification multi-label, permet de dégager les trois critères de test Multi-label suivants :

a- Critère-P

Les exemples tests sont labellisés avec le label de chaque classe ayant obtenu un score positif (P indique positif). Si aucun score n'est positif, les données seront classées dans une classe « inconnu ».

b- Critère-T

Ce critère suppose que toutes les données doivent appartenir à au moins une des classes existantes. Il reste similaire au critère P pour l'attribution des labels lorsque les scores sont positifs. Si les scores des N classes sont tous négatifs alors il est attribué à la donnée le label correspondant au score le moins négatif (Top score).

c- Critère-C

La décision d'attribution du label dépend de la proximité entre les M Top scores. Ainsi, si les M Top scores sont assez proches, alors les labels des classes correspondantes sont attribués à l'exemple considéré. Le principe du maximum a posteriori (MAP) peut être utilisé pour décider du seuil à partir duquel les scores seraient considérés comme suffisamment proches.

Conclusions

Nous avons détaillé dans ce chapitre les méthodes de classification de signaux constituant la base théorique des algorithmes développés dans le cadre de la classification en genres de signaux musicaux.

Les méthodes à Vecteurs de Support ont été particulièrement développées car elles constituent notamment couplés avec les descripteurs ondelettes un système de classification robuste et performant pour le traitement multiclassés des signaux musicaux. Nous montrons dans le chapitre résultats les performances de cette approche.

Nous avons également exploité une nouvelle piste novatrice en classification de signaux audio, à savoir la classification en étiquettes multiples (Multi-Labels), les résultats préliminaires développés et commentés dans le chapitre suivant montrent l'efficacité de cette voie et justifient les promesses qui en sont attendues. Cette tâche doit être poursuivie et constitue une de nos perspectives.

Références Bibliographiques du Chapitre

- [1] Duda, R., Hart, P. E., et Stork, *Pattern Classification (2nd edition)*. John Wiley & Sons, 2000
- [2] Fritz, M., Leibe, B., Caputo, B., et Schiele, B. *Integrating representative and discriminative models for object category detection*. Dans Proc. ICCV.2005
- [3] Louradour J, *Noyaux de séquence pour la vérification du locuteur par Machines à Vecteurs de Support*. Thèse de Doctorat de l'Université de Toulouse III, IRIT, Toulouse 2007
- [4] Arias, J., Pinquier, J., et André-Obrecht, R. , *Evaluation of classification techniques for audio indexing*. Dans Proc. EUSIPCO, 2005
- [5] Bouchard, G. . *Les modèles génératifs en classification supervisée et applications à la catégorisation d'images et à la fiabilité industrielle*. Thèse de doctorat, Université Joseph Fourier (Grenoble I)
- [6] Burges, C. , *A tutorial on support vector machines for pattern recognition*. In Data Mining and Knowledge Discovery, 2(2) :955ï 974. 1998
- [7] Campbell, W., Reynolds, D., et Campbell, J. , *Fusing discriminative and generative methods for speaker recognition : Experiments on switchboard and nfi/tno field data*. Dans Proc. IEEE Odyssey. 2004
- [8] Aucouturier J.J., *Ten Experiments on the Modelling of Polyphonic Timbre*. PhD thesis, Laboratoire d'Informatique de Paris 6, France, June 2006.
- [9] Balkema W., *Variable-size gaussian mixture models for music similarity measures*. In Proceedings of the 8th International Conference on Digital Audio Effects (DAFx), pages 491ï 494, September 2007.
- [10] Pinquier J., *Indexation Sonore: recherche de composantes primaires pour une structuration audiovisuelle*, PhD Thesis, Université Paul Sabatier , Toulouse III. (2004).
- [11] Ulusay I., Bishop M. C., *Comparison of Generative and Discriminative Techniques for Object Detection and Classification in* LNCS4170 Springer-Verlag. (2006)

- [12] Tzanetakis G. and Cook P.R., *Musical genre classification of audio signals*. IEEE Transactions on Speech and Audio Processing, 10(5), July 2002.
- [13] Harb H., *Classification du signal sonore en vue d'une indexation par le contenu des documents multimédia*, Thèse de doctorat de l'école centrale de Lyon . septembre 2004 .
- [14] Essid S., *Classification automatique de signaux audio-fréquence : reconnaissance des instruments de musique* — PhD Thesis, Université Pierre et Marie Curie . (2005).
- [15] Reed J. and Lee C., *A study on music genre classification based on universal acoustic models*. In Proceedings of the 7th International Conference on Music Information Retrieval, pages 89–94, October 2006.
- [16] Pampalk E., Flexer A., and Widmer G., *Improvements of audio-based music similarity and genre classification*. In Proceedings of the 6th International Conference on Music Information Retrieval, pages 628–633, October 2005.
- [17] E. Pampalk E., *Islands of Music: Analysis, Organization, and Visualization of Music Archives*. Master's thesis, Vienna University of Technology, Austria, December 2001.
- [18] Pachet F., Roy P., and Cazaly D., *A combinatorial approach to content based music selection*. IEEE Multimedia, 7(1):44–51, March 2000.
- [19] F. Pachet and D. Cazaly. *A taxonomy of musical genres*. In Content- Based Multimedia Information Access Conference (RIAO) proceedings, 2000.
- [20] Balkema W., *Design and Realisation of an Efficient Content Based Music Playlist Generation System* PhD Thesis TU Delft/Universiteit Twente, 2007
- [21] Carvalho A.CPLF and Freitas A. A., *A Tutorial on Multi-Label Classification Techniques* in
- [22] Rabiner, L. *A tutorial on hidden markov models and selected applications in speech recognition*. Proc. of the IEEE, 77(2) :257–286, 1989
- [23] Shakhnarovich, Darrell, et Indyk, editors, *Nearest-Neighbor Methods in Learning and Vision*. MIT Press.
- [24] Ganchev, T., Tasoulis, D., Vrahatis, M., et Fakotakis, N. , *Locally recurrent probabilistic neural network for text-independent speaker verification*. Dans Proc. Eurospeech. 2003
- [25] Lippens S., Martens J., Leman M., Baets B., Meyer H., and Tzanetakis G., *A comparison of human and automatic musical genre classification*. In Proceedings of the IEEE International Conference on Audio, Speech and Signal Processing (ICASSP), volume 4, pages 233–236, May 2004.
- [26] Aucouturier J.J. and Pachet F., *Music similarity measures: What is the use?* In Proceedings of the 3rd International Conference on Music Information Retrieval, pages 157–163, October 2002.

- [27] Flexer A., Pampalk E., and Widmer G., *Hidden markov models for spectral similarity of songs*. In Proceedings of the 8th International Conference on Digital Audio Effects (DAFx), Madrid, Spain, September 2005.
- [28] Mitri G., Uitdenbogerd A. L., and Ciesielski V., *Automatic music classification problems*. In Proceedings of the 27th conference on Australasian computer science, volume 26, pages 315–322, 2004.
- [29] Mandel M.I. and Ellis D., *Song-level features and support vector machines for music classification*. In Proceedings of the 6th International Conference on Music Information Retrieval (ISMIR 2005), pages 594–599, October 2005.
- [30] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals", In IEEE Trans. On Speech and Audio Processing, Vol. 10, No. 5, pp. 293-302, 2002.
- [31] K. West and S. Cox, "Features and classifiers for the automatic classification of musical audio signals", In Proceedings of the 5th International Conference on Music Information Retrieval, Barcelona, October 2004.
- [32] T. Lambrou, P. Kudumakis and all, "Classification of audio signals using statistical features on time and wavelet transform domains", In Proceedings of IEEE International Conference on acoustics, speech and signal processing, Vol.6, pp. 3621-3624, USA, May 1998.
- [33] J.J. Aucouturier and F. Pachet, "Musical genre: a survey", Journal of new music research, 32(1), 2003.
- [34] M.R. Boutell, J. Luo, X. Shen and C.M. Brown, "Learning multi-label scene classification", *Pattern Recognition*, 37(9):1757-1771, 2004.
- [35] K. Brinker, J. Furnkranz and E. Hullermeier, "A unified model for multilabel classification and ranking", In *Proceedings of the 17th European Conference on Artificial Intelligence (ECAI '06)*, pages 489–493, Riva del Garda, Italy, Aug/Sept 2006.

CHAPITRE 6

Résultats et évaluations de performances
--

INTRODUCTION

Avant de décrire en détail les différentes expériences et simulations ayant permis de valider les approches théoriques, il convient de rappeler que la chaîne de traitements pour l'indexation de signaux audio que nous proposons, comprend successivement les modules suivants : les signaux audio numérisés sont convertis en vecteurs acoustiques grâce aux transformations décrites dans le chapitre dédié aux descripteurs ; Ces vecteurs acoustiques comprennent diverses classes sonores qu'il convient de séparer dans le temps : c'est la tâche réalisée par le module de segmentation musique/non musique. Les données récupérées contiennent alors des mélanges sonores de musique instrumentale de divers genres et de la voix chantée (parole) que nous pouvons séparer au moyen des algorithmes de séparation de sources. Les données contenant de la musique instrumentale pure passent à travers le module de classification en genres. Nous comparons alors plusieurs méthodes de classification. Il convient de préciser que nous avons utilisé la même base de données audio pour les toutes les tâches qui succèdent au module de segmentation musique/non musique. En conséquence, nous avons divisé nos résultats en trois entités : La segmentation de signaux audio, La séparation de sources sonores et enfin La classification en genres musicaux.

VI.1 SUR LA SEGMENTATION DE SIGNAUX AUDIO

VI.1.1 La Base de Données Audio «KASSEM»

Afin d'évaluer notre méthode nous avons construit une base de données de signaux contenant diverses composantes de mélanges sonores (parole + musique + bruit+ silence) divisée en trois parties : Une base de référence contenant les fichiers sonores exemples, une base de fichiers dédiée au développement de la méthode et finalement une base de signaux test afin d'estimer objectivement les performances. Leur description est détaillée dans le tableau VI.1:

Tableau VI.1 : Description de la base de données

Fichier	Durée (s)	%M	%P	%S	%B
Mref.wav	-	100	0	0	0
Pref	-	0	100	0	0
Sref	-	0	0	100	0
Bref	-	0	0	0	100
Dev1	300	35	50	5	10
Dev2	600	50	35	10	5
Dev3	900	0	90	5	5
Dev4	1200	90	0	5	5
Test1	1800	25	30	20	25
Test2	2700	50	20	15	15

- Le Critère de Performance

Le critère de performance adopté est la mesure F définie comme étant la combinaison de deux métriques appelés Précision et Rappel empruntées à la RI (Information Retrieval) [21,22] : Le Rappel mesure la capacité d'un système à sélectionner les hypothèses pertinentes :

$$\text{Rappel}_i = \frac{\text{Trames correctement attribuées à la classe } i}{\text{Nombre de trames appartenant à la classe } i}$$

La précision mesure la capacité du système à rejeter les hypothèses non pertinentes :

$$\text{Précision}_i = \frac{\text{Trames correctement attribuées à la classe } i}{\text{Nombre de trames attribuées à la classe } i}$$

La combinaison harmonique de ces deux métriques donne F mesure:

$$F = \frac{2 * \text{précision} * \text{rappel}}{\text{précision} + \text{rappel}} \quad (\text{VI.1})$$

Les signaux sonores sont échantillonnés à 35Khz et la paramétrisation est effectuée par l'outil HTK Tools. Nous avons choisi d'utiliser les coefficients MFCC évalués au centi-seconde (fenêtre de Hamming de 32 ms) et dont le nombre est variable (Tableaux VI.4, VI.5, VI.6, VI.7). Dans ce qui suit nous allons présenter les différentes expériences réalisées afin de quantifier les performances de notre méthode. La décision est prise trame par trame. Nous utiliserons les notations suivantes : ND pour Non Détection, PD pour Pas de Détection, PP pour Pas de Parole et PM pour Pas de Musique.

VI.1.2 Expériences sur les signaux de développement

- Influence de la normalisation de l'énergie

Ci-dessous sont présentés les résultats pour des fichiers sans normalisation de l'énergie (Tableau VI.2), puis avec normalisation de l'énergie (Tableau VI.3) :

Tableau VI.2. Données avec énergies non normalisées

F.mes Fichier	Bruit	Silence	Musique	Parole
Dev1	83.22%	92.77%	76.32%	82.32%
Dev2	83.22%	95.16%	72.16%	87.12%
Dev3	79.34%	87.13%	Pas de Musique	89.67%
Dev4	87.32%	86.36%	91.17%	Pas de Parole.

Tableau VI.3. Données avec énergies normalisées

F mes. Fichier	Bruit	Silence	Musique	Parole
Dev1	93.55%	96.77%	97.17%	98.32%
Dev2	93.55%	95.16%	95.64%	95.24%
Dev3	85%	87.13%	P.M.	99.07%
Dev4	94.49%	86.36%	98.83%	P.P.

Nous observons que les résultats s'améliorent sensiblement avec la normalisation de l'énergie.

- *Influence de la dimension des descripteurs*

Dans l'expérience qui suit, nous faisons varier la dimension des descripteurs MFCC et nous observons l'évolution de F mesure correspondante à la détection de chaque classe. Les tableaux qui suivent (IV.4, IV.5, IV.6 et IV.7) illustrent les pourcentages de segmentation obtenus pour différents fichiers de développement :

Tableau VI.4 : Influence de la dimension des descripteurs : F mesure (Classe Bruit)

Dim. Fichier	8	12	16	20	24	30	35	40
Dev1	36.25	93.55	93.55	93.55	36.25	34.52	93.55	93.55
Dev2	16.57	93.55	93.55	93.55	16.57	16.20	93.55	93.55
Dev3	85	85	85	85	85	70.83	85	85
Dev4	10.25	94.49	94.49	10.25	10.32	10.32	94.49	94.49

Tableau VI.5 : Influence de la dimension des descripteurs : F mesure (Classe Silence)

Dim. Fichier	8	12	16	20	24	30	35	40
Dev1	16.13	ND	96.77	96.77	19.87	25.21	96.77	ND
Dev2	35.19	8.82	95.16	95.16	40.27	47.97	95.16	8.82
Dev3	9.75	17.24	87.13	87.13	9.75	10.01	87.13	17.24
Dev4	86.36	15.79	87.10	87.10	86.36	86.36	86.36	15.79

Tableau VI.6 : Influence de la dimension descripteurs : F mesure (Classe Musique)

Dim. Fichier	8	12	16	20	24	30	35	40
Dev1	ND	97.17	97.17	97.17	ND	ND	97.17	97.17
Dev2	ND	95.64	95.64	94.22	ND	ND	95.64	95.64
Dev3	PM	PM	PM	PM	PM	PM	PM	PM
Dev4	ND	98.83	98.83	98.83	98.8	1.47	98.83	98.83

Tableau VI.7 : Influence de la dimension descripteurs : F mesure (Classe Parole)

Dim. Fichier	8	12	16	20	24	30	35	40
Dev1	ND	98.32	98.32	98.32	37.84	56.46	98.32	98.32
Dev2	ND	95.24	95.24	93.46	36.29	57.72	95.24	95.24
Dev3	1.22	98.7	98.7	99.22	1.22	3.14	99.07	98.19

- Influence de la durée des fichiers de référence*

Dans ces expériences nous varions la durée des fichiers de référence et observons l'évolution de la F mesure correspondant à chaque classe (Tableaux VI.8, VI.9, VI.10 et VI.11). Nous rappelons que les fichiers de référence sont considérés comme des échantillons de classes sonores et qu'il est souhaitable que leur taille soit la plus petite possible afin de ne pas peser sur les charges de calculs de l'algorithme.

Tableau VI.8 : Influence de la durée des fichiers de référence : F mesure (Classe Bruit)

Durée(s) Fichier	10	20	30	40	50
Dev1	93.55	93.55	36.25	34.52	93.55
Dev2	93.55	93.55	16.96	16.20	93.55
Dev3	85	85	85	65.38	85
Dev4	94.49	94.49	10.32	10.25	94.49

Tableau VI.9 : Influence de la durée des fichiers de référence : F mesure (Classe Silence)

Durée (s) \ Fichier	10	20	30	40	50
Dev1	96.77	96.77	16.13	16.85	96.77
Dev2	95.16	95.16	36.04	37.85	95.16
Dev3	87.13	87.13	9.75	10.01	87.13
Dev4	87.10	87.10	81.43	86.36	87.10

Tableau VI.10 : Influence de la durée des fichiers de référence : F mesure (Classe Musique)

Durée(s) \ Fichier	10	20	30	40	50
Dev1	93.64	97.17	ND	ND	97.17
Dev2	91.94	94.22	ND	ND	94.37
Dev3	PM	PM	PM	PM	PM
Dev4	98.83	98.83	ND	ND	98.83

Tableau VI.11 : Influence de la durée des fichiers de référence : F mesure (Classe Parole)

Durée (s) \ Fichier	10	20	30	40	50
Dev1	95.50	98.32	ND	ND	98.32
Dev2	88.89	93.46	14.10	14.10	93.20
Dev3	87.54	87.70	1.22	1.22	96.12
Dev4	PP	PP	PP	PP	PP

- *Influence de la taille de la fenêtre de lissage*

Suite à l'étape de classification un fichier de labels est créé. Il s'agit de regrouper d'abord les labels identiques adjacents, ensuite d'appliquer une fenêtre de lissage glissante dans le temps afin d'éliminer les segments issus de l'assemblage et qui sont non significatifs, c'est-à-dire dont la durée est inférieure à une durée minimale critique. Dans les expériences suivantes

(Tableaux VI.12, VI.13, VI.14 et VI.15), nous faisons varier la taille de la fenêtre de lissage et nous relevons la valeur de F mesure correspondant à la détection de chaque classe. L'objectif est d'ajuster ce paramètre à la valeur minimale qui donne un maximum de F mesure selon la classe considérée.

Tableau VI.12 : Influence de la taille de la fenêtre de lissage : F mesure (Classe Bruit)

Fichier \ Taille(s)	1	3	5	8
Dev1	PD	96.67	96.67	93.55
Dev2	PD	96.67	96.67	93.55
Dev3	PD	90.24	89.41	85
Dev4	PD	97.56	97.56	94.49

Tableau VI.13 : Influence de la taille de la fenêtre de lissage : F mesure (Classe Silence)

Fichier \ Taille(s)	1	3	5	8
Dev1	PD	93.33	93.33	96.77
Dev2	PD	100	98.33	95.16
Dev3	PD	91.49	91.49	87.13
Dev4	PD	96.72	96.67	87.10

Tableau VI.14 : Influence de la taille de la fenêtre de lissage : F mesure (Classe Musique)

Fichier \ Taille(s)	1	3	5	8
Dev1	PD	PD	97.14	97.17
Dev2	PD	PD	97.33	95.64
Dev3	PM	PM	PM	PM
Dev4	PD	PD	99.44	98.83

Tableau VI.15 : Influence de la taille de la fenêtre de lissage : F mesure (Classe Parole)

Taille(s) Fichier	1	3	5	8
Dev1	PD	PD	98.01	98.32
Dev2	PD	PD	96.21	95.24
Dev3	PD	PD	98.44	98.7
Dev4	PP	PP	PP	PP

Les expériences de développement qui précèdent nous permettent de conclure qu'une paramétrisation avec 16 coefficients MFCC et énergie normalisée est celle qui donne les meilleures performances quant à la discrimination des classes (F mesure élevée) pour l'ensemble des fichiers de développement. Un mauvais choix du type de paramétrisation et/ou du nombre de paramètres peut amener à ne pas détecter la classe ciblée (ND : Non Détection). D'autre part, une taille moyenne de 10 ou 20 secondes des fichiers de référence ainsi qu'une fenêtre de lissage de 5 secondes donnent des performances acceptables. En effet, nous pouvons remarquer que des longueurs de fenêtre de filtrage inférieures à cette valeur critique ne permettent pas la détection d'une classe (PD : Pas de Détection, PP : Pas de Parole et PM ; Pas de Musique). Les expériences sur les fichiers de développement ont permis de mettre en évidence les paramètres optimaux de la méthode pour les fichiers sonores de la base de données considérée : 16 coefficients MFCC avec énergie normalisée, une taille des fichiers de référence de 10 secondes et une taille de la fenêtre de lissage de 5 secondes.

VI.1.3 Expériences sur les signaux de Test

Dans cette partie nous considérons deux fichiers sonores de taille et complexité plus importantes par rapport aux fichiers de développement. Ces fichiers serviront à la validation de la méthode en tenant compte de l'ajustement des paramètres effectué lors de la phase de développement. Le critère de performance adopté dans ce cas est la matrice de confusion. Les tableaux ci-dessous (Tableaux VI.16, VI.17) traduisent les résultats de la méthode pour le fichier Test1.wav avec les paramètres suivants : (16 Coefficients MFCC avec normalisation d'énergie ; la durée des fichiers de référence est de 20 secondes ; la taille de la fenêtre de lissage est de 5 secondes avec un recouvrement de 50%). Ces résultats confirment clairement la segmentation correcte des classes sonores.

Tableau VI.16 : Matrice de confusion pour le fichier Test1.wav

Classe Estimée Classe réelle	Bruit	Silence	Musique	Parole
Bruit	89.41	2.83	3.53	4.23
Silence	2.66	94.70	2.64	0
Musique	1.98	0.57	96.23	1.12
Parole	2.65	0	2.23	95.12

Tableau VI.17 : Matrice de confusion pour le fichier Test2.wav

Classe Estimée Classe réelle	Bruit	Silence	Musique	Parole
Bruit	88.36	3.33	4.03	4.23
Silence	3.26	94.02	2.74	0
Musique	0.98	0.57	97.23	1.12
Parole	2.77	0	3.23	94.00

VI.2 SUR LA SEPARATION DE SOURCES AUDIO

Afin d'évaluer nos méthodes de séparation de sources (EMD et ACI), nous avons utilisé des fichiers audio de la base de données « Kassem ». Ces signaux sont des mélanges musique, parole et bruit. Les méthodes de séparation de sources vont effectuer la séparation musique/parole. La parole représentant dans ce cas la voix chantée. Deux méthodes de séparation de sources ont été explorées : La séparation de sources par Décomposition Empirique Modale (EMD) et la séparation de sources par Analyse en Composantes Indépendantes (ACI). Les sources représentant bien évidemment la musique et la parole et éventuellement le bruit. Comment évaluer la « qualité de séparation » de chacune des méthodes ? Nous proposons les critères objectifs d'évaluation : le DSR (Distorsion to Signal Ratio), et l'ISR (Interference to Signal Ratio).

VI.2.1 Application des méthodes ACI/FastICA & EMD à différents mélanges et différentes matrices de mélange

Nous avons choisi trois types de mélanges :

- Le premier mélange est constitué de deux signaux de parole.
- Le deuxième est un mélange de parole et de bruit.
- Le troisième mélange contient du bruit et de la musique.

Les signaux sont des signaux réels (enregistrés). Les mélanges réalisés sont des mélanges synthétiques, linéaires et instantanés. La matrice de mélange A est générée de façon aléatoire par programmation. Le but étant d'étudier l'influence de cette dernière ainsi que l'influence du type de mélange sur la qualité de séparation de chaque méthode, en étudiant l'évolution des deux critères DSR et ISR exprimés par les deux équations (VI.2) et (VI.3) :

$$DSR = \frac{\|\check{\mathbf{S}}_k\|^2 - \left| \left\langle \check{\mathbf{S}}_k, \frac{\mathbf{s}_k}{\|\mathbf{s}_k\|} \right\rangle \right|^2}{\left| \left\langle \check{\mathbf{S}}_k, \frac{\mathbf{s}_k}{\|\mathbf{s}_k\|} \right\rangle \right|^2} \quad (VI.2)$$

$$ISR = \frac{\|e_{interf}\|^2}{\left\| \left\langle \check{\mathbf{S}}_k, \frac{\mathbf{s}_k}{\|\mathbf{s}_k\|} \right\rangle \frac{\mathbf{s}_k}{\|\mathbf{s}_k\|} \right\|^2} \quad (VI.3)$$

Tel que :

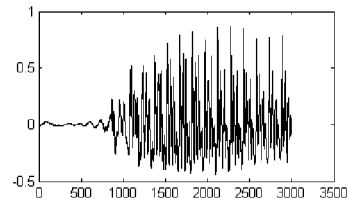
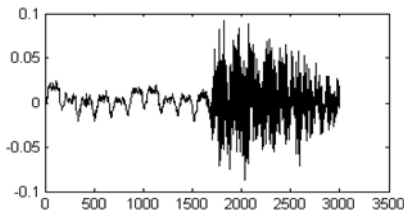
$$e_{interf} = \sum_{l \neq k} \left\langle \check{\mathbf{S}}_k, \frac{\mathbf{s}_l}{\|\mathbf{s}_l\|} \right\rangle \frac{\mathbf{s}_l}{\|\mathbf{s}_l\|} \quad \check{\mathbf{S}} \text{ représente le signal restitué et } \mathbf{s} \text{ le}$$

signal original.

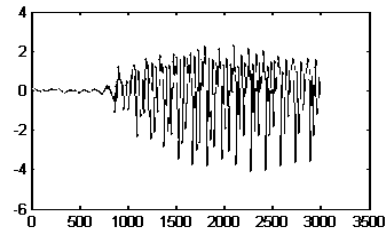
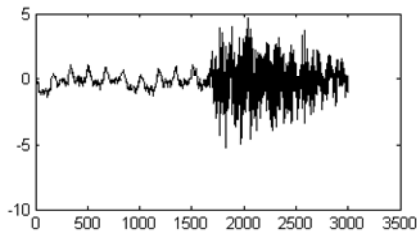
• Evaluation subjective de la séparation de sources

Les Figures VI.1.a et VI.1.b représentent deux signaux réels, respectivement un signal de parole et un signal de musique. Les Figures VI.2.a et VI.2.b représentent deux mélanges synthétiques des deux signaux précédents. Ils sont obtenus grâce à des matrices de mélange générées aléatoirement. Enfin, les Figures VI.3.a et VI.3.b illustrent la séparation de sources par EMD : La Figure VI.3.a représente le signal de parole estimé après EMD et la Figure VI.3.b le signal de musique estimé. Les Figures VI.4.a et VI.4.b illustrent la séparation de sources par FastICA : La Figure VI.4.a représente le signal de parole estimé après EMD et la Figure VI.4.b le signal de musique estimé. Visuellement, donc subjectivement, les signaux

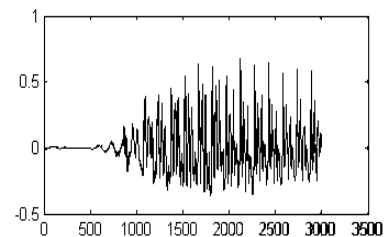
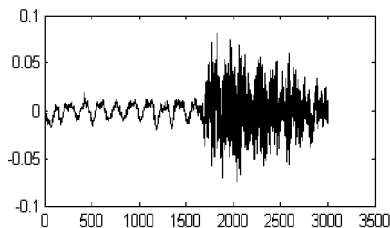
originaux sont identiques aux signaux estimés, à un facteur d'échelle et une inversion de phase près pour la méthode FastICA . De même des tests d'écoute ont confirmé cette concordance. Aucune différence notable entre les deux méthodes n'est directement observable.



Figures VI.1.a et VI.1.b : Sources Originales (parole et musique)



Figures VI.2.a et VI.2.b : Sources Estimées par FastICA



Figures VI.3.a et VI.3.b: Sources estimées par EMD

- **Evaluation objective de la séparation de sources**

Nous récapitulons dans le tableau VI.18, les valeurs des DSR et ISR pour chacun des deux algorithmes FastICA et EMD et ce pour différentes nature de mélange et différentes matrices de mélanges. Plusieurs constatations peuvent être établies à partir du tableau précédent :

- La méthode FastICA, quelle que soit la nature du mélange ou la matrice de mélange, présente un DSR constant pour les deux sources et de valeur nettement inférieure à

celui calculé pour la méthode EMD. La même conclusion peut être retenue pour le critère ISR.

- L'évaluation objective est cohérente avec celle subjective. Elle peut donc la remplacer et de fait automatiser le processus d'évaluation de la qualité de séparation.

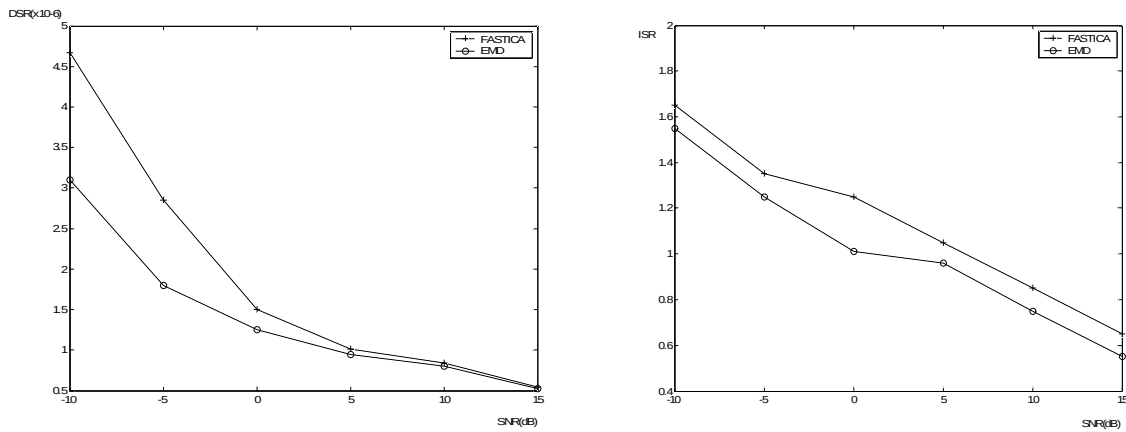
Puisque nous avons montré, dans le chapitre IV, que la séparation était d'autant meilleure que les deux critères DSR et ISR tendaient vers zéro, nous pouvons conclure que la méthode FastICA sépare mieux les signaux que la méthode EMD. Au fait, ce résultat était prévisible puisque l'ICA s'applique essentiellement aux mélanges linéaires et que l'EMD pourrait éventuellement apporter une solution à la séparation de mélanges convolutifs. Dans ce cadre, un point important doit être relevé : de quelle nature sont les enregistrements utilisés? Deux cas peuvent se présenter : Des enregistrements en « Live » réalisés grâce à des micros ou un mixage de la voix chantée et de la musique en studio sur table de mixage. Dans le premier cas, les sons obtenus tendraient à représenter des mélanges convolutifs; Dans le deuxième cas, ils seraient plus près de mélanges linéaires. Et donc, la méthode de séparation doit s'adapter au type de mélange.

Tableau VI. 18 : Critères ISR et DSR pour les deux méthodes FastICA et EMD.

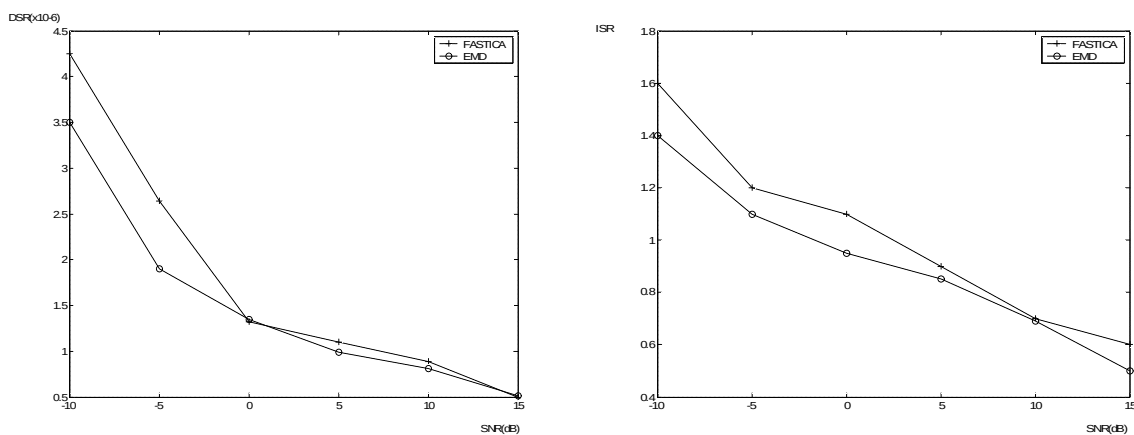
			Mélange 1		Mélange 2		Mélange 3	
			Parole /Parole		Parole /Bruit		Parole /Musique	
			DSR	ISR	DSR	ISR	DSR	ISR
FastICA	Matrice 1	Source 1	0.0044	5.9 ^e -4	0.0035	1.1 ^e -4	0.0201	1.9 ^e -4
		Source 2	0.0043	4.4 ^e -4	0.0012	0.207	0.0085	0.96
	Matrice 2	Source 1	0.0044	6 ^e -4	0.0035	1.1481 ^e -4	0.0201	1.9874 ^e -4
		Source 2	0.0043	4.4 ^e -4	0.0012	0.207	0.0085	0.96
	Matrice 3	Source 1	0.0044	5.9 ^e -4	0.0035	1.1 ^e -4	0.0203	2.007 ^e -4
		Source 2	0.0043	4.4 ^e -4	0.0012	0.207	0.0086	0.96
EMD	Matrice 1	Source 1	0.36	0.45	0.74	0.033	0.21	2.3 ^e -4
		Source 2	0.99	0.062	0.11	0.13	0.017	0.96
	Matrice 2	Source 1	0.109	0.0079	0.11	0.0052	0.74	0.013
		Source 2	0.67	0.038	0.11	0.15	0.037	0.071
	Matrice 3	Source 1	0.065	0.0040	0.107	0.0069	0.34	0.0057
		Source 2	0.026	2.08 ^e -4	0.11	0.16	0.0302	0.13

- **Evaluation en présence de bruit**

Les Figures VI.4.a et VI.4.b représentent l'évolution des critères DSR et ISR pour différentes valeurs du Rapport Signal à Bruit (SNR) pour les deux méthodes FASTICA et EMD. Les mélanges utilisés (parole+musique) sont noyés dans du bruit blanc gaussien. Le bruit est alors considéré comme une source supplémentaire. La matrice de mélange étant aléatoire, les critères utilisés représentent alors des moyennes pour chaque paire de mélanges. Nous observons, que les valeurs de DSR et ISR diminuent lorsque le SNR augmente, ce qui fournit une meilleure séparation des sources sonores. Nous remarquons cependant que les courbes de DSR et ISR associés à la méthode FASTICA sont toujours plus faibles que ceux de l'EMD, notamment pour des signaux fortement bruités (SNR négatif). Ceci nous permet de conclure que la méthode de décomposition par modes empiriques (EMD) fournit des résultats bien plus performants que l'Analyse en Composantes Indépendantes bien que leurs performances semblent équivalentes pour des valeurs élevées du SNR.



Figures VI.4.a et VI.4.b: Evolution des critères DSR et ISR en fonction du RSB (parole+ musique+bruit)



Figures VI.5.a et VI.5.b: Evolution des critères DSR et ISR en fonction du RSB (parole+bruit)

Les Figures VI.5.a et VI.5.b représentent aussi l'évolution du DSR et ISR en fonction du Rapport Signal à Bruit (SNR) pour les deux méthodes comparées (EMD et FASTICA). Les mélanges sont composés de signaux de parole corrompus par du bruit gaussien. Ces résultats confirment les conclusions précédentes, à savoir que l'EMD fournit une meilleure séparation de mélanges en situation de signaux fortement bruités quel que soit le type de mélange.

VI.3 SUR LA CLASSIFICATION EN GENRES MUSICAUX

VI.3.1 Description de la Base de données « Zouheir »

La classification en genres nécessite la mise en œuvre d'une base de données assez riche afin que les jeux d'apprentissage et de test soient raisonnables. Pour l'élaboration de notre base de données concernant la musique algérienne et afin de déterminer ses principaux genres, nous avons procédé à une écoute intensive des différentes chansons diffusées sur la station radiophonique EL BAHDJA et nous avons aussi effectué une enquête auprès d'un panel d'étudiants sur les genres de musique les plus écoutés. La taxonomie retenue est la suivante : Chaabi, Rai, Haouzi, Kabyle, Staifi, Chaoui.

Pour avoir une représentation exhaustive de chaque genre musical, nous avons pris 10 chansons pour chaque style. Afin d'assurer la variété de la qualité des enregistrements, les fichiers musicaux proviennent de différents supports : radio ou compacts disques et ont été enregistrés sous différents formats : .wav ou .MP3. Ainsi, notre base de données est constituée de 60 morceaux musicaux (Tableau VI.19), sans voix chantée, de 3 secondes chacun ; Il a en effet été prouvé que l'écoute de 3 secondes de musique suffisait à un être humain pour reconnaître un genre dans 70 %. Cette durée est généralement admise par nombre de chercheurs. Il y a 10 fichiers par genre musical et 6 genres musicaux. La fréquence d'échantillonnage est fixée à 22050 Hz (fréquence standard pour les signaux musicaux) en mode mono. Chaque fichier contient 66150 échantillons.

Tableau VI.19 : Récapitulatif de la base de données

Chaabi	Kabyle	Chaoui	Haouzi	Staïfi	Raï
c001.wav	c002.wav	c003.wav	c004.wav	C005.wav	c006.wav
c007.wav	c008.wav	c009.wav	c010.wav	C011.wav	c012.wav
c013.wav	c014.wav	c015.wav	c016.wav	C017.wav	c018.wav
c019.wav	c020.wav	c021.wav	c022.wav	C023.wav	c024.wav
c025.wav	c026.wav	c027.wav	c028.wav	C029.wav	c030.wav
c031.wav	c032.wav	c033.wav	c034.wav	C035.wav	c036.wav
c037.wav	c038.wav	c039.wav	c040.wav	C041.wav	c042.wav
c043.wav	c044.wav	c045.wav	c046.wav	C047.wav	c048.wav
c049.wav	c050.wav	c051.wav	c052.wav	C053.wav	c054.wav
c055.wav	c056.wav	c057.wav	c058.wav	C059.wav	c060.wav

Etant donné la taille des fichiers utilisés, il est nécessaire de réduire la masse de données. Nous associerons à chaque extrait musical, choisi par segmentation manuelle, un vecteur acoustique. Ce vecteur acoustique est une représentation compacte et réduite du signal musical auquel il est associé. Différents vecteurs acoustiques seront ainsi mis en œuvre et leur pertinence sera évaluée lors de l'étape de classification. Les différents vecteurs acoustiques se doivent de traduire au moins une partie des caractéristiques perceptives de la musique (timbre, rythme, contenu harmonique). Ils sont constitués par concaténation de diverses combinaisons de métriques « signal » : coefficients LPC, coefficients MFCC, paramètres spectraux, paramètres temporels, perceptuels et coefficients d'ondelettes. Une nouvelle base de données, constituée des vecteurs acoustiques décrits précédemment, de taille évidemment réduite, constituera le cœur de notre système de classification en genres. Les métriques signal ou descripteurs utilisés sont décrites dans ce qui suit.

VI.3.2 Extraction des descripteurs

Les extraits correspondants aux fichiers cités ci-dessus ont une durée de 3s : c'est la fenêtre *texturale*. Or certains descripteurs nécessitent d'être calculés sur des durées pour lesquelles le signal puisse être considéré comme stationnaire ; C'est pourquoi il est parfois nécessaire de subdiviser la fenêtre texturale en fenêtres *d'analyse* dont la durée varie entre 20 et 30ms. Les fenêtres d'analyse se recouvrent de 50%.

Trois jeux de descripteurs ont été établis. Le premier est composé des 14 descripteurs suivants (jeu **Temporel-Spectral-Perceptuel- MFCC** soit **TSPMFCC**):

- ❖ Taux de passage par zéro, taux de trames de basse énergie, centre de gravité spectral, étalement spectral, spectral skewness, spectral kurtosis, décroissance spectrale, roulement spectral, flux spectral, MFCC, delta-MFCC, delta-delta-MFCC, sonie, acuité.

Deux stratégies ont été envisagées pour l'extraction des vecteurs acoustiques : soit garder l'ensemble des descripteurs, calculés sur chaque fenêtre d'analyse et qui seront organisés sous forme de matrice (approche instantanée), ou alors ne garder que les moyennes et variances de chaque descripteur sur l'ensemble des fenêtres (approche globale).

Le deuxième est composé des descripteurs ondelettes suivants (jeu **OND**):

- ❖ Moyenne et variance par échelle en valeurs absolues, Rapport des moyennes des sous bande adjacentes par échelle en valeurs absolues et Ecart type par échelle en valeurs absolues ;

Ce deuxième jeu de descripteurs est directement calculé sur la fenêtre texturale (toute la durée de l'échantillon musical) : ceci tient au fait que la transformée en ondelettes échappe aux contraintes de non stationnarité imposées par la transformée de Fourier.

Enfin, le troisième jeu de descripteurs (jeu **Temporel-Spectral-Perceptuel-MFCC Indépendants** soit **TSPMFCC-IND**) dont la construction est intimement liée au jeu **TSPMFCC** ; En effet, les vecteurs acoustiques issus du jeu **TSPMFCC** sont redondants entre eux et portent parfois la même information, nous avons donc proposé de les rendre indépendants (et donc d'éliminer toute redondance d'information) en utilisant l'Analyse en Composantes Indépendantes. Nous avons utilisé les descripteurs TSPMFCC issus de l'approche globale (moyenne et variance de chaque descripteur) ; Nous constituons alors des vecteurs acoustiques formés des moyennes (ou des variances) de chaque descripteur pour l'ensemble de la base de données. Chaque vecteur est alors considéré non plus comme un signal mais plus encore comme un mélange de sources (composantes) indépendantes; Il suffit alors de séparer les sources pour retrouver les sources indépendantes en utilisant l'ACI ; Ces sources indépendantes ou signaux sont également des vecteurs descripteurs indépendants : les descripteurs **TSPMFCC-IND**.

Comment évaluer les jeux de descripteurs précédents ? Tout d'abord, il faut se placer dans le cadre d'une application précise : nous avons choisi la classification en genres musicaux. Il apparaît alors que l'outil de comparaison ne saurait être que le taux de classification. Il faut donc passer par une étape de classification. Nous avons travaillé avec des machines d'apprentissage supervisé (réseaux de neurones et machines à vecteurs support) qui nécessitent au préalable un apprentissage que nous détaillons dans les sections suivantes. *Nous tenons à préciser que nous ne cherchons pas à comparer les machines d'apprentissage mais les jeux de descripteurs. La machine d'apprentissage ne représentera donc qu'un outil dont il faudra bien évidemment optimiser l'utilisation.*

VI.3.3 Réseaux de neurones, Approche globale et descripteurs TSPMFCC

Nous avons commencé par implémenter un classificateur neuronal du type perceptron multicouche (**PMC**). Comme toutes les méthodes de classification supervisée, la classification par réseaux de neurones nécessite une étape d'apprentissage, qui sera suivie d'une étape de validation et enfin de la phase test.

Nous avons scindé notre base de données en deux : 80% de la base de données sera utilisé pour l'apprentissage et les 20% restant serviront à tester le réseau. Nous avons utilisé pour entraîner le PMC l'algorithme de rétro-propagation du gradient (RPG). L'apprentissage consistera à présenter à l'entrée du réseau un exemple de la base de données (sous forme de vecteur acoustique), et en sortie le label (qui est un vecteur binaire de taille fixe égale à 6), correspondant au genre de l'exemple. Le réseau apprend l'exemple (il converge) et nous refaisons cela pour tous les exemples. Lorsque tous les exemples ont été appris, la matrice des poids obtenue est sauvegardée.

Nous rechercherons l'architecture optimale, nous offrant les meilleurs taux de classification en genres (nombres de neurones de la couche cachée, pas d'apprentissage et nombre d'itérations). Nous explorerons également les différentes structures du vecteur acoustique pour la classification et comparerons leurs performances vis à vis de la classification en genres de signaux musicaux.

A- Vecteurs acoustiques

Comme nous avons choisi l'approche globale, les segments de 3s (fenêtre texturale) sont subdivisés en plusieurs petites fenêtres d'une durée de 20ms à 30ms (fenêtre d'analyse) se recouvrant de 50%. Le vecteur acoustique global est constitué de la moyenne et la variance des paramètres calculés, sur chaque fenêtre d'analyse, sur la durée de la fenêtre texturale. Nous avons calculé sur chaque fenêtre d'analyse les descripteurs suivants :

- Descripteurs spectraux: Centre de gravité Spectral, Flux Spectral, Roulement Spectral.
- Descripteurs temporels : Taux de Passage par Zéro (TPZ), Energie .
- Coefficients MFCC : nous avons pris successivement 5, 9 et 13 coefficients.
- Coefficients LPC : l'ordre de prédiction a été fixé à 5, 9 puis 13.

Le vecteur acoustique final est donc constitué comme suit:

- ✓ Moyenne et Variance du : Centre de gravité Spectral (CS), Flux spectral (FS), Roulement Spectral (SR), Taux de Passage par Zéro (TPZ) : 8 coefficients
- ✓ Energie (LE) : 1 coefficient.
- ✓ Moyenne et Variance des coefficients MFCC (10, 18, 26).
- ✓ Moyenne et Variance des coefficients LPC (10, 18, 26).

La dimension du vecteur de timbre dépendra du choix de nombre de descripteurs utilisés. Les descripteurs sont rangés dans une matrice dont les dimensions sont définies par le nombre d'extraits (60) et la taille des vecteurs acoustiques. A titre d'illustration, les tableaux VI.20, VI.21, VI.22, VI.21 et VI.22 représentent les moyennes ($M_{_}$) et variances ($V_{_}$) des différents descripteurs cités précédemment, pour une chanson de chaque genre musical. Chaque tableau est suivi d'un commentaire.

Tableau VI.20: Moyenne et Variance des paramètres temporels et spectraux

	M_{CS}	V_{CS}	M_{FS}	V_{FS}	M_{SR}	V_{SR}	M_{TPZ}	V_{TPZ}	LE
Chaabi	331.2267	0.0448	0.4	0.809	15.849	3.1613	1.0212	0.0621	0.54
Kabyle	330.2524	1.019	0.5	2.3247	24.557	1.0629	0.8976	0.0495	0.58
Chaoui	330.2834	3.2786	0.4964	19.8447	25.793	0.577	1.1283	0.2187	0.81
Haouzi	329.6986	2.8511	0.4418	8.0959	23.504	4.4361	0.5907	0.0513	0.77
Staifi	330.7397	0.5618	0.6357	1.7972	25.612	2.2399	2.16	0.173	0.58
Rai	330.4432	1.6195	0.4187	1.7609	25.75	0.2045	0.8716	0.1361	0.55

Dans le tableau précédent, nous remarquons que les moyennes du Centre de gravité spectral, du Flux Spectral et du Roulement Spectral sont du même ordre de grandeur, quel que soit le genre musical considéré (elles sont même pratiquement identiques pour le CS). La moyenne du TPZ est elle différente suivant le genre. Le paramètre LE est plus élevé pour les genres Chaoui et Haouzi que pour les autres genres. Par contre, les variances de chacun des descripteurs sont nettement différentes pour chaque genre.

Tableau VI.21 : Moyenne des 9 premiers coefficients MFCC

	M_{mfcc1}	M_{mfcc2}	M_{mfcc3}	M_{mfcc4}	M_{mfcc5}	M_{mfcc6}	M_{mfcc7}	M_{mfcc8}	M_{mfcc9}
Chaabi	-11.97	-0.1943	-0.8261	0.1252	-0.3805	-0.0918	-0.3013	0.045	0.0255
kabyle	-12.97	-0.7441	0.0902	0.03	-0.0036	0.0541	0.1197	0.0068	-0.3547
Chaoui	-14.49	-2.129	-0.7264	0.1683	0.0669	0.4002	0.5075	0.1501	0.3128
Haouzi	-13.097	0.5146	-0.3017	0.3288	-0.3328	-0.1867	-0.6244	0.0701	-0.0523
Staifi	-14.64	-3.386	0.0503	-0.1815	-0.428	0.5501	0.2269	-0.3297	0.2754
Rai	-14.06	-0.4704	0.4851	0.4327	-0.4267	-0.3047	-0.4784	-0.4607	0.1465

Tableau VI.22 : Variance des 9 premiers coefficients MFCC

	<i>V_mfcc1</i>	<i>V_mfcc2</i>	<i>V_mfcc3</i>	<i>V_mfcc4</i>	<i>V_mfcc5</i>	<i>V_mfcc6</i>	<i>V_mfcc7</i>	<i>V_mfcc8</i>	<i>V_mfcc9</i>
Chaabi	0.5223	0.3011	0.2111	0.1665	0.0932	0.1266	0.0949	0.1328	0.1293
Kabyle	0.9223	0.1013	0.1839	0.1091	0.1013	0.102	0.0967	0.1445	0.1153
Chaoui	1.866	0.3134	0.5588	0.1292	0.1165	0.0947	0.0695	0.1181	0.0918
Haouzi	1.135	0.3713	0.1161	0.1036	0.0804	0.0899	0.0829	0.0699	0.0837
Staifi	3.4401	0.8047	0.2561	0.1349	0.1319	0.1546	0.1458	0.106	0.118
Rai	1.438	0.4138	0.1906	0.1058	0.1612	0.1275	0.0895	0.132	0.1201

Les tableaux VI.21 et VI.22 montrent que les moyennes et variances des 9 premiers coefficients MFCC sont différentes pour chaque genre. Ils peuvent donc être utilisés dans la construction du vecteur acoustique global. Mais seule une Analyse de la Variance garantirait statistiquement que la différence entre ces groupes (ici genres) est significative.

Tableau VI.23: Moyenne des 9 premiers coefficients LPC

	<i>M_lpc1</i>	<i>M_lpc2</i>	<i>M_lpc3</i>	<i>M_lpc4</i>	<i>M_lpc5</i>	<i>M_lpc6</i>	<i>M_lpc7</i>	<i>M_lpc8</i>	<i>M_lpc9</i>
Chaabi	-0.7745	0.1898	0.5779	-1.139	1.142	-0.535	-0.3648	1.041	-1.112
Kabyle	-0.0151	0.0128	-0.0008	0.0006	-0.0001	0	0	0	0
Chaoui	-0.0813	0.0884	-0.1006	-0.0526	-0.039	-0.061	0.0486	0.0095	0.1027
Haouzi	0.3968	0.9709	0.6503	0.2242	-0.8865	0.7776	0.0427	-0.9477	1.239
Staifi	-0.816	0.0328	-0.0065	0.004	-0.0012	0.0008	-0.0003	0.0002	-0.0001

Tableau VI.24 : Variance des 9 premiers coefficients LPC

	<i>V_lpc1</i>	<i>V_lpc2</i>	<i>V_lpc3</i>	<i>V_lpc4</i>	<i>V_lpc5</i>	<i>V_lpc6</i>	<i>V_lpc7</i>	<i>V_lpc8</i>	<i>V_lpc9</i>
Chaabi	14.099	27.066	17.0029	12.13	19.82	15.79	8.265	15.64	19.32
Kabyle	0.0127	0	0	0	0	0	0	0	0
Chaoui	0.4148	1.057	1.810	2.243	2.215	1.938	1.759	1.865	2.127
Haouzi	115.9	183.4	41.26	37.40	178.8	118.4	5.452	117.2	178.8
Staifi	0.0264	0.003	0.0008	0.0002	0	0	0	0	0
Rai	96.62	234.4	199.9	47.09	13.57	151.3	246.09	150.3	13.04

Les mêmes conclusions que pour les 2 tableaux précédents, à savoir que les moyennes et variances des 9 premiers coefficients LPC sont différentes suivant les genres. Cependant, nous pouvons remarquer que les variances des coefficients LPC sont plus grandes que celles obtenues pour les autres coefficients : il y aurait donc une plus grande disparité de la valeur de ces coefficients autour de leur moyenne.

Au vu des résultats précédents, il semble qu'un vecteur constitué des différents coefficients spectraux et temporels, MFCC et LPC puisse être caractéristique de chaque genre. Cependant une étude statistique serait nécessaire pour corroborer ce résultat (Analyse de la Variance par exemple). Nous allons donc combiner les différents descripteurs décrits ci-dessus pour construire des vecteurs dont nous testons la pertinence lors de la phase classification ; Ils seront notés comme suit :

- **VTS5MFCC** : vecteur comportant les descripteurs temporels, spectraux et 5MFCC (19 coefficients).
- **VTS5LPC** : vecteur comportant les paramètres temporels, spectraux et 5LPC (19 coefficients).
- **V9MFCC** : vecteur comportant les 9 premiers coefficients MFCC : moyenne et variance (18 coefficients).
- **V9LPC** : vecteur comportant les 9 premiers coefficients LPC : moyenne et variance (18 coefficients).
- **VTS** : vecteur comportant les paramètres temporels et spectraux (9 coefficients).
- **V5MFCC** : vecteur comportant les 5 premiers coefficients MFCC (10).
- **V5LPC** : vecteur comportant les 5 premiers coefficients LPC (10).
- **V13MFCC** : vecteur comportant les 13 premiers coefficients MFCC (26).
- **V13LPC** : vecteur comportant les 13 premiers coefficients LPC (26).

B- Architecture du réseau

Nous avons choisi pour notre classificateur un perceptron multicouche PMC à une seule couche cachée. Le nombre de neurones de la couche d'entrée est conditionné par la taille du vecteur de paramètres acoustique. Il sera donc compris entre 9 et 19 (voir section précédente). Quant au nombre de neurones de la couche de sortie il a été défini par la taille du vecteur LABEL de sortie. Dans notre cas il est de 6 puisque nous disposons de 6 genres seulement. Le nombre de neurones de la couche cachée sera optimisé de façon à obtenir la meilleure convergence et le meilleur taux de classification. Pour assurer la convergence, les vecteurs acoustiques seront normalisés entre -1 et 1.

C- Evaluation de nombre de neurones de la couche cachée (NHID)

Nous fixons le pas d'apprentissage ALPHA à 0.1 et le nombre d'itérations NIT à 100, nous faisons varier le nombre de neurones de la couche cachée NHID et le type de vecteurs acoustiques puis observons l'évolution du taux moyen de reconnaissance. Nous présentons à chaque fois deux taux de classification : l'un pour les exemples d'apprentissage et l'autre pour les exemples de test.

Tableau VI.25 : Taux de classification en fonction du nombre de neurones de la couche cachée-VTS5MFCC

NHID	1	2	4	5	6	7	8	9
Taux de classification (apprentissage)	64.58	89.58	100	100	100	100	100	100
Taux de classification (test)	33.33	66.66	75	58.33	83.33	75	83.33	83.33

Tableau VI.26 : Taux de classification en fonction du nombre de neurones de la couche cachée -VTS5LPC

NHID	1	2	4	5	6	7	8	9
Taux de classification (apprentissage)	45.83	85.41	93.75	93.75	95.83	95.83	95.83	97.91
Taux de classification (test)	41.66	50	50	50	58.33	50	58.33	53.33

Nous remarquons que pour les exemples d'apprentissage, le taux moyen de classification augmente en fonction du nombre de neurones pour la couche cachée, jusqu'à atteindre 100%, et ce pour les deux tableaux : ceci traduit la faculté de mémorisation du réseau. Pour les exemples de test, le taux moyen de classification augmente continuellement pour les deux tableaux jusqu'à 6 neurones de la couche cachée. Au-delà, ce taux fluctue mais ne dépasse jamais celui obtenu pour NHID=6. Nous adopterons donc 6 neurones pour la couche cachée.

D- Evaluation de nombre d'itérations

Nous fixerons NHID à 6 et le pas d'apprentissage ALPHA à 0.1 ; Les tableaux (VI.27) et (VI.28) retracent l'évolution des taux moyens de classification en fonction du nombre d'itérations NIT, pour les exemples d'apprentissage et de test.

Tableau VI.27: Taux de classification en fonction du nombre d'itérations- VTS5MFCC.

NIT	10	20	30	40	50	60	70	80	90
Taux de classification (apprentissage)	31.25	89.58	95.83	97.91	97.91	100	100	100	100
Taux de classification (test)	33.33	58.33	91.66	91.66	83.33	75	83.33	83.33	83.33

Tableau VI.28 : Taux moyens de classification en fonction du nombre d'itérations- VTS5LPC.

<i>NIT</i>	<i>10</i>	<i>20</i>	<i>30</i>	<i>40</i>	<i>50</i>	<i>60</i>	<i>70</i>	<i>80</i>	<i>90</i>
Taux de classification (apprentissage)	45.8	77	79.1	87.5	95.8	95.8	95.8	95.8	95.8
Taux de classification (test)	41.6	33.3	58.3	58.3	58.3	58.3	58.3	58.3	58.3

Première constatation : le taux de classification augmente en fonction du nombre d'itérations pour les deux tableaux. Deuxième constatation : au-delà de 70 itérations, il est pratiquement de 100% pour la base d'apprentissage, et il reste constant et maximal (respectivement 58.33 et 83.33%) pour la base test. Nous fixons donc NIT à 70 itérations pour la suite de nos expériences.

D- Evaluation du pas d'apprentissage

Pour NHID=6 et NIT=70, nous faisons varier le pas d'apprentissage. Nous avons associé aux Tableaux VI.29 et VI.30, les histogrammes des Figure VI.6 et VI.7 respectivement.

Tableau VI.29: Taux de classification en fonction du pas d'apprentissage- VTS5MFCC

<i>ALPHA</i>	<i>1</i>	<i>0.1</i>	<i>0.01</i>	<i>0.001</i>	<i>0.0001</i>
Taux de classification (apprentissage)	93.8	100	85.41	68.75	58.33
Taux de classification (test)	58.3	83.33	50	50	41.66

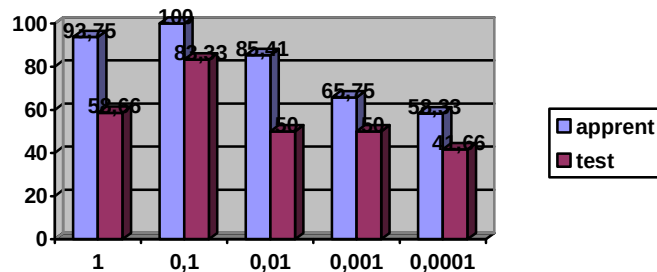


Figure VI.6: Taux de classification en fonction du pas d'apprentissage-VTS5MFCC.

Tableau VI.30 : Taux moyens de classification en fonction du pas d'apprentissage-VTS5LPC.

ALPHA	1	0.1	0.01	0.001	0.0001
Taux de classification (apprentissage)	75	95.83	100	95.83	95.83
Taux de classification (test)	41.66	58.33	41.66	33.33	33.33

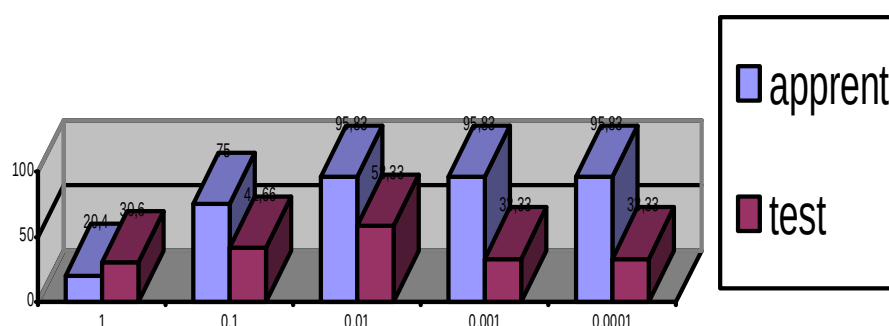


Figure VI.7: Taux de classification en fonction du pas d'apprentissage - VTS5LPC

Les 2 tableaux montrent clairement que les meilleurs taux de reconnaissance sont obtenus dans les deux cas pour un pas d'apprentissage de 0.1. C'est cette valeur qui sera adoptée pour l'apprentissage de notre réseau ALPHA=0.1 .

E- Effet du partitionnement de la base de données sur le taux de classification

Nous avons voulu voir si un partitionnement différent de la base de données en base d'apprentissage et base de test pouvait avoir un effet significatif sur les taux de reconnaissance. Pour ce faire, nous avons réalisé diverses partitions de cette base : sur les 60 extraits, 48 seront utilisés pour l'apprentissage et 12 pour le test. Différents choix des 12 extraits seront réalisés. Les résultats obtenus sont résumés par le tableau VI.31 et la figure VI.8. Nous avons fixés : NHID =6, NIT =60 et ALPHA = 0.1 ;

Tableau VI.31: Effet du partitionnement sur le taux de classification - VTS5MFCC et VTS5LPC.

les partitions de test	[1:12]	[13:24]	[37:48]	[49:60]	[41:60]	[1:15]
%VTS5MFCC	83.33	83.33	91.66	83.33	85	66.66
%VTS5LPC	66.66	100	75	58.33	70	73.33

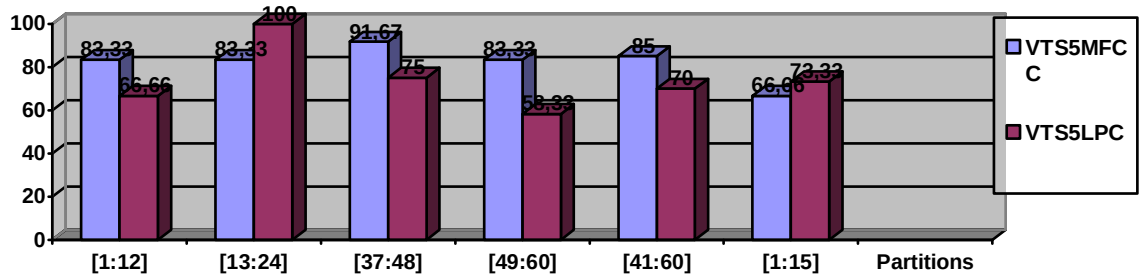


Figure VI.8: Taux de classification pour différentes partitions de la base de données

Comme pour tous les tableaux précédents, le taux de reconnaissance obtenus pour le vecteur VTS5MFCC sont toujours supérieurs à ceux obtenus pour le vecteur VTS5LPC. Nous retrouvons d'ailleurs pour la partition [49:60] les résultats obtenus précédemment. Les conclusions sont différentes suivant le vecteur acoustique considéré. Pour le vecteur VTS5MFCC, les taux de classification demeurent élevés quelle que soit la partition sauf pour la partition [1:15] ; Quant au vecteur VTS5LPC, les taux de classification sont plus élevés que ceux de la partition [49:60]. Nous décidons donc d'utiliser un taux de classification moyen pour contrebalancer l'influence de la partition.

F- Classification en genres et Evaluations

L'architecture de notre réseau est à présent définie et l'apprentissage terminé. Il reste à tester les capacités de généralisation de ce réseau : permet-il la classification en genres d'extraits tests inconnus ? Nous disposons pour cette partie de 12 extraits test pour chaque genre (soit en tout 72 extraits).

- **Taux de classification**

La Figure VI.9 montre les meilleurs taux moyens de classification obtenus pour chacun des six genres proposés. Ces taux sont obtenus pour le vecteur acoustique constitué des descripteurs temporels, spectraux et de 5 coefficients MFCC. Le score global pour l'ensemble des genres est de 86.11%. Nous avons également constaté que les genres: Haouzi, Chaoui et Chaabi obtiennent les meilleurs taux (100%) alors que le genre Rai obtient le plus mauvais score (58.33%).

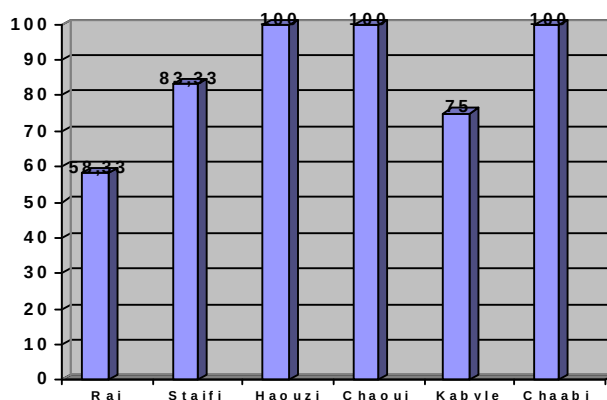


Figure VI.9: Taux de Classification

- **Matrice de confusion**

Le tableau VI.32.a détaille les taux de classification obtenus sous forme de matrice de confusion. Les colonnes de la matrice correspondent au vrai genre et les lignes au genre prédit. Les pourcentages de bonne classification sont sur la diagonale. Les meilleurs scores sont de 87.5% pour le Haouzi et le Chaoui. Les plus mauvais scores sont de 50% pour le Staifi, le Rai et l'Andalou. La confusion entre les genres est essentiellement due aux choix des extraits musicaux (étape de segmentation manuelle et choix des titres) ainsi qu'aux limites confuses entre genres (malouf et andalou ou Staifi et Rai par exemple). Nous suggérerions de considérer ces paires de genres comme sous-classe d'un genre plus général : le genre moderne et le genre classique.

- **Importance de chaque descripteur au sein du vecteur acoustique**

La Figure VI.10 et le tableau VI.32.b représentent les taux de classification obtenus pour différentes combinaisons de descripteurs et pour chaque genre musical. Le meilleur taux (86.11%) est obtenu pour la combinaison descripteurs temporels, descripteurs spectraux et 5 MFCC.

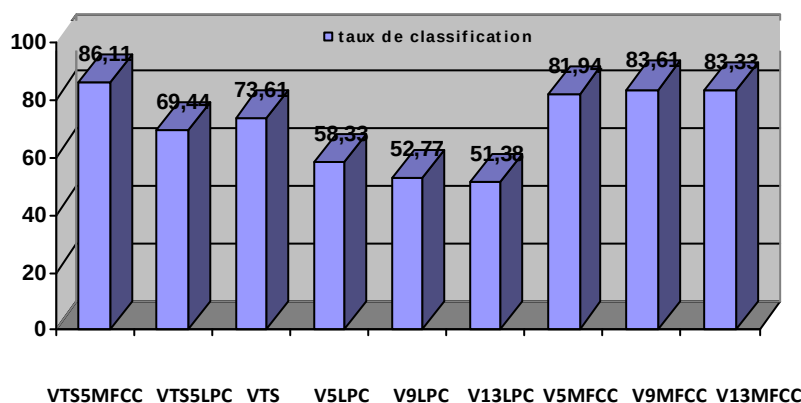


Figure VI.10 : Taux moyens à Différents vecteurs acoustiques

Tableau VI.32.a: Matrice de confusion

	Chaabi	Haouzi	Kabyle	Malouf	Chaoui	Staifi	Rai	andalou
Chaabi	62.5%	12.5%	0%	0%	0%	0%	0%	25%
Haouzi	0%	87.5%	0%	0%	12.5%	0%	0%	0%
Kabyle	0%	12.5%	75%	0%	0%	12.5%	0%	0%
Malouf	0%	12.5%	0%	62.5%	25%	0%	0%	0%
Chaoui	0%	0%	0%	12.5%	87.5%	0%	0%	0%
Staifi	0%	0%	0%	0%	25%	50%	0%	25%
Rai	12.5%	25%	0%	0%	0%	0%	50%	12.5%
Andalou	0%	12.5%	12.5%	25%	0%	0%	0%	50%

Tableau VI.32.b : Taux moyens - Différents vecteurs acoustiques

Vecteurs acoustiques	VTS5MFC C	VTS5LP C	VTS	V5LP C	V9LP C	V13LP C	V5MFC C	V9MFC C	V13MFC C
taux de classification %	86.11	69.44	73.6 1	58.33	52.77	51.38	81.94	83.61	83.33

VI.3.3 Machines à vecteurs support, Approche instantanée et descripteurs TSPMFCC

L'approche instantanée contrairement à l'approche globale utilise tous les descripteurs calculés (pas de moyenne ou de variance). Les descripteurs audio cités ci-dessus sont regroupés ligne par ligne dans une matrice spécifique à chaque fichier audio, de 47 lignes (le nombre des descripteurs) et 256 colonnes (le nombre de fenêtres).

Nous avons mis en place une machine d'apprentissage à vecteurs support (SVM) multiclasse. L'algorithme du SMO (Sequential Minimal Optimization) est utilisé pour résoudre le problème d'optimisation. La stratégie « un contre un » est adoptée pour le cas multiclasse. Deux types de noyaux ont été testés : RBF et Polynomiaux. La phase d'apprentissage dans le cas des SVM nous a permis d'ajuster d'une part les paramètres du noyau (γ pour le noyau RBF et D pour le noyau polynomial), et d'autre part le paramètre de pénalité C du SVM. L'ajustement des paramètres est réalisé grâce à une grille de recherche ; c'est une grille à deux entrées (C et γ) pour le noyau RBF et à une entrée d pour le noyau polynomial. Nous avons également testé la normalisation ou non des données.

A- SVM et Noyau RBF

Nous avons testé le noyau RBF et l'effet de la normalisation des données ; En premier une grille de recherche nous permet de fixer les paramètres C et σ lors de la phase d'apprentissage (sur les exemples soit 80% de la base de données). Ensuite, nous passons à la phase test (sur les 20% restants). Les tableaux VI.33 et VI.34 représentent respectivement la grille de recherche lors de la phase d'apprentissage et la matrice de confusion pour la phase de test pour des données non-normalisées. Les tableaux VI.35 et VI.36, quant à eux, représentent respectivement la grille de recherche lors de la phase d'apprentissage et la matrice de confusion pour la phase de test pour des données normalisées.

Le tableau VI.33 nous permet de constater que le meilleur taux de classification en phase d'apprentissage atteint 70% pour $C = 10^6$ et $\sigma = 10^{-9}$; Ces deux valeurs ont donc été adoptées pour la phase de test. Le tableau VI.34 montre les taux de classification obtenus pour chaque genre. Nous observons que le meilleur score obtenu est de 85% pour le genre Chaabi suivi du genre Chaoui avec 82%. Les genres Kabyle et Staifi obtiennent des scores moyens avec respectivement 68% et 55%. Le plus mauvais score est réalisé par le genre Rai avec 30% : Le Taux moyen de classification pour tous les genres est de 66,66%.

Tableau VI.33 : Evolution du taux de classification en fonction de (C, σ) , sans normalisation.

$\sigma \backslash C$	10^0	10^1	10^2	10^3	10^4	10^5	10^6	10^7	10^8	10^9
10^0	0,166	0,166	0,166	0,166	0,166	0,166	0,166	0,166	0,166	0,166
10^{-1}	0,166	0,166	0,166	0,166	0,166	0,166	0,166	0,166	0,166	0,166
10^{-2}	0,176	0,176	0,176	0,176	0,176	0,176	0,176	0,176	0,176	0,176
10^{-3}	0,166	0,166	0,166	0,166	0,166	0,166	0,166	0,166	0,166	0,166
10^{-4}	0,176	0,178	0,178	0,178	0,178	0,178	0,178	0,178	0,178	0,178
10^{-5}	0,376	0,378	0,378	0,378	0,378	0,378	0,378	0,378	0,378	0,378
10^{-6}	0,508	0,508	0,464	0,462	0,462	0,462	0,462	0,462	0,462	0,462
10^{-7}	0,438	0,500	0,544	0,549	0,546	0,536	0,536	0,536	0,536	0,536
10^{-8}	0,382	0,412	0,480	0,570	0,634	0,666	-	-	-	-
10^{-9}	0,353	0,388	0,401	0,447	0,574	0,662	0,700	-	-	-

Tableau VI.34 : Matrice de confusion (phase de test), sans normalisation.

	Chaabi	Kabyle	Chaoui	Haouzi	Staïfi	Raï	
Chaabi	0,854	0,00520	0,0130	0,0559	0,0130	0,0585	1
Kabyle	0,00781	0,677	0,00130	0,0338	0,145	0,134	1
Chaoui	0,0559	0,0338	0,821	0,0195	0,0507	0,0182	1
Haouzi	0,115	0,0208	0,00130	0,791	0,0182	0,0520	1
Staïfi	0,0221	0,221	0	0,0989	0,552	0,105	1
Raï	0,0833	0,0976	0,0572	0,235	0,226	0,299	1

Tableau VI.35: Evolution du taux de classification en fonction de (C, $\hat{u}$), avec normalisation.

Σ C	10^0	10^1	10^2	10^3	10^4	10^5	10^6	10^7	10^8	10^9
10^0	0,195	0,272	0,375	0,477	0,592	0,666	0,719	0,718	0,715	0,717
10^{-1}	0,190	0,202	0,235	0,337	0,445	0,567	0,658	0,679	0,674	0,674
10^{-2}	0,183	0,189	0,202	0,230	0,335	0,394	0,533	0,535	0,557	0,550
10^{-3}	0,182	0,183	0,189	0,202	0,230	0,335	0,405	0,425	0,420	0,424
10^{-4}	0,182	0,182	0,183	0,189	0,203	0,232	0,349	0,355	0,351	0,355
10^{-5}	0,181	0,181	0,181	0,182	0,189	0,205	0,261	0,240	0,240	0,260
10^{-6}	0,181	0,181	0,181	0,181	0,181	0,190	0,211	0,208	0,212	0,213
10^{-7}	0,176	0,176	0,176	0,176	0,176	0,176	0,183	0,145	0,141	0,141
10^{-8}	0,189	0,189	0,189	0,189	0,189	0,189	0,189	0,189	0,189	0,189
10^{-9}	0,189	0,189	0,189	0,189	0,189	0,189	0,189	0,189	0,189	0,189

Tableau VI.36: Matrice de confusion (Phase de test), avec normalisation

	Chaabi	Kabyle	Chaoui	Haouzi	Staïfi	Raï	
Chaabi	0,923	0,00130	0,00911	0,0312	0,0130	0,0221	1
Kabyle	0,00390	0,701	0	0,0143	0,108	0,171	1
Chaoui	0,0286	0,0130	0,886	0,00520	0,0638	0,00260	1
Haouzi	0,0664	0,00260	0,00130	0,861	0,0169	0,0507	1
Staïfi	0,00520	0,148	0,00260	0,0859	0,618	0,139	1
Raï	0,0468	0,139	0,0546	0,261	0,177	0,320	1

Dans le cas de données normalisées, le taux de classification pour la phase d'apprentissage a augmenté : il atteint 72% pour $C = 10^6$ et $\gamma = 1$ (voir Tableau VI.35) ; Ces valeurs sont donc adoptées pour la phase de test. La matrice de confusion du tableau VI.36 montre une nette augmentation des taux de classification pour chaque genre. Le meilleur score est également obtenu par le genre Chaabi avec 92%, suivi du Chaoui avec 87% ; Les genres Kabyle et Staïfi montent à 70% et 62% respectivement ; Enfin le Raï atteint 32%. Le taux moyen de classification pour tous les genres est de 71.87%. ***Nous observons que la normalisation des données apporte une amélioration aux taux moyens de classification pour chaque genre et par conséquent à taux moyen pour tous les genres.***

B- SVM et Noyau Polynomial

Nous avons décidé de tester le noyau polynomial. Les données sont normalisées. Le tableau VI.37 illustre les taux de classification obtenus lors de la phase d'apprentissage pour différentes valeurs du degré d du noyau polynomial. Le tableau VI.38 représente la matrice de confusion de la phase de test.

Tableau VI.37 : Taux de classification en fonction de d

d	1	2	3	4	5
Taux de Classification	0.198	0.311	0.468	0,617	0,641

Tableau VI.38 : Matrice de confusion (Phase de test), Noyau Polynomial

	Chaabi	Kabyle	Chaoui	Haouzi	Staïfi	Raï	
Chaabi	0,937	0	0,0104	0,0312	0,00911	0,0117	1
Kabyle	0	0,825	0	0,0559	0,0546	0,0638	1
Chaoui	0,0182	0,0377	0,850	0,0299	0,0572	0,00651	1
Haouzi	0,131	0,0195	0,00130	0,779	0,0104	0,0572	1
Staïfi	0,00651	0,363	0	0,222	0,205	0,201	1
Raï	0,0807	0,106	0,0859	0,298	0,180	0,247	1

La durée d'exécution des programmes lors de la phase d'apprentissage est très longue, ce qui explique le peu de résultats. Le meilleur taux de classification lors de la phase d'apprentissage est obtenu pour $d=5$: il est faible et vaut 64%. Pour $d=5$, la matrice de confusion du tableau VI.38 montre que le score du genre Chaabi atteint 94%, suivi par celui du Chaoui avec 85% et du Kabyle avec 82%. Le Haouzi obtient un score de 78% ; Les plus mauvais scores sont ceux des genres Staïfi et Raï avec respectivement 20% et 25%. Le taux moyen global de classification est de 66%.

C- SVM, Noyau RBF et Fusion de classes

Ayant remarqué que les deux genres Raï et Staïfi présentaient des scores faibles qui affectaient le score global, et que de plus les deux genres présentaient des similarités instrumentales et rythmiques, nous avons décidé de fusionner ces deux genres en un seul genre, que nous avons labellisé genre moderne. Nous utilisons des données normalisées et un noyau RBF. Les tableaux VI.39 et VI.40 illustrent la grille de recherche lors de l'apprentissage ainsi que la matrice de confusion de la phase test.

Tableau VI.39 : Matrice de confusion (Phase test)

	Chaabi	Kabyle	Chaoui	Haouzi	Staïfi	
Chaabi	0.955	0	0.00520	0.0325	0.00651	1
Kabyle	0.00130	0.834	0	0.0221	0.141	1
Chaoui	0.0260	0.0104	0.902	0.00520	0.0559	1
Haouzi	0.0742	0.0273	0	0.890	0.00781	1
Moderne	0.00781	0.180	0.00130	0.122	0.687	1

Tableau VI.40 : Evolution du taux de classification en fonction de (C, σ).

$\sigma \backslash C$	10^0	10^1	10^2	10^3	10^4	10^5	10^6	10^7	10^8	10^9
10^0	0.225	0.315	0.435	0.556	0.698	0.786	0.854	0.857	0.845	0.849
10^{-1}	0.221	0.234	0.274	0.395	0.521	0.670	0.781	0.797	0.797	0.792
10^{-2}	0.214	0.221	0.234	0.273	0.393	0.463	0.637	0.651	0.668	0.669
10^{-3}	0.214	0.214	0.221	0.233	0.274	0.392	0.479	0.501	0.499	0.504
10^{-4}	0.214	0.214	0.214	0.221	0.234	0.279	0.409	0.419	0.407	0.419
10^{-5}	0.213	0.213	0.213	0.213	0.220	0.237	0.306	0.279	0.296	0.303
10^{-6}	0.212	0.212	0.212	0.212	0.212	0.219	0.229	0.226	0.220	0.222
10^{-7}	0.209	0.209	0.209	0.209	0.209	0.209	0.217	0.209	0.224	0.228
10^{-8}	0.211	0.211	0.211	0.211	0.211	0.211	0.211	0.211	0.211	0.211
10^{-9}	0.211	0.211	0.211	0.211	0.211	0.211	0.211	0.211	0.211	0.211

Le tableau VI.39 montre que le meilleur taux de classification lors de la phase d'apprentissage est de 85.72% pour $C=10^7$ et $\sigma=1$; Ces valeurs ont été alors utilisées lors de la phase test et le tableau VI.40 détaille les scores obtenus pour chaque genre. Le Chaabi obtient le score de 95% , le Chaoui 90%, le Haouzi 89%, le Kabyle 83 et enfin le genre moderne 69%. Le taux moyen global pour l'ensemble des genres est de 85%.

La fusion des deux genres en un seul genre moderne a permis de faire passer le aux moyen global de classification à 85%.

VI.3.4 Machines à vecteurs support, Approche globale et descripteurs TSPMFCC

Le vecteur acoustique utilisé est constitué des statistiques (moyenne et variance) des descripteurs audio cités ci-dessus auquel nous avons rajouté les dérivées secondes des MFCC et la sonie globale (soit 51 descripteurs). Le SVM est le même que celui de la section précédente (multiclasse, noyau RBF). Nous décrivons dans ce qui suit différentes expériences sur l'influence du type et le nombre de descripteurs. Nous décrivons également les résultats obtenus pour les descripteurs indépendants introduits au Chapitre II.

A- SVM, Approche globale et descripteurs TSPMFCC

Le tableau VI.41 représente l'évolution du taux de classification en fonction de C et σ pour un vecteur acoustique de 26 descripteurs (7 descripteurs spectraux, 2 descripteurs temporels,

13MFCC et 4 delta-MFCC). Le choix du nombre et de la combinaison est aléatoire car pratiquement tous les vecteurs plafonnent au même score. Les données sont bien évidemment normalisées. Le tableau VI.42 présente le détail de la classification en genres pour la phase test.

Tableau VI.41 : Evolution du taux de classification en fonction de (C, $\hat{u}$), 26 descripteurs.

$\backslash C$ Sigma	2^0	2^1	2^2	2^3	2^4	2^5	2^6	2^7	2^8	2^9
2^0	0.200	0.200	0.200	0.200	0.200	0.200	0.200	0.200	0.200	0.200
2^{-1}	0.200	0.200	0.200	0.200	0.200	0.200	0.200	0.200	0.200	0.200
2^{-2}	0.200	0.200	0.200	0.200	0.200	0.200	0.200	0.200	0.200	0.200
2^{-3}	0.250	0.250	0.250	0.250	0.250	0.250	0.250	0.250	0.250	0.250
2^{-4}	0.300	0.300	0.300	0.300	0.300	0.300	0.300	0.300	0.300	0.300
2^{-5}	0.350	0.350	0.350	0.350	0.350	0.350	0.350	0.350	0.350	0.350
2^{-6}	0.350	0.350	0.350	0.350	0.350	0.350	0.350	0.350	0.350	0.350
2^{-7}	0.300	0.250	0.250	0.250	0.250	0.250	0.250	0.250	0.250	0.250
2^{-8}	0.300	0.300	0.300	0.300	0.300	0.300	0.300	0.300	0.300	0.300
2^{-9}	0.300	0.350	0.350	0.350	0.350	0.350	0.350	0.350	0.350	0.350

Tableau VI.42: Matrice de confusion (Phase test).

	Chaabi	Kabyle	Chaoui	Haouzi	Staïfi	Raï	
Chaabi	0.333	0	0.333	0.333	0	0	1
Kabyle	0	0.666	0.3333	0	0	0	1
Chaoui	0	0	1.00	0	0	0	1
Haouzi	0	0	0.666	0.333	0	0	1
Staïfi	0.250	0.500	0.250	0	0	0	1
Raï	0.250	0.250	0.500	0	0	0	1

La phase d'apprentissage est très insatisfaisante puisque le meilleur taux obtenu est de 35%. Plusieurs couples (C, $\hat{u}$) permettent d'obtenir ce résultat (voir tableau VI.40). Nous avons retenu le couple : $C=2^0$ et $\hat{u}=5^{-5}$. La matrice de confusion du tableau VI.42 montre un

effondrement total des taux de classification sauf pour le Chaoui qui atteint 100%. Le taux moyen global de classification est de 35%. Ceci était prévisible vu le faible taux atteint durant la phase d'apprentissage.

B- SVM, Approche globale et descripteurs TSPMFCC-IND

Nous avons alors introduit la notion de descripteurs indépendants TSPMFCC; Les vecteurs précédents sont rendus indépendants par le biais d'une Analyse en Composantes Indépendantes grâce à l'algorithme FastICA. Les tableaux VI.43 représente la nouvelle évolution du taux de classification en fonction de (C, σ). La matrice de confusion de la figure VI.44 illustre les taux obtenus genre par genre.

Tableau VI.43 : Evolution du taux de classification en fonction de (C, σ), descripteurs TSPMFCC-IND

$\sigma \backslash C$	2^0	2^1	2^2	2^3	2^4	2^5	2^6	2^7	2^8	2^9
2^0	0.800	0.800	0.800	0.750	0.750	0.750	0.750	0.750	0.750	0.750
2^{-1}	0.800	0.850	0.900	0.900	0.900	0.8500	0.850	0.850	0.850	0.850
2^{-2}	0.750	0.850	0.850	0.850	0.850	0.850	0.850	0.850	0.850	0.850
2^{-3}	0.750	0.850	0.900	0.850	0.850	0.800	0.800	0.800	0.800	0.800
2^{-4}	0.800	0.700	0.850	0.900	0.850	0.850	0.850	0.850	0.800	0.800
2^{-5}	0.700	0.700	0.800	0.900	0.900	0.900	0.850	0.850	0.850	0.850
2^{-6}	0.600	0.700	0.700	0.800	0.850	0.900	0.900	0.900	0.850	0.850
2^{-7}	0.600	0.600	0.700	0.700	0.800	0.900	0.900	0.900	0.900	0.850
2^{-8}	0.600	0.600	0.650	0.700	0.700	0.800	0.900	0.900	0.900	0.850
2^{-9}	0.600	0.600	0.600	0.650	0.700	0.700	0.800	0.900	0.950	0.950

Tableau VI.44: Matrice de confusion (Phase test).

	Chaabi	Kabyle	Chaoui	Haouzi	Staïfi	Raï	
Chaabi	1.00	0	0	0	0	0	1
Kabyle	0	1.00	0	0	0	0	1
Chaoui	0	0	1.00	0	0	0	1
Haouzi	0	0	0	1.00	0	0	1
Staïfi	0	0	0	0	1.00	0	1
Raï	0	0	0	0.2500	0	0.750	1

Le tableau VI.42 fait apparaître un taux de classification de 95% lors de la phase d'apprentissage pour $C=2^8$ et $\eta=10^{-9}$. Avec ces deux valeurs de C et η nous sommes passés à la phase de test. Les résultats obtenus sont détaillés par la matrice de confusion du tableau VI.43. Nous remarquons que tous les genres réalisent des scores de 100% mis à part le genre Rai qui obtient 75%. Le score moyen global de classification est de 95%.

C- Influence du nombre de descripteurs TSPMFCC-IND

Nous avons représenté sur la Figure VI.11, l'évolution du taux de classification en fonction du nombre de descripteurs.

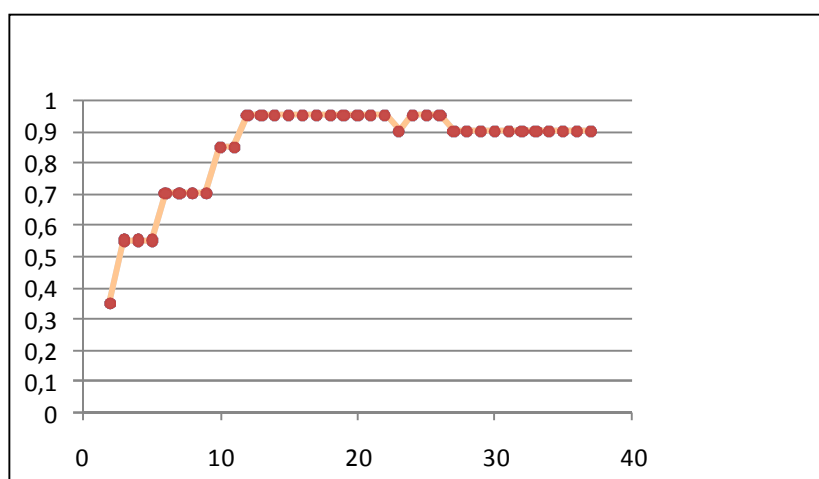


Figure VI.11: Evolution du taux de classification en fonction du nombre de descripteurs.

Le taux de classification augmente au fur et à mesure que nous augmentons le nombre de descripteurs et **passé la barre des 90% dès que nous introduisons les MFCC**.

D- Influence du type de descripteurs TSPMFCC-IND

Pour confirmer le rôle joué par les descripteurs MFCC, nous représentons les taux de classification obtenus dans les deux cas de figure suscités (descripteurs classiques, descripteurs indépendants) et ce pour quatre jeux de 26 descripteurs :

- 7 paramètres spectraux, 2 temporels, 13 delta-MFCC et 4 delta- delta-MFCC.
- 13 MFCC et 13 delta-MFCC.
- 13 MFCC et 13 delta- delta-MFCC.
- 7 descripteurs spectraux, 2 temporels, 13 delta-MFCC, 1 delta-delta-MFCC, la sonie, la sonie spécifique et l'acuité.

Les résultats obtenus sont représentés par la Figure VI .12

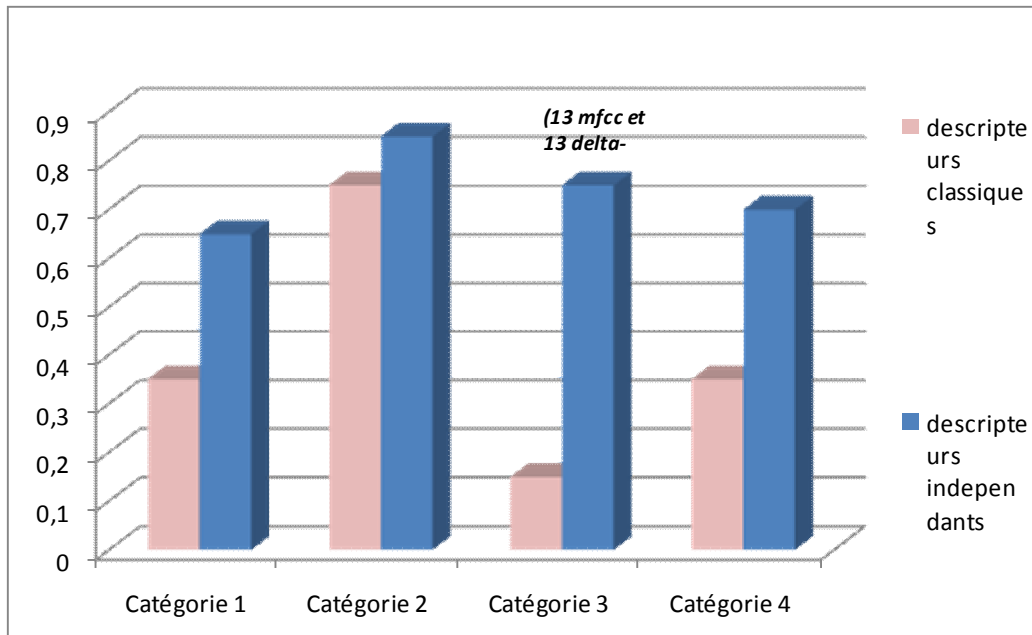


Figure VI.12 : Influence du type de descripteurs

Quelle que soit la catégorie de descripteurs utilisée, les descripteurs indépendants donnent les meilleurs résultats que les descripteurs classiques avec un taux de classification de 95%. De plus, les catégories où sont introduits les coefficients MFCC contribuent aux meilleurs scores.

VI.3.5 Machines à vecteurs support et descripteurs ondelettes (OND)

Les descripteurs ondelettes sont construits à partir des coefficients de la transformée en ondelettes discrète. Plusieurs types de filtres ont été utilisés. Le vecteur acoustique est constitué des statistiques des coefficients à savoir : moyenne et variance de la valeur absolue des coefficients pour chaque échelle, rapport de la valeur absolue des moyennes des échelles adjacentes et écart type des coefficients pour chaque échelle. Nos résultats sont disposés de la manière suivante : pour chaque vecteur acoustique obtenu nous avons varié les filtres d'ondelettes. Les vecteurs acoustiques sont :

- Moyenne et variance par échelle en valeurs absolues : **MVPEA**
- Rapport des moyennes des sous bande adjacentes par échelle en valeurs absolues : **MCPEA**
- Ecart type par échelle avec valeurs absolues : **STDPEA**

Les filtres ondelettes utilisés sont : Daubechies1(db1), Daubechies5 (db2), Meyer (Meyer), Symlet2 (Sym2) et enfin Coiflet1 (Coif1).

A- SVM, Filtre-Ondelette Meyer, Noyau RBF, MCPEA

Le tableau VI.45 illustre un exemple de classification en genres par SVM, noyau RBF et données non normalisées avec l'ondelette de Meyer. Il s'agit de l'étape d'apprentissage : recherche du couple (C, γ) optimal. Le tableau VI.46 représente la matrice de confusion pour la phase test.

Tableau VI.45 : Taux de classification en fonction de (C, γ) , MCPEA, Ondelette Meyer

$\begin{matrix} C \\ \backslash \\ \gamma \end{matrix}$	10^0	10^1	10^2	10^3	10^4	10^5	10^6	10^7	10^8	10^9
10^0	61,1	72,2	72,2	72,2	66,6	66,6	66,6	66,6	66,6	66,6
10^{-1}	55,5	61,1	77,7	77,7	77,7	83,3	83,3	83,3	83,3	83,3
10^{-2}	55,5	55,5	61,1	61,1	77,7	77,7	77,7	77,7	77,7	77,7
10^{-3}	50,0	50,0	50,0	50,0	61,1	77,7	77,7	77,7	77,7	77,7
10^{-4}	50,0	50,0	50,0	50,0	50,0	77,7	77,7	77,7	77,7	77,7
10^{-5}	50,0	50,0	50,0	50,0	50,0	61,1	77,7	77,7	77,7	77,7
10^{-6}	50,0	50,0	50,0	50,0	50,0	50,0	61,1	77,7	83,3	77,7
10^{-7}	55,5	55,5	55,5	55,5	55,5	55,5	55,5	66,6	66,6	66,6

Tableau VI.46 : Matrice de confusion (Phase test)

	Chaabi	Kabyle	Chaoui	Haouzi	Staifi	Rai
Chaabi	100	0	0	0	0	0
Kabyle	0	100	0	0	0	0
Chaoui	0	0	100	0	0	0
Haouzi	0	0	0	100	0	0
Staifi	0	66.67	0	0	33.33	0
Rai	0	0	0	33.33	0	66.67

Le meilleur taux de classification pour la phase d'apprentissage est de 83.33% pour un certain nombre de couples (C, γ) ; Nous avons opté pour les valeurs $C=10^7$ et $\gamma=10^{-2}$ (voir

Tableau VI.45). La matrice de confusion (Tableau VI.46) montre un taux de classification de 100% pour les genres Chaabi, Kabyle, Chaoui et Haouzi ; Le taux de classification du genre Staifi est de 33.33%, celui du genre Rai augmente à 66.67%. Ceci implique un taux global moyen de classification de 83.33%.

B- Influence du type d'ondelette, Données non normalisées et noyau RBF

Pour évaluer l'influence du type d'ondelette utilisée, nous avons représenté sur la Figure VI.13, les taux de classification obtenus lors de la phase test pour les différents vecteurs acoustiques (MVPEA, MCPEA, et STDPEA) et pour 5 ondelettes (db1, db5, Meyer, Sym2 et Coif1).

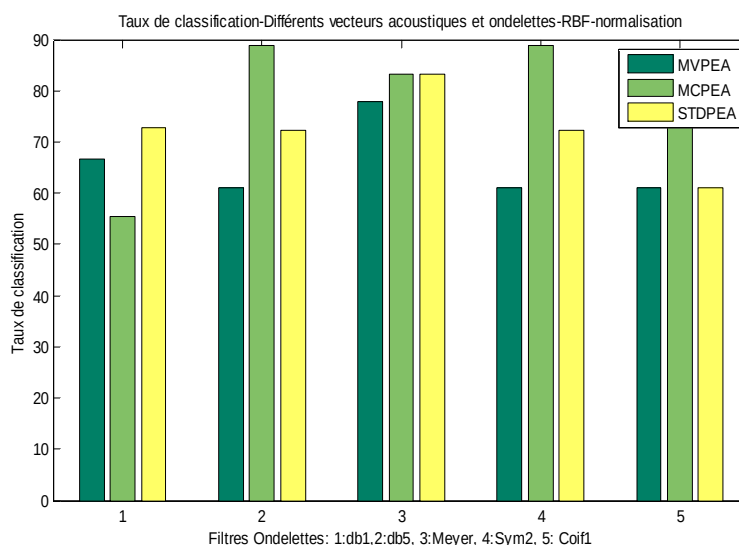


Figure VI.13 : Influence du type d'ondelette

Nous constatons que les filtres db5 et Sym2 obtiennent les meilleurs scores avec un même taux de classification de 88.89%. Ces taux sont obtenus pour le même vecteur acoustique MCPEA.

C- Influence du type de noyau, Données non normalisées

Pour évaluer l'influence du type de noyau (RBF, Polynomial), nous avons représenté sur la Figure VI.14, les taux de classification obtenus lors de la phase test, pour les différents vecteurs acoustiques (MVPEA, MCPEA, et STDPEA) et pour 5 ondelettes (db1, db5, Meyer, Sym2 et Coif1) pour le noyau polynomial.

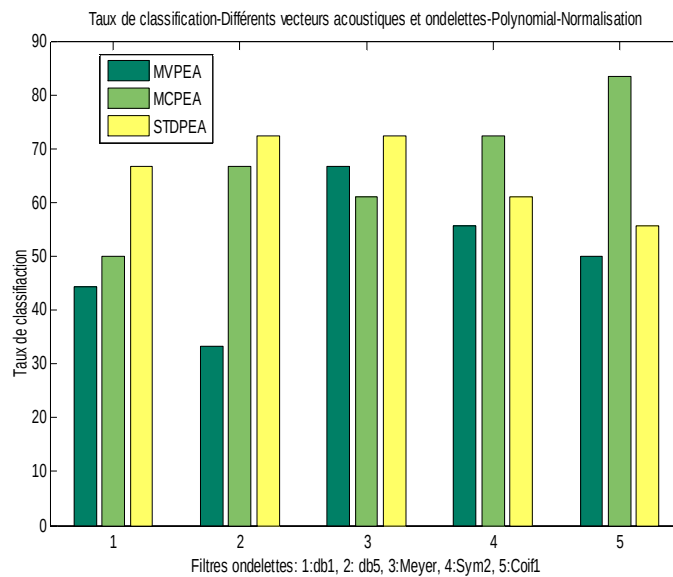


Figure VI.14 : Influence du type de noyau

En comparant les Figures VI.13 (noyau RBF) et VI.14 (noyau Polynomial), nous pouvons constater que les scores obtenus pour le noyau RBF sont supérieurs à ceux obtenus pour le noyau polynomial. Cette conclusion reste valable pratiquement pour toutes les ondelettes et tous les vecteurs acoustiques, à une exception près le vecteur MCPEA et l'ondelette de Coif 1 (il se maintient dans les deux cas à 83.33%).

D- Influence de la normalisation, noyau RBF et vecteur acoustique MCPEA

Les données peuvent être normalisées ou non. La Figure VI.15 montre la comparaison entre les scores obtenus dans chaque cas.

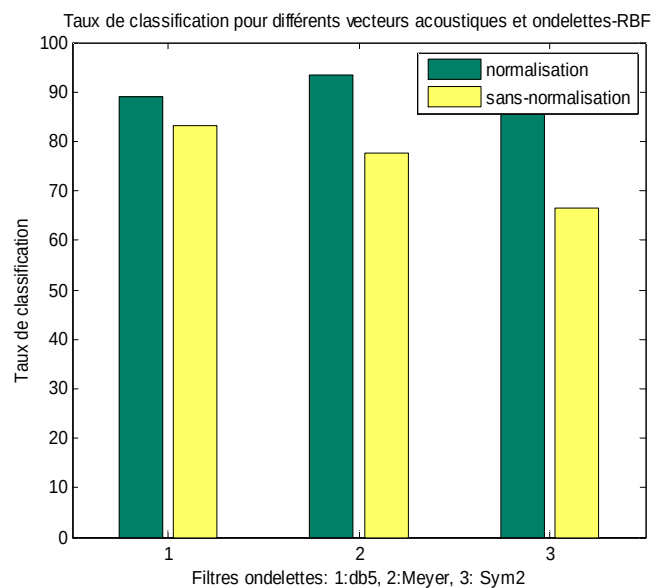


Figure VI.15 : Influence de la normalisation

Nous pouvons constater que la normalisation apporte une nette amélioration des scores pour toutes les ondelettes et tous les vecteurs.

E- Influence du type de vecteur acoustique, Noyau RBF et données non normalisées

Trois vecteurs acoustiques Ondelettes ont été envisagés : MVPEA, MCPEA et STDPEA ; La figure VI.16 montre l'évolution du taux de classification pour les trois vecteurs acoustiques. Différentes ondelettes ont été testées ; Nous avons utilisé pour cette expérience des données normalisées et un noyau RBF.

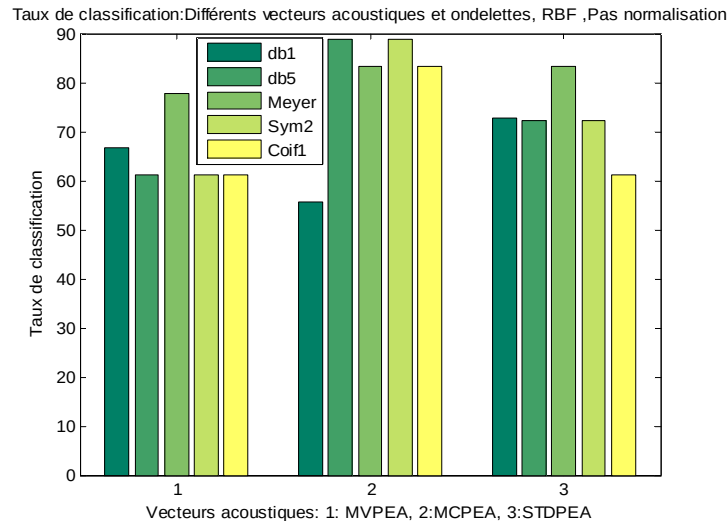


Figure VI.16 : Influence du type de vecteur acoustique

Les différentes expériences menées permettent de conclure que le vecteur MCPEA associé au filtre ondelette db5 ou Sym2 donne le meilleur score soit 88.89%. Les données se doivent d'être normalisées et l'utilisation du noyau RBF est préconisée.

VI.3.6 SVM Multilabels et descripteurs ondelettes (OND)

Les vecteurs acoustiques sont les mêmes que ceux décrits dans la section précédente (descripteurs ondelettes MVPEA, MCPEA, et STDPEA). Le changement se situe au niveau du classifieur. En effet même s'il s'agit toujours d'une machine d'apprentissage à vecteurs support Multiclasse. L'appellation Multilabels provient du fait que certains exemples au niveau du classifieur, et plus spécialement au niveau de l'apprentissage, seront considérés comme appartenant à deux classes : ils auront deux labels et seront appris deux fois (une fois pour chaque label). Par contre au niveau du test, chaque exemple ne pourra appartenir qu'à une seule classe. En nous basant sur les résultats de la section précédente, nous avons appliqué les multilabels à des données normalisées ; Le vecteur acoustique choisi est le vecteur MCPEA ; L'ondelette Sym2 et un noyau RBF ont été adoptés pour cette expérience. Le tableau VI.47 illustre la grille de recherche du couple (C, γ) et le tableau VI.48 la matrice de confusion associée à la phase test.

Tableau VI.47 : Evolution du taux de classification en fonction de (C, $\hat{u}$), cas Multilabels

$\begin{matrix} C \\ \text{Sigma} \end{matrix}$	10^0	10^1	10^2	10^3	10^4	10^5	10^6	10^7	10^8	10^9
10^0	85,33	92,33	92,33	92,33	92,33	92,33	92,33	92,33	92,33	92,33
10^{-1}	85,33	92,33	92,33	92,33	92,33	92,33	92,33	92,33	92,33	92,33
10^{-2}	75,55	75,55	72,22	77,77	77,77	77,77	77,77	77,77	77,77	77,77
10^{-3}	75,55	75,55	75,55	66,66	66,66	66,66	66,66	66,66	66,66	66,66
10^{-4}	75,55	75,55	75,55	75,55	82,22	66,66	66,66	66,66	66,66	66,66
10^{-5}	75,55	75,55	75,55	75,55	82,22	66,66	66,66	66,66	66,66	66,66
10^{-6}	75,55	75,55	75,55	75,55	82,22	82,22	82,22	82,22	82,22	82,22
10^{-7}	75,55	75,55	75,55	75,55	75,55	75,55	75,55	61,11	61,11	61,11

Tableau VI.48 : Matrice de confusion (Phase test)

	Chaabi	Kabyle	Chaoui	Haouzi	Staifi	Rai
Chaabi	100	0	0	0	0	0
Kabyle	0	100	0	0	0	0
Chaoui	0	0	100	0	0	0
Haouzi	0	0	0	100	0	0
Staifi	0	66.67	0	0	77.77	0
Rai	0	0	0	33.33	0	77.77

Le tableau VI.47 montre que le taux de classification durant la phase d'apprentissage atteint 92.33% et ce pour un certain nombre de couples (C, $\hat{u}$). Nous avons choisi le couple $C=10^6$ et $\hat{u}=10^{-1}$. La matrice de confusion du tableau VI.48 montre les scores obtenus pour chaque genre. Ils sont de 100% pour 4 genres (Chaabi, Kabyle, Chaoui et Haouzi) et de 77.77% pour 2 genres (Staifi et Rai). Le taux de classification moyen global est de 92.59%.

VI.3.7 Comparaison entre les différents jeux de descripteurs et Conclusions

La Figure VI.17 résume les taux de classification obtenus pour les différentes combinaisons de descripteurs, les différentes approches (instantanée ou globale) ainsi que les différentes machines d'apprentissage (SVM multiclasse, SVM multilabels, PMC).

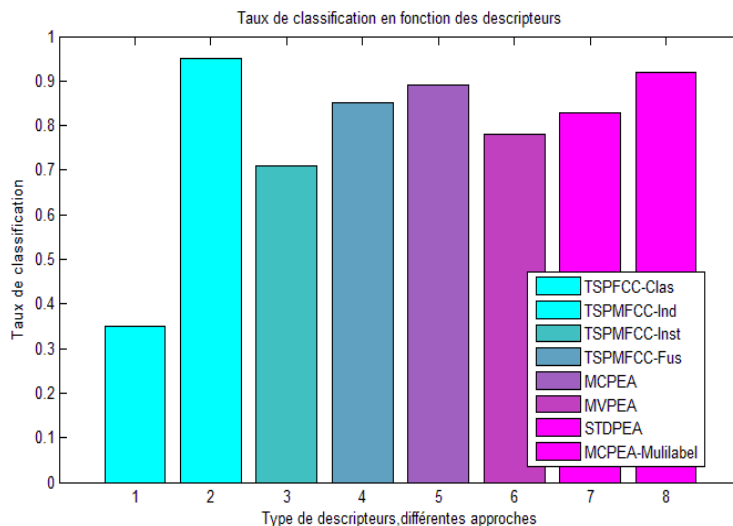


Figure VI.17 : Comparaison entre les différents jeux de descripteurs

Nous pouvons constater que le meilleur taux de classification est obtenu par le jeu de descripteurs TSPMFCC-Ind (approche globale, SVM Multiclasse) avec un score de 95%. Il est suivi de près par le jeu MCPEA (approche globale, SVM Multilabels) avec un score de 92.69%. Les deux jeux précédents sont issus de l'approche globale, ce qui tend à prouver la supériorité de cette approche par rapport à l'approche instantanée ; Sans omettre, que de plus, l'approche globale est moins coûteuse en temps de calcul et est moins gourmande en capacité de stockage. Nous avons quand même réussi à rehausser le score de l'approche instantanée à 85% en adoptant une stratégie de fusion des classes. Quant aux jeux de descripteurs eux-mêmes, nous avons pu mettre en évidence l'impact des coefficients MFCC sur les taux de classification. En effet dans la catégorie TSPMFCC, les combinaisons qui contiennent les MFCC (leurs dérivées premières ou secondes), présentent toujours des scores supérieurs aux autres. L'introduction des descripteurs indépendants a permis d'introduire une nouvelle classe de descripteurs qui bien qu'améliorant le taux de classification a l'inconvénient de créer des descripteurs sans sens physique réel. Ce type de descripteurs aurait donc un intérêt limité pour d'autres applications liées au signal musical. Le jeu de descripteurs Ondelettes MCPEA donne de bonnes performances tout en ne nécessitant que peu de capacités de calculs et de stockage. Son avantage pour les signaux musicaux réside dans le fait que les coefficients obtenus peuvent être interprétés et associés aux caractéristiques perceptuelles du signal musical. Quant aux machines d'apprentissage, nous en avons testé deux : le PMC et les SVM. Pour le même jeu de descripteurs TSPMFCC et l'approche globale, le PMC culmine au score de 86.11% alors que le SVM obtient 35% ; Cependant le PMC nécessite une étape d'apprentissage et un choix d'architecture dont le SVM nous libère. Entre le SVM et SVM Multilabels et pour le même type de descripteurs (MCPEA), le second permet une nette

amélioration du score de classification. Il apparaît donc que le choix de la machine d'apprentissage pourrait être un facteur important pour la classification en genres sans pour autant surpasser celle du choix des descripteurs.

CONCLUSIONS

Nous avons exploré trois modules d'une chaîne d'indexation et de classification de signaux audio : le module segmentation de signaux audio, le module séparation de sources sonores et le module classification en genres de signaux musicaux.

Le module segmentation de signaux audio que nous avons proposé permet la séparation des classes de signaux audio en deux classes : la classe musique et la classe non-musique. En effet, les signaux audio contiennent principalement trois types de sons : bruit, musique, parole ; Et puisque nous visons comme application finale la classification en genres de signaux musicaux, il était nécessaire d'extraire la musique (instrumentale ou chantée) du reste. L'algorithme mis en œuvre est appliqué à la base de données KASSEM. Il utilise l'Analyse en Composantes Principales combinée au Critère d'Alignement. Il réalise une segmentation temporelle des segments audio. A la sortie de ce module, nous savons où se situent les sons musicaux. La F mesure (une combinaison entre le Rappel et la précision) est utilisée pour évaluer la qualité de la segmentation. L'influence de divers paramètres a été étudiée : normalisation de l'énergie, dimension des descripteurs, durée des fichiers de référence et taille de la fenêtre de lissage. Cette étude a été menée sur des fichiers de développement. Elle a permis de mettre en évidence les paramètres optimaux de la méthode pour des fichiers sonores de la base de données considérée : 16 coefficients MFCC avec énergie normalisée, une taille des fichiers de référence de 10 secondes et une taille de la fenêtre de lissage de 5 secondes.

Le module séparation de sources sonores se doit de suivre le module segmentation ; En effet, les segments obtenus contiennent de la musique et de la parole et éventuellement du bruit : il faudra donc séparer ces trois sources. Deux méthodes de séparation de sources ont été mises en œuvre : l'Analyse en Composantes Indépendantes et la Décomposition Empirique Modale. Pour évaluer les deux méthodes, nous avons utilisé plusieurs types de mélanges et procédé à la séparation. Les signaux obtenus sont visuellement identiques aux signaux originaux ; Des tests d'écoute le confirment également. Mais cela reste subjectif et ne permet pas une automatisation du procédé. Nous avons alors introduit deux critères objectifs de séparation : les critères DSR et ISR. Ces deux critères nous ont donc permis de tester la qualité de séparation en présence de bruit. Ainsi pour différents rapports signal à bruit, nous avons relevé les valeurs des ISR et DSR et observé que les valeurs de DSR et ISR diminuent lorsque le SNR augmentait (ce qui prouve une meilleure qualité de séparation des sources sonores). Nous remarquons cependant que les courbes de DSR et ISR associés à la méthode FASTICA sont toujours plus faibles que ceux de l'EMD, notamment pour des signaux fortement bruités (SNR négatif). Ceci nous permet de conclure que la méthode de décomposition par modes empiriques (EMD) fournit des résultats bien plus performants que

l'Analyse en Composantes Indépendantes bien que leurs performances se rejoignent pour des valeurs élevées du SNR. L'EMD présente en plus l'avantage, du point de vue théorique, de ne pas prendre d'hypothèse sur la nature des mélanges : elle fonctionne aussi bien pour des mélanges convolutifs que linéaires.

Le module classification en genres est une suite possible aux autres modules sans être obligatoire : c'est une application parmi d'autres que nous avons envisagée. Nous avons décidé de concentrer nos efforts sur le choix des descripteurs constituant les vecteurs acoustiques représentatifs de chaque extrait musical. Nous avons étudié l'évolution du taux de classification pour différentes combinaisons de descripteurs. Les descripteurs utilisés sont des descripteurs temporels, spectraux et perceptuels ainsi que des descripteurs ondelettes. Nous avons également introduit la notion de descripteurs indépendants : un jeu de descripteurs pour lequel les descripteurs à l'intérieur du vecteur acoustique sont indépendants. Deux approches ont été testées pour constituer le vecteur acoustique : l'approche globale qui ne conserve que les moyennes et variances des coefficients sur la durée du signal et l'approche instantanée qui conserve tous les coefficients calculés sur chaque fenêtre d'analyse. Le but étant tout de même une classification, nous avons mis en œuvre deux machines d'apprentissage : le perceptron multicouche (PMC) et les machines à vecteurs de support (SVM). Une variante des SVM, les SVM Multilabels, dont la formulation nous a semblé correspondre au type de problème posé, a été proposée. Nous sommes arrivés à la conclusion que l'association du vecteur acoustique composé de descripteurs temporels, spectraux et perceptuels, rendus indépendants grâce à l'Analyse en composantes indépendantes (TSPMFCC-Ind) et du classifieur SVM permettait d'atteindre le score de 95%. Nous avons quand même émis une réserve quant à l'interprétation physique de ces descripteurs. Nous avons donc privilégié l'association vecteurs acoustiques ondelettes (MCPEA : Rapport des moyennes des sous bande adjacentes par échelle en valeurs absolues) et SVM Multilabels, qui présente le score de classification 92.96%, tout en préservant le sens physique des descripteurs.

Conclusion Générale et Perspectives

Conclusions

Cette thèse a porté sur l'étude et l'implémentation d'un système d'analyse et d'indexation de signaux audio avec un intérêt particulier pour les signaux musicaux. De telles travaux sont assez rare dans la littérature dans le sens où notre contribution porte sur les aspects méthodologiques et pratiques : nos travaux ont cerné l'ensemble des modules intervenant dans la chaîne d'indexation, depuis la paramétrisation acoustique jusqu'à la classification en genres musicaux et sur l'aspect applicatif, étant donné la nature et sources des signaux audio considérés (base de données audio orientées de genres et thèmes variés).

Concernant le module de la paramétrisation acoustique nos travaux ont d'abord cerné l'ensemble des descripteurs de l'état de l'art caractéristiques des signaux musicaux, puis nous avons introduit des variations et combinaisons permettant d'inclure des descripteurs issues de l'analyse temps-échelle et d'étudier leurs décorrélations par les méthodes d'analyse en composantes indépendantes.

Poursuivant notre chaîne de traitements, nous avons étudié les méthodes de segmentation audio et notamment les algorithmes de séparation parole/musique et musique/non musique. Dans ce cadre, nous avons proposé un nouvel algorithme original de segmentation audio basé sur le critère de l'alignement de matrices issues de la théorie des noyaux. Cette méthode de classification semi supervisée fait appel à une méthode combinant l'analyse en composantes principales et le critère d'alignement de noyaux introduit et souvent exploité dans l'apprentissage statistique. Cette contribution constitue une des rares applications de ce critère en traitement de signaux audio.

Nous avons montré la faisabilité de cette méthode et évalué ces performances sur une base de signaux sonores issues d'enregistrements radiodiffusés dont nous avons varié les durées et le contenu. Les résultats établis sont prometteurs et permettent d'envisager son utilisation alternative aux autres méthodes d'indexation de l'état de l'art.

A l'issue de ce module, nous récupérons des signaux de audio composés de mélanges musique et voix chantée et de bruits divers. Afin de séparer ces composantes, nous avons implémenté un module de séparation de sources audio se basant sur les méthodes de la décomposition empirique en modes et l'analyse en composantes indépendantes. Une étude comparative de ces deux méthodes a été réalisée.

Nous avons par la suite concentré nos efforts sur les méthodes de classification en genres musicaux. Nous avons analysé et comparé des méthodes génératives (GMM, knn et RN) et discriminatives (SVM) et évalué leurs performances sur une base de signaux audio algérienne radiodiffusés de thématiques et genres divers. Les résultats obtenus sont prometteurs et permettent d'envisager leurs investigations et extensions futures, notamment en classification multi-labels qui constitue à notre connaissance, une voie novatrice pour le traitement de signaux audio.

Perspectives

A l'issue de ce travail, loin d'être complet, nous envisageons des travaux futurs sur les modules d'extraction et de combinaison de descripteurs acoustiques. Nous avons entamé un travail pertinent sur la décorrélation de descripteurs et de réduction de dimensionnalité avec l'analyse en composantes indépendantes qui mérite un développement et une investigation plus poussée afin d'aboutir un travail à soumettre à la communauté internationale.

Une autre perspective intéressante est notamment celle relative à la classification multi-labels dont les résultats préliminaires exposés dans le chapitre résultats ont montré la faisabilité et la pertinence.

Enfin et ce n'est pas des moindres tâches, un travail sur les modules d'indexation lui-même mérite d'être poursuivie afin de fédérer harmonieusement l'ensemble des modules de la chaîne de traitements.

Bibliographie Globale

- [1] Zills A., *Extraction des descripteurs musicaux : une approche évolutionniste*
Thèse de doctorat de l'université de Paris 6 . septembre 2004.
- [2] Vincent E., *Modèles d'instruments pour la séparation des sources et la transcription d'enregistrement musicaux*, thèse de doctorat, université Paris VI à Pierre et Marie curie, Décembre 2003
- [3] Saunders I., « *Real time discrimination of broadcast speech/music* » In Proc. Int. Conf. Acoustics, Speech, Signal Processing (ICASSP) Pages 993 à 996. 1996
- [4] Scheirer E. and Slaney M., *Construction and evaluation of a robust multifeature speech/music discriminator*, in Proc. Int. Conf. Acoustics, Speech, Signal Processing (ICASSP) pages. 1331 à 1334, 1997
- [5] Foote J.T., *A similarity measure for automatic audio classification*, In Proc. Of AAAI 1997 Spring Symposium on Intelligent and Use of Text, Image, Video, and Audio Corpora, Stanford Mars 1997
- [6] Tzanetakis G. and Cook P., *Musical Genre Classification of audio signals* in *IEEE Transactions on Speech and Audio Processing*, 2002.
- [7] Carré M. et Philippe P., *indexation audio , un état de l'art*, in Annales des Telecommunications, septembre 2000.
- [8] Brown J.C., *Determination of the meter of musical scores by autocorrelation*, in Journal of the Acoustical Society of America, 94(4) :1 pp.1957, 1993
- [9] Large E.W. , *Beat tracking with a nonlinear oscillator*, In Working Notes of the IJCAI Workshop on AI and Music, pages 24 à 31, 1995.
- [11] Tzanetakis G., Essl G. and Cook P., « *Audio analysis using the discrete wavelet transform* In Proc. Conf. Acoustics and Music Theory Applications, Sept. 2001.
- [12] Tolonen T. and Karjalainen M. , *A computationally efficient multipitch analysis model*, in *IEEE Trans. Speech Audio Processing*, vol. 8, pages 708 à 716, Nov. 2000.
- [13] Mallat S.G. , *A Wavelet Tour of Signal Processing*. New York: Academic Press 1999.
- [14] Harb H., *Classification du signal sonore en vue d'une indexation par le Contenu des documents multimédia* Thèse de doctorat de l'école centrale de Lyon . septembre 2004 .
- [15] Ajmera J. « *Robust Audio Segmentation* » In PhD Thesis, Ecole Polytechnique Fédérale de Lausane EPFL , 2004
- [16] Bishop M. C., « *Pattern Recognition and Machine Learning* » Springer, Jordan M., Kleinberg J., Schölkopf B. Eds. (2006)
- [17] Burges C. « *A tutorial on Support Vector Machines for pattern recognition* » in Data Mining and Knowledge Discovery 2 (2). . (1998),

- [18] Cristiannini N., Shaw-Taylor J., Elisseef A. and Kandola J. *On kernel target alignment* Advances in Neural Informations Processing Systems, Vol.14, pp. 367-373. (2002).
- [19] Duda R.O., Hart P.E., Stork G. *Pattern Classification* in 2nd Ed., Wiley, NY. . (2001)
- [20] Essid S. *Classification automatique de signaux audio-fréquence : reconnaissance des instruments de musique* PhD Thesis, Université Pierre et Marie Curie . (2005).
- [21] Fergani B. *Application des Méthodes à Vecteurs de Support pour l'indexation en locuteurs de documents audio* PhD Thesis Université Technique d'Alger (USTHB), (2007)
- [22] Fergani L., Fergani B., Houacine A., *Segmentation de signaux audio : Une nouvelle approche utilisant le critère d'alignement*, in Canadian Acoustics, Vol 38 N°2, pp. 3-10. (2010)
- [23] Fergani L. et Houacine A., *Genre classification of musical signals* in Conférence IMESE'06 Djelfa, Algeria, 2006.
- [24] Fergani L. et Houacine A., *Noisy Audio Sources Separation by Empirical Mode Decomposition and Independent Component Analysis* in 19th International Congress on Acoustics, Madrid ,Sept.2007
- [25] Fergani L. et Houacine A., *Artificial Neural Network Genre Classification of Musical Signals*, in SETIT'07 ; Tunisia 2007.
- [26] Fergani L. et A. Houacine, *On the choice of audio descriptors for artificial neural network genre classification of musical signals* in ICCISS International Congress, Vienna 2007
- [27] Fergani L. et Houacine A., *Empirical Mode Decomposition and Independent Component Analysis for noisy audio sources separation* in 2nd International Conference on Electrical and Electronic Engineering ICEEE'08, Laghouat, Algeria 2008.
- [28] Petiot Jean-François, Fergani L. et Lucas marie, *Analyse signal et étude perceptive de sons de trompettes* in Congres Français d'Acoustique (CFA) Troie, 2006
- [29] Herbrich R. *Learning Kernel Classifiers: Theory and Algorithms* MIT Press, Cambridge USA. (2002),
- [30] Llorente Godino J. I. and Co *Discriminative methods for the detection of voice disorders* NoLisp, April 2005.
- [31] Mauclair Julie, *Mesures de confiance en traitement automatique de la parole et applications*, PhD Thesis Université du Maine, Laboratoire d'Informatique de l'Université du Maine. (2006)
- [32] Meignier S., Moraru D., Fredouille C., Besacier L., Bonastre J.F. (2004) *Benefits of prior acoustic segmentation for automatic speaker segmentation* IEEE ICASSP 2004, Montreal Canada.
- [33] Pinquier J.. *Indexation Sonore: recherche de composantes primaires pour une structuration audiovisuelle* PhD Thesis, Université Paul Sabatier , Toulouse III. (2004).
- [34] Ulusay I., Bishop M. C.), *Comparison of Generative and Discriminative Techniques for Object Detection and Classification* LNCS4170 Springer-Verlag. (2006)

- [35] Vert J.P., Tsuda K., Schölkopf B. *A primer on kernel methods* Kernel Methods in Computational Biology, Schölkopf B, Tsuda K and Vert J.P. Eds. Cambridge MA: the MIT Press, pp.35-70. . (2004).
- [36] Young S. and all , *The HTK Book* Ver.3.2.1 (2002),
- [37] Hyvarinen A., *Survey on independant component analysis* Neural Computing Surveys, 2 : 94-128, 1999.
- [38] Cédric Févotte, *Approche temps fréquence pour la séparation aveugle de source non stationnaires*, PhD thesis, Ecole Centrale de Nantes et Université de Nantes, Octobre 2003.
- [39] Flandrin P., Rilling G. and Gonçalves P. «*Empirical Mode Decomposition as a filter bank*», in IEEE Sig.Proc.Lett.,Vol.11, N°2, pp. 112-114.
- [40] Aissa-El-Bey A., Abed-Meraim K. et Grenier Y., «*Séparation de sources aveugle sous-déterminée basée sur l'algorithm EMD*», 20ème Colloque GRETSI, Septembre 2005.
- [41] Aissa-El-Bey A., Abed-Meraim K. et Grenier Y., «*Blind Separation of audio sources using Modal Decomposition*», 8th ISSPA, Vol.2, pp.451-454, Sydney
- [42] Bello, J.P. & Pickens J., *A robust mid-level representation for harmonic content in music signals* in Proc. Of Int. conf. on Music Information Retrieval (ISMIR), pp. 304-311, 2005.
- [43] Cornuz G., Ravelli E., Leveau P. & Duadet L., *Object coding of harmonic sounds using sparse and structured representations* in Proc. of COST G-6 Conference on Digital Audio Effects (DAFX), 2007.
- [44] Daudet L., *Sparse and structured decompositions of signals with the molecular matching pursuit* in IEEE Trans. on Audio, Speech and Language Processing, pp.1808-1819, 2006.
- [45] Ellis D. & Poliner G.J., *Classification-based melody transcription* Machine Learning 65(2):439-456, 2006.
- [46] Escoda O., Granai L. & Vandergheynst P., *On the use of a priori information for sparse signal approximations* in IEEE Trans. on Signal Processing (54):3468-3482, 2006.
- [47] Klapuri A. & Davy M. (eds.), *Signal Processing Methods for Music Transcription* Springer, New York, NY, 2006.
- [48] Li T. & Ogihara M., *Music genre classification with taxonomy* in Proc. of IEEE Int. Conf. on Audio, Speech and Signal Processing (ICASSP) 2005.
- [49] Serra, X. , *A System for Sound Analysis/Transformation/Synthesis based on a Deterministic plus Stochastic Decomposition*, PhD thesis, Stanford University, 1999.
- [50] Vincent, E., *Musical source separation using time-frequency source priors* in IEEE Trans. on Audio, Speech and Language Processing 14(1) : 91-98, 2006.
- [51] Virtanen, T., *Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria* in IEEE Trans. on Audio, Speech

and Language Processing **15**(3) : 1066–1074, 2007.

[52] Rouvier M., Linares G. & Matrouf D., “Robust audio-based classification of video genre” In Proceedings of INTERSPEECH, p. 1159–1162, 2009

[53] DeCoro, C., Z. Barutcuoglu, and R. Fiebrink., “Bayesian aggregation for hierarchical genre classification” in *Proceedings of the International Conference on Music Information Retrieval*. 77–80, 2007.

[54] Hu, X., J. S. Downie, and A. F. Ehman., “Lyric text mining in music mood classification” in *Proceedings of the International Society for Music Information Retrieval Conference*. 411–416, 2009.

[55] Lidy, T., A. Rauber, A. Pertusa, and J. M. Iñesta. , “Improving genre classification by combination of audio and symbolic descriptors using a transcription system” in *Proceedings of the International Conference on Music Information Retrieval*. 61–6, 2007.

[56] McKay, C., “Automatic genre classification of MIDI recordings” in *M.A. Thesis*. McGill University, Canada, 2004.

[57] McKay, C., and I. Fujinaga., “WebMiner: A web-based feature extractor” in *Proceedings of the International Conference on Music Information Retrieval*. 113–4, 2007.

[58] Aristomenis S. Lampropoulos, Paraskevi S. Lampropoulou, George A. Tsihrintzis: “Music genre classification based on ensemble of signals produced by source separation methods” in *Intelligent Decision Technologies Journal*, IOS Press, Volume 4, Number 3, pp.229-237, 2010.

[59] M. Mandel, G. Poliner, and D. Ellis, “Support vector machine active learning for music retrieval,” in *ACM Multimedia Systems*, vol. 12, no. 1, pp. 3 – 13, 2006.

[60] E. Tsunoo, G. Tzanetakis, N. Ono, and S. Sagayama, “Audio genre classification using percussive pattern clustering combined with timbral features,” in *Proc. of ICME*, 2009, pp. 382 – 385.

[61] E. Tsunoo, N. Ono, and S. Sagayama, “Musical bass-line pattern clustering and its application to audio genre classification,” in *Proc. Of ISMIR*, 2009.

[62] D. Turnbull, L. Barrington, D. Torres, and G. Lanckriet, “Towards musical query-by-semantic-description using the cal500 data set,” in *Proc. of ACM SIGIR*, 2007, pp. 439 – 446.

[63] G. Tzanetakis, *Marsyas-0.2: A Case Study in Implementing Music Information Retrieval System*, chapter 2, pp. 31 – 49, Idea Group Reference, 2007, Shen, Shepherd, Cui, Liu (eds).

[65] I. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*, Morgan Kaufmann, 2005.

[66] J. Bergstra, N. Casagrande, D. Erhan, D. Eck, and B. Kégl, “Aggregate features and ADABOOST for music classification,” *Machine Learning*, vol. 65, no. 2-3, pp. 473 – 484, 2006.

- [67] A. Holzapfel and Y. Stylianou, "Musical genre classification using nonnegative matrix factorization-based features," *IEEE Trans. on Audio, Speech and Lang. Proc.*, vol. 16, no. 2, pp. 424–434, 2008.
- [68] S. Kiranyaz, A. F. Qureshi, and M. Gabbouj, "A generic audio classification and segmentation approach for multimedia indexing and retrieval," in *IEEE Trans. on Audio, Speech and Lang. Proc.*, vol. 14, no. 3, pp. 1062–1081, 2006.
- [69] E. Ravelli, G. Richard, and L. Daudet, "Union of MDCT bases for audio coding," in *IEEE Trans. on Audio, Speech and Lang. Proc.*, vol. 16, no. 8, pp. 1361–1372, Nov. 2008.
- [70] S. Hainsworth, "Techniques for the automated analysis of musical audio," Ph.D. dissertation, Dept. Eng., Cambridge University, 2004.
- [71] M. D. Plumbley, S. A. Abdallah, T. Blumensath, and M. E. Davies, "Sparse representations of polyphonic music," *Signal Processing*, vol. 86, no. 3, pp. 417–431, Mar. 2006.
- [72] L. Benaroya, F. Bimbot, and R. Gribonval, "Audio source separation with a single sensor," *IEEE Trans. Audio, Speech and Language Processing*, vol. 14, pp. 191–199, Jan. 2006.
- [73] A. Ozerov, P. Philippe, F. Bimbot, and R. Gribonval, "Adaptation of Bayesian models for single channel source separation and its application to voice / music separation in popular songs," *IEEE Trans. Audio, Speech and Language Processing*, vol. 15, no. 5, pp. 1564–1578, Jul. 2007.
- [74] P. D. O'Grady, B. A. Pearlmutter, and S. T. Rickard, "Survey of sparse and non-sparse methods in source separation," *International Journal of Imaging Systems and Technology*, vol. 15, no. 1, pp. 18–33, 2005.
- [75] L. Daudet and B. Torresani, "Hybrid representations for audiophonic signal encoding," *Signal Processing, special issue on Image and Video Coding Beyond Standards*, vol. 82, no. 11, pp. 1595–1617, 2002.
- [76] S. Mallat, *A Wavelet Tour of Signal Processing: The Sparse Way*, 3rd ed. Academic Press, 2009.
- [77] S. Arberet, R. Gribonval, and F. Bimbot, "A robust method to count and locate audio sources in a stereophonic linear instantaneous mixture," in *Proc. 6th Intl. Conf. on Independent Component Analysis and Blind Signal Separation (ICA 2006)*, Charleston, SC, USA, ser. LNCS 3889, J. Rosca et al., Eds. Springer, Mar. 2006, pp. 536–543.
- [78] B. Manjunath, P. Salembier, and T. Sikora. *Introduction to MPEG-7: Multimedia Content Description Language*. John Wiley & Sons, 2002.
- [79] A. Pikrakis, S. Theodoridis, and D. Kamarotos, "Classification of musical patterns using variable duration hidden Markov models," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 5, pp. 1795–1807, 2006.

- [80] M. Helén and T. Lahti, "Query by example methods for audio signals," in *Proceedings of the 7th Nordic Signal Processing Symposium (NORSIG 2006)*, pp. 302–305, Reykjavik, Iceland, June 2006.
- [81] M. Mandel and D. Ellis, "Song-level features and support vector machines for music classification," in *Proceedings of the 6th International Conference on Music Information Retrieval (ISMIR 2005)*, London, UK, September 2005.
- [82] Guo, G., Li, S.Z.: Content-based Audio Classification and Retrieval by Support Vector Machines. *IEEE Transactions on Neural Networks* 14(1), 209–215 (2003)
- [83] Lin, C.C., Chen, S.H., Truong, T.K., Chang, Y.: Audio Classification and Categorization Based on Wavelets and Support Vector Machine. *IEEE Transactions on Speech and Audio Processing* 13(5), 644–651 (2005).
- [84] Ogle, J. and Ellis, D., "Fingerprinting to identify repeated sound events in long-duration personal audio recordings," in *Proc. ICASSP 2007*.
- [85] Master, A. S. and Liu, Y.-W. "Non-stationary sinusoidal modeling with efficient estimation of linear frequency chirp parameters," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2003.
- [86] Lagrange, S., Delanoue, N., and Jaulin, L., "On sufficient conditions of injectivity, development of a numerical test via interval analysis," in *Reliable Computing*, 13(5) :409–421, 2007.
- [87] Jensen, J., Heusdens, R., and Jensen, S. H. , "A perceptual subspace approach for modeling of speech and audio signals with damped sinusoids," in *IEEE Trans. Speech Audio*, 12(2) :121–132, 2004.
- [88] de Cheveigné, A. and Kawahara, H. Yin, "a fundamental frequency estimator for speech and music," in *Journal of the Acoustical Society of America*, 111(4), 2002.
- [89] Cano, P., Batlle, E., Gaomez, E., Gomes, L., and Bonnet, M., "Audio fingerprinting : concepts and applications," in *Studies in computational intelligence*, 2 :233–245, 2005.
- [90] Betser, M., Collen, P., David, B., and Richard, G., "Review and discussion on STFT-based frequency estimation methods," in *Audio Engineering Society 120th Convention*, 2006.
- [91] Badeau, R., "Méthodes à haute résolution pour l'estimation et le suivi de sinusoïdes modulées. Application aux signaux de musique," In PhD thesis, Telecom Paris, 2005.
- [92] D. FitzGerald. Automatic Drum Transcription and Source Separation. PhD thesis, Dublin Institute of Technology, 2004.
- [93] R. Fiebrink et I. Fujinaga. Feature selection pitfalls and music classification. In *Proceedings of the 7th International Conference on Music Information Retrieval (ISMIR 06)*, 2006.

- [94] M. Davy, S. Godsill, et J. Idier. Bayesian analysis of polyphonic western tonal music. *Journal of the Acoustical Society of America*, 119(4):2498-2517, April 2006.
- [95] M. Davy et S. Godsill. Detection of abrupt spectral changes using support vector machines : an application to audio signal segmentation. In *Proceedings of the 2002 IEEE Conference on Acoustics, Speech and Signal Processing (ICASSP02)*, 2002.
- [96] S. Canu et A. Smola. Kernel methods and the exponential family. In *Proceedings of the 13th European Symposium on Artificial Neural Networks (ESANN05)*, 2005.
- [97] P. Chen, C. Lin, et B. Schölkopf. A tutorial on ℓ_1 -support vector machines. In *Applied Stochastic Models in Business and Industry*, volume 21, 2, pages 111-136, 2005.
- [98] C. C. Chang et C. J. Lin. LibSVM : a library for Support Vector Machines. Software available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm>, 2001.
- [99] M. Cooper et J. Foote. Automatic Music Summarization via Similarity Analysis. In *Proceedings of the 3rd International Conference on Music Information Retrieval (ISMIR02)*, 2002.
- [100] M. Casey. MPEG-7 sound-recognition tools. *IEEE Transactions on Circuits and Systems for Video Technology*, 11(6):737-747, 2001.

Contributions de l'auteur

Publications dans revues internationales

- [1] **Fergani L.**, Fergani B., Houacine A., "*Segmentation de signaux audio : Une nouvelle approche utilisant le critère d'alignement*" in *Canadian Acoustics*, Vol 38 N°2, pp. 3-10. (2010)

Conférences et Communications nationales et internationales

- [1] **Fergani L.** et Houacine A., "*Genre classification of musical signals*" in *Conférence IMESE06* Djelfa, Algeria, 2006.
- [2] **Fergani L.** et Houacine A., "*Noisy Audio Sources Separation by Empirical Mode Decomposition and Independent Component Analysis*" in 19th International Congress on Acoustics, Madrid ,Sept.2007
- [3] **Fergani L.** et Houacine A., "*Artificial Neural Network Genre Classification of Musical Signals*" in *SETIT07* ; Tunisia 2007.
- [4] **Fergani L.** et Houacine A., "*Empirical Mode Decomposition and Independent Component Analysis for noisy audio sources separation*" in 2nd International Conference on Electrical and Electronic Engineering ICEEE08, Laghouat, Algeria 2008.
- [5] Petiot Jean-François, **Fergani L.** et Lucas marie, "*Analyse signal et étude perceptive de sons de trompettes*" in *Congres Français d'Acoustique (CFA)* Troie, France 2006