UNIVERSITÉ DU QUÉBEC À CHICOUTIMI

MÉMOIRE

PRÉSENTÉ À

L'UNIVERSITÉ DU QUÉBEC À CHICOUTIMI

COMME EXIGENCE PARTIELLE

DU PROGRAMME DE MAÎTRISE EN INGÉNIERIE EN VUE DE L'OBTENTION DU GRADE DE MAÎTRE ÈS EN SCIENCES APPLIQUÉES (M. SC. A.).

PAR

GLENN HALL

IDENTIFICATION D'INSTRUMENTS DE MUSIQUE À L'AIDE DE MÉTHODES STATISTIQUES ET D'ALGORITHMES D'INTELLIGENCE ARTIFICIELLE

mars 2013

RÉSUMÉ

Avec l'explosion des médias d'information, particulièrement celui d'Internet et des formats audio numériques, la quantité de musique disponible sur le marché induit une tâche colossale de maintenance, de classification et d'authentification, d'où l'importance incontestable de méthodes efficaces d'analyse automatique de la musique. L'idée première de cette analyse est de dériver les informations d'un signal sonore brut (préalablement numérisé) et de les transposer sous forme de données symboliques (nom de l'instrument, partition, style musical, etc.) intelligibles et réutilisables par différents procédés informatiques.

Ce présent mémoire traite de l'identification des instruments de musique; l'idée première est de construire un système automatisé capable de déterminer quel instrument de musique est joué à partir d'un son musical. Le matériel expérimental fut constitué d'un ensemble de 6 698 notes isolées provenant de 29 instruments de musique occidentaux, lesquels sont fréquemment utilisés comme référence. Les travaux les plus cités dans le domaine de l'identification des instruments de musique sont introduits et les descripteurs classiques ainsi que les classificateurs usuels sont décrits.

Différents descripteurs classiques, tels que les coefficients MFCC et LPC, les moments spectraux, les moments et la pente de l'enveloppe, le temps d'attaque et le taux de passage par zéro, ont été utilisés pour construire les vecteurs d'observations d'un système de classification. Un nouveau descripteur, le chromatimbre, fut introduit et évalué. De plus, les performances de chacun des groupes de paramètres prient individuellement furent comparées. L'effet de la normalisation sur les vecteurs d'observation fut examiné avec les normalisations mu-sigma et min-max. Deux classificateurs usuels, les k plus proches voisins ainsi que le modèle de mélange de gaussiennes, furent utilisés. Différentes variantes d'un algorithme de sélection séquentielle des paramètres permirent d'augmenter les performances des systèmes de classification. Entre autre, un système de classification hiérarchique, ayant obtenu un score d'identification des instruments de 88,32% et un score d'identification des familles de 94,74%, fut comparé à un système de classification directe : un gain de plus de 2% fut observé entre les deux approches. Différentes expérimentations mirent en évidence l'importance d'adapter la sélection des paramètres à chaque nœud de la classification hiérarchique contrairement à l'utilisation d'un vecteur d'observation statique, dont les paramètres ne varient pas en fonction du nœud.

REMERCIEMENTS

Je désire remercier ma conjointe Nadine, pour sa compréhension et son appui offerts pendant la réalisation de ce projet de Maîtrise.

Mes remerciements s'adressent également à mon directeur de thèse, M. Hassan Ezzaidi, pour ses efforts soutenus, sa patiente, son amabilité et sa compétence. Il fut un guide exceptionnel. Je tiens aussi à remercier mon co-directeur, M. Mohammed Bahoura, ainsi que les évaluateurs, M. Jean Rouat et M. Christophe Volat.

Je tiens finalement à remercier mes parents pour leurs encouragements et leur soutient indéfectible tout au long de ma formation académique m'ayant permis d'atteindre et de poursuivre ce programme de Maîtrise.

TABLE DES MATIÈRES

LISTE DES FIGURES

LISTE DES TABLEAUX

LISTE DES ABREVIATIONS

Acronyme	Signification	
ADSR	Attack, decay, sustain, release.	Attaque, diminution, maintient, relâchement
AM	Amplitude Modulation	Modulation en amplitude
ANSI	American National Standard Institute	-
CASA	Computational Auditory Scene Analysis	Analyse computationnelle de scène auditive
CQT	Constant-Q Transform	Transformée à facteur *Q* constant
DCT	Descrete Consine Transform	Transformée en cosinus discrète
EM	Expectation-Maximization	Espérance-maximisation
EOF	Empirical Orthogonal Function	Décomposition orthogonale aux valeurs propres
FFT	Fast Fourier Transform	Transformée de Fourier rapide
FIR	Finite Impulse Response	Réponse impulsionnelle finie
HL	Hearing Level	Seuil d'audition
k-NN	k-Nearest Neighbour	k plus proches voisins
LDA	Linear Discriminant Analysis	Analyse discriminante linéaire
LPC	Linear Prediction Coefficients	Coefficients de prédiction linéaire
MFCC	Mel-Frequency Cepstral Coefficients	Coefficients cepstraux sur l'échelle MEL
MIR	Music Information Retrieval	Récupération d'information musicale
MIS	Musical Instrument Sample	-
MSE	Mean Squared Error	Erreur quadratique moyenne
MUMS	McGill University Master Samples	-
PCA	Principal Component Analysis	Analyse en composante principale
PSE	Point of Subjective Equality	Point d'égalité subjective
RWC	Real World Computing	-
SC	Spectral Centroid	Barycentre spectral
SFS	Sequential Forward Selection	Sélection séquentielle en avant
SK	Spectral Kurtosis	Asymétrie spectrale
SL	Sensation Level	Niveau de sensation
SOM	Self-Organizing Map	Carte auto-organisatrice

Acronyme	Signification	
SPL	Sound Pressure Level	Niveau de pression acoustique
SS	Spectral Skewness	Asymétrie spectrale
SVD	Singular Value Decomposition	Décomposition en valeurs singulières
SW	Spectral Width	Largeur spectrale
VSL	Vienna Symphonic Library	-
WAV	WAVeform Audio File Format	-
ZRC	Zero Crossing Rate	Taux de passages par zéro
LTI	Linear Time-Invariant	Linéaire et invariant dans le temps
ARMA	Autoregressive Moving-Average	Autorégressif et moyenne mobile

CHAPITRE 1

INTRODUCTION

1.1 Problématique

L'analyse musicale faite par le musicologue, le musicien ou l'audiophile se concrétisera sous différents aspects techniques : pour l'un l'obtention du rythme, de l'intonation, de l'échelle musicale ainsi que de l'interprétation contiennent l'information pertinente. Pour l'autre, l'écriture de la partition et la sélection instrumentale seront appropriées à la description musicale tandis que la majorité s'intéressera aux styles musicaux (classification, expérience et expressions de l'auditeur, émotions, etc.). Toutefois, chacun d'eux se servira du signal sonore pour acquérir les expertises respectives. Avec l'explosion des médias d'information, particulièrement celui d'Internet et des formats audio numériques, la quantité de musique disponible sur le marché induit une tâche gigantesque de maintenance, de classification et d'authentification, d'où l'importance incontestable de méthodes efficaces et effectives d'analyse automatique de la musique. L'idée première de cette analyse est de dériver les informations d'un signal sonore numérique et de les transposer sous forme de données symboliques intelligibles et réutilisables par différents procédés informatiques.

Ces méthodes d'analyse automatique de la musique, encore au stade embryonnaire, évoquent des défis spécifiques dont plusieurs sont abordés par les chercheurs sous différents angles, tant

par la mise en œuvre de théories classiques que l'élaboration de nouvelles techniques originales. Les propositions proviennent en grande partie des domaines connexes inhérents aux sciences du traitement du signal, de la psycho-acoustique et de l'intelligence artificielle, comme c'est souvent le cas pour les avancées en traitement de la parole.

Certains parallèles sont possibles entre parole et musique (Tableau 1.1). La prosodie d'une élocution est l'équivalent d'un contour mélodique d'une phrase musicale et le texte est à la parole ce que la partition est à la musique. La tâche de reconnaissance des instruments de musique est comparable à celle de la reconnaissance du locuteur. Dans les faits, le traitement de la parole fait l'objet de recherches abondantes et les performances des algorithmes sont excellentes dans bien des contextes, ce qui n'est malheureusement pas encore le cas avec la reconnaissance des instruments de musique. En effet, si l'identité du locuteur est unique, celle d'un instrument de musique varie selon des facteurs contextuels (musicien, fabricant, ajustements, etc.). Il n'existe également pas de consensus sur les descripteurs sonores qui déterminent la signature de l'instrument tandis que l'utilisation du cepstre est acceptée globalement pour la reconnaissance du locuteur.

Tableau 1.1 Parallèles entre parole et musique.

Parole	**Musique**
Anatomie	Organologie descriptive
Physiologie acoustique	Mécanique acoustique
Consonne	Transitoire d'attaque
Voyelle	Maintien ou relâchement
Phonème	Sonème
Phone	Sone
Prosodie	Contour mélodique et rythme
Texte	Partition
Reconnaissance du locuteur	Reconnaissance instrumentale
Reconnaissance de la parole	Reconnaissance des partitions

Les applications d'analyse musicale et le développement de méthodes d'identification d'instrument de musique permettent d'étudier la perception auditive humaine et les propriétés psycho-acoustiques du son. Les recherches effectuées sur le sujet mèneront indéniablement vers une connaissance plus générale de l'analyse sémantique de l'environnement sonore, qu'il soit perçu par un humain ou par un processus computationnel. Notre compréhension actuelle est basée essentiellement sur l'analyse du système auditif humain; les modèles algorithmiques qui en sont dérivés ne peuvent qu'indiquer la voie à suivre. On est encore loin d'une entité artificielle autonome pouvant décrire symboliquement la production d'un son tant la mécanique de l'oreille et de son entendement est mal comprise.

Plus encore, l'acquisition d'information musicale a plusieurs avenues commerciales. Trois principaux consommateurs sont identifiés comme bénéficiaires des avancés technologiques : les industriels qui collectionnent et filtrent, le grand public qui exige une recherche simple et personnalisée et les professionnels tels les musiciens, les enseignants, musicologues et avocats

spécialisés en droits d'auteurs. Plusieurs applications de la vie courante en bénéficient donc. On peut extraire l'information, faire des tris par genre, des recherches par similitudes ou encore par mélodie - fredonnée par exemple -, vérifier les droits d'auteurs, marquer le refrain pour générer des étampes musicale lors de parcours dans une liste de chansons, produire des partitions à partir de musique poly-instrumentale, retirer la plage vocale et régénérer les pistes pour un système de karaoké, obtenir automatiquement la dynamique de la musique (gamme, tempo, etc.) et générer une mélodie d'accompagnement soit pour supporter une vidéo ou encore une improvisation musicale. L'information pourrait être utilisée dans des systèmes experts pour le rehaussement de la qualité de l'enregistrement (alignement temporel des pistes instrumentales, ajout d'effets aux emplacements clés comme les refrains, etc.). Avec l'identification des instruments de musique, on peut notamment effectuer des requêtes sur l'orchestration des pièces musicales mais de plus, l'extraction des notes bénéficie de l'information sur le nombre de sources musicales et le codage audio paramétrique peut adapter les modèles de présentation au contenu instrumental.

1.2 Objectifs

Le présent mémoire traite de l'identification des instruments de musique; en particulier, une classification hiérarchique par famille d'instruments est effectuée comme première étape de ségrégation pour mener à l'identification de l'instrument même. L'étude est située par conséquent majoritairement au niveau de l'analyse du signal sonore, portant sur les traits similaires entre les signaux des instruments de même famille et sur ceux d'un même instrument. Une nouvelle méthode d'analyse du signal est proposée, le chromatimbre, pour représenter l'empreinte de l'instrument. Une attention particulière est aussi portée à la sélection, la

modélisation et l'apprentissage des algorithmes de classification même si ces derniers ne sont pas étudiés et décrits de manière exhaustive.

1.3 Méthodologie

L'environnement de développement Matlab fut utilisé pour la totalité des simulations. Les descripteurs ne faisant pas partie des librairies ont été implantés à partir des librairies de Matlab, en particulier la librairie « Signal Processing Toolbox ». Le processus de classification hiérarchique a été complètement implanté à partir de zéro. Les notes d'un instrument étant jouées séquentiellement par le musicien et donc contenues dans un seul fichier, le code nécessaire à la segmentation de tous les fichiers de la base de données fut également implanté à partir de zéro. L'annexe A contient la liste des librairies utilisées dans le cadre des simulations.

1.4 Structure du mémoire

Le Chapitre 2 donne une description de la nature des instruments de musique qui seront traités par les algorithmes d'identifications. La base de données choisie pour les simulations est également décrite et un résumé des articles principaux traitant de la reconnaissance des instruments de musique avec des notes isolées permet de positionner le présent mémoire dans son contexte de recherche. Dans le Chapitre 3, les notions théoriques des descripteurs usuels du timbre sont abordées sur les plans temporel et fréquentiel. Dans le Chapitre 4, les notions théoriques concernant la reconnaissance de motifs par les méthodes d'apprentissage supervisé sont présentées. Deux méthodes sont proposées, soit la méthode des k plus proches voisins (k-NN) ainsi que la méthode du modèle de mélange de gaussiennes (GMM). Le Chapitre 5 détaille les

simulations qui ont été effectuées et discute de l'analyse des résultats d'une classification directe et d'une classification hiérarchique.

CHAPITRE 2

ÉTAT DE L'ART

Ce chapitre traite des familles d'instruments de musique et présente la taxonomie utilisée dans un contexte de classification hiérarchique. La base de données utilisée pour les travaux de ce mémoire est également présentée ainsi que le protocole d'expérimentation. Un résumé des publications les plus pertinentes sur la reconnaissance des instruments de musique introduit et commente les performances des travaux les plus cités.

2.1 Catégorisation des instruments

En premier lieu, il est à noter que les études de ce mémoire sont faites exclusivement sur les instruments de musique occidentaux, en excluant les percussions. Les instruments de musique occidentaux ont été sujets de plusieurs études tant aux niveaux acoustique que psycho-acoustique. Cela permet d'approfondir les connaissances disponibles et d'apporter des conclusions plausibles en comparant les résultats obtenus aux performances des recherches actuelles.

La catégorisation classique par famille des instruments de musique connus du grand public est celle présentée par Von Hornbostel et Sachs [1]. Plusieurs variantes sont proposées de nos jours ; cette catégorisation, sans faire l'unanimité, reste quand même la référence de base dans la musicologie et chez les auditeurs en général. Le Tableau 2.1 présente un résumé de cette

classification. Les instruments sont ainsi classifiés en trois grandes familles soit les cordophones (instruments à cordes), les aérophones (instruments à vent) et les percussions (cloches, xylophones, drums, etc.).

Tableau 2.1 Catégorisation selon Von Hornbostel et Sachs[1].

Familles	Sous-familles	Instruments
Cordophones	Frottées	Violon Alto Violoncelle Contrebasse
	Pincées	Guitare Harpe Mandoline Ukulélé Clavecin Banjo
	Frappées	Piano
Aérophones	Voix	Chant
	Bois	Hautbois Flûte à bec Piccolo Cor anglais Basson Clarinette Flûte de Pan Saxophone
	Cuivres	Trompette
Percussions	Membranophones	Tambour
	Idiophones à son déterminé	Xylophone Métallophone
	Idiophones à son indéterminé	Triangle

2.1.1 Taxonomie naturelle

La définition de la catégorisation des instruments de musique n'étant pas l'objectif de ce mémoire, l'utilisation de la taxonomie proposée par Martin [2] et réutilisée par Eronen [3] a été considérée. Cela permet, en outre, d'effectuer un comparatif sur les modèles présentement étudiés. De plus, cette taxonomie permet non seulement de regrouper les instruments similaires selon leur mode de production du son mais aussi de les regrouper selon la forme de leur enveloppe, qui est primordiale dans l'entendement du timbre. Cette taxonomie est appelée *taxonomie naturelle* puisqu'elle s'inspire de la classification populaire de Von Hornbostel et Sachs [1].

La taxonomie naturelle de la Figure 2.1, sépare lors de la première étape de ségrégation, les instruments *pizzicato*, dont l'attaque est abrupte, par rapport aux instruments *soutenus*, dont le temps de maintien est constant. Les instruments pizzicato ont comme particularité que la source d'excitation est donnée par une impulsion et le temps de maintien dépend de l'intensité de cette impulsion. Les instruments soutenus ont comme particularité que la source d'excitation est appliquée de façon constante jusqu'au relâchement de la note. Cette caractéristique associée à la continuité d'une note sera référé dans le présent ouvrage par le terme « articulation ».

Au deuxième niveau, les instruments sont regroupés par famille et mode de production. Dans la sous-classe des instruments pizzicato, une seule famille est présente, soit les instruments à cordes. Pour la sous-classe des instruments soutenus, quatre groupes d'instruments sont présents : les cuivres, les flûtes et le piccolo, les instruments à anche et les instruments à cordes.

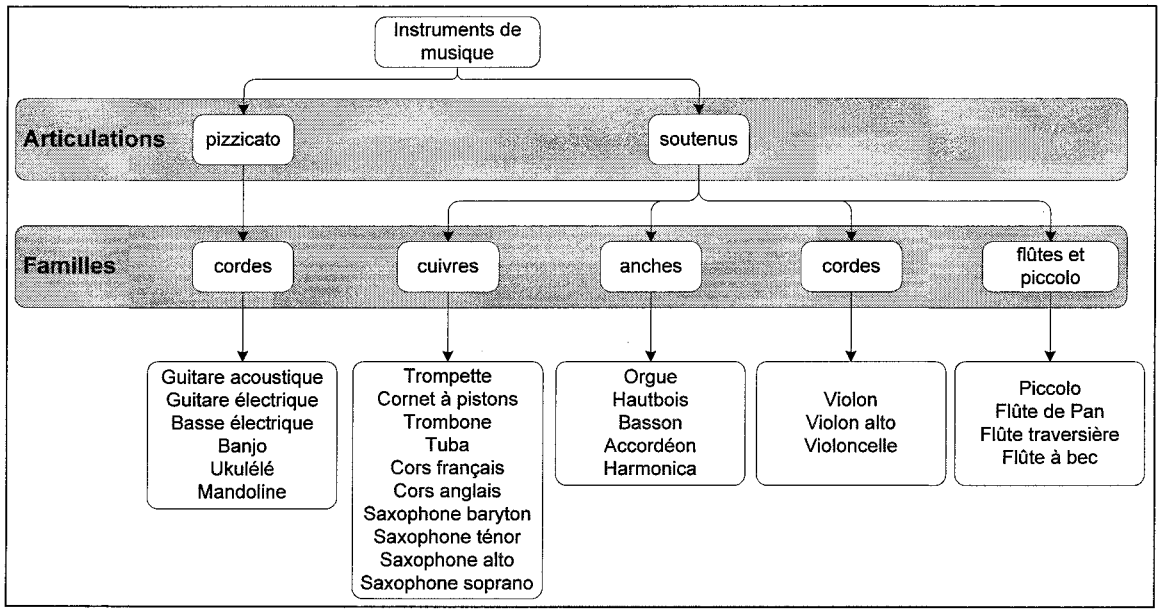

Figure 2.1 Taxonomie de la classification hiérarchique utilisée par Martin [2] et Eronen [3].

2.1.2 Taxonomie automatique

Dans sa thèse, Essid [4] compara les performances de la taxonomie naturelle avec une taxonomie inférée automatiquement à partir d'exemples. Ce processus d'inférence nécessite de déterminer les descripteurs à utiliser pour la construction de la taxonomie ainsi que le choix d'un critère de proximité entre instruments de même classe. Un critère de proximité convenable consiste à utiliser une distance probabiliste, c'est-à-dire une distance entre distributions de probabilités des classes [5]. Essid choisit dans son étude la distance de Bhattacharryya et la divergence (version symétrisée de la distance de Kullback-Leibler). En conclusion de son étude, la classification hiérarchique des instruments de musique basée sur la taxonomie automatique donne de meilleurs résultats que celle basée sur la taxonomie naturelle.

Pour vérifier les performances de la taxonomie naturelle proposée par Martin [2], une taxonomie construite automatiquement à partir de nuages de points fut générée puis utilisée à chaque itération de la classification hiérarchique (Figure 2.2). L'algorithme des *k*-moyennes fut utilisé pour générer les groupes de nuages de points de la taxonomie (en utilisant la corrélation comme mesure de distance entre les points). En particulier, seuls les instruments soutenus furent regroupés par l'algorithme des *k*-moyennes ; le premier niveau hiérarchique, c'est-à-dire l'articulation, démontre d'excellents taux de reconnaissance et fut conservé dans la hiérarchie.

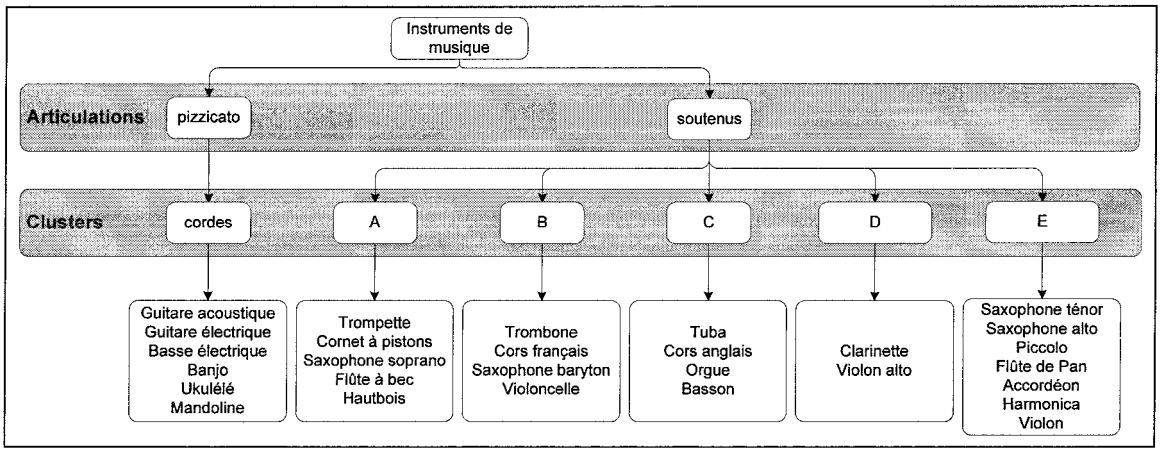

Figure 2.2 Taxonomie hiérarchique obtenue par nuages de points selon les *k*-moyennes.

2.2 Bases de données expérimentales

Cette section présente les données musicales utilisées dans les expérimentations réalisées dans ce mémoire. Une description des bases de données MUMS et RWC est présentée dans un but de comparaison. Un aperçu des bases de données disponibles qui sont régulièrement utilisées dans des contextes de recherches analogues à celui de ce mémoire est également présenté.

La sélection des données expérimentales est fondamentale dans l'évaluation d'un modèle, non seulement à l'élaboration du modèle, mais à la qualification de son rendement. Une surabondance de données entraîne une analyse difficile et des calculs excessifs tandis qu'une quantité limitée de données entraîne des résultats qui reflètent difficilement la réalité. Enfin, des données fortement homogènes conduisent à une hyperspécialisation.

Les résultats des publications actuelles doivent être interprétés avec prudence ; tandis que la construction des modèles est rigoureuse, les simulations sont parfois effectuées avec des données arbitraires : la singularité (un seul instrument par famille) et l'invariance (un seul modèle d'instrument) sont à l'origine de résultats optimistes. De plus, un enregistrement sonore de haute qualité ne reflète pas fidèlement toutes les conditions des milieux acoustiques : présence de nuisances comme bruits, réverbérations, échos, bande passante réduite, signaux incomplets, superpositions des sources sonores, etc. Il existe aussi une multitude de variantes dans la technique de jeu, le modèle, le style et l'âge de l'instrument. On doit pouvoir également comparer les résultats de chacune des recherches non seulement qualitativement mais quantitativement. Les comparaisons sont couramment difficiles puisque les données respectives à chaque étude sont différentes.

2.2.1 Sources

Les principales sources de données sont présentées ici afin de justifier l'ensemble de données qui a été retenu pour réaliser les travaux de recherche de ce mémoire.

2.2.1.1 McGill University Master Samples Collection

La première version de cette base de données date de 1987 et offrait une variété d'instruments disponibles en support CD [6]. Plus récemment, depuis 2006, la base de données contient presque tous les instruments standards classiques et populaires et est distribuée en support DVD [7]. Cette base de données se veut une source centrale dans la recherche, elle est l'une des plus citées dans les publications traitant la classification et la reconnaissance des instruments de musique [8] et est utilisée dans plus de 200 universités à travers le monde pour des objectifs académiques et de recherches [9].

Les fichiers contenus sur les DVDs sont divisés entre les instruments à cordes, claviers, instruments à vent (bois et cuivres) et percussions. En principe, chaque note est enregistrée séparément (44,1 kHz, 24 bits) en stéréo. Malgré une impressionnante couverture de sons, il y a approximativement 29 enregistrements par instrument ce qui implique que la plage de tous les tons possibles n'est pas nécessairement jouée pour tous les instruments. L'uniformité des enregistrements a été remise en question par Eerola et Ferrer [8] ; ces derniers ont identifié des erreurs parmi les fichiers : libellés erronés, hauteur de note dans la mauvaise classe chromatique et dans le mauvais octave, instruments mal accordés. Finalement, l'absence d'une convention de nommage des fichiers et l'inexistence d'un index rendent difficile l'utilisation de la base de données et à la distinction des erreurs dans celle-ci.

2.2.1.2 RWC Music Database: Musical Instrument Sound

Les données de ce présent mémoire proviennent de la base de données « RWC (Real World Computing) Music Database» [10]. Cette base de données est privilégiée par ses droits d'auteur affranchis, sa grande couverture d'instruments de musique et par le nombre de variations pour chaque instrument. Également, une convention de nommage formelle est utilisée et facilite ainsi l'étiquetage des données.

Chaque fichier sonore contient le signal d'un seul instrument joué avec des notes isolées. La base de données fournit plusieurs enregistrements pour chaque instrument. Différents fabricants pour le même instrument et différents musiciens ont participé pour générer les enregistrements et ainsi fournir un éventail de plusieurs signatures instrumentales.

En principe, la base de données contient trois variations pour chaque instrument : trois manufacturiers, trois intensités et trois musiciens différents. Pour chaque instrument, le musicien joue chaque note individuellement à un intervalle d'un demi-ton sur toute la plage possible de l'instrument. Pour ce qui est des instruments à cordes, la plage complète pour chaque corde est jouée. La dynamique est également variée avec des intensités *forte, mezzo* et *piano*.

2.2.1.3 Autres

Il existe quelques alternatives aux bases de données MUMS et RWC tel que la base de données « Musical Instrument Sample » (MIS) [11] de l'Université d'Iowa et la collection « Vienna Symphonic Library » (VSL) produite et commercialisée par l'entreprise autrichienne Vienna Symphonic Library GmbH [12]. Ces deux sources ne sont pas nécessairement conçues pour la

recherche scientifique mais plutôt pour la réalisation d'œuvres musicales à l'aide d'échantillonneurs. De plus, les fichiers de la base de données VSL ne peuvent être lus que par des logiciels spécialisés et les paramètres de production des sons ne sont pas spécifiés.

2.3 Protocole d'évaluation

2.3.1 Prétraitement des données

Les fichiers sources contenant les enregistrements des instruments de musiques sont encodés en format « wav » à une fréquence d'échantillonnage de 44.1 kHz sur un seul canal. Chaque fichier contient la plage complète d'un seul instrument; pour être capable d'effectuer les tests et les analyses (inférences/apprentissage) des modèles, ceux-ci ont été segmentés en plusieurs fichiers ne contenant qu'une seule tonalité (ou bien un seul accord) (Figure 2.3). Un exemple de segmentation d'un fichier contenant les notes d'un accordéon est présenté dans la Figure 2.4.

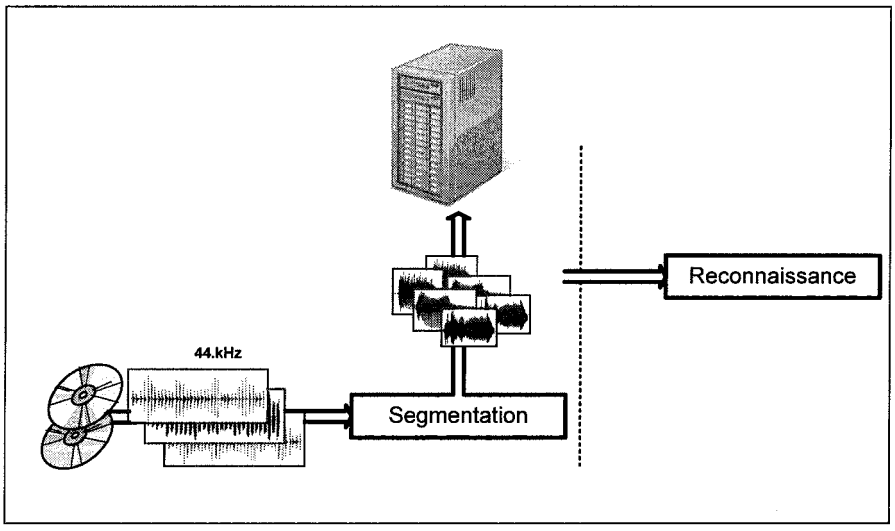

Figure 2.3 Prétraitements des fichiers de données.
Un prétraitement des fichiers de données permet de créer de nouveaux fichiers ne contenant qu'une seule tonalité chacun.

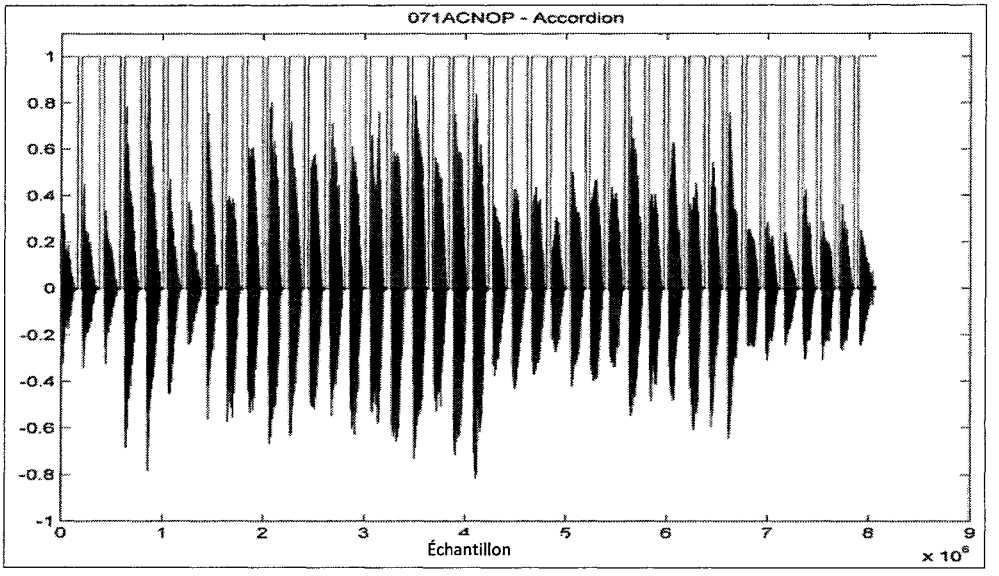

Figure 2.4 Exemple de segmentation automatique d'un fichier sonore.
Accordéon, signal échantillonné à 44,1 kHz. 41 notes ont été extraites du fichier sonore.

La segmentation automatique fut réalisée à l'aide d'un algorithme comparant l'énergie locale d'une fenêtre de 100 ms du signal à un seuil préalablement défini (voir annexe B pour le pseudo-code). Une retouche manuelle fut effectuée pour certaines segmentations automatiques ayant échouées. Par exemple, certains fichiers ont dû être segmentés manuellement puisque l'algorithme de la segmentation automatique fut incapable d'effectuer la segmentation correctement. Également, seules les articulations des instruments dites « normales », c'est-à-dire excluant les styles de jeu tels glissando (glissement continu d'une note à une autre), pizzicato ou encore vibrato (modulation en fréquence), ont été utilisées. Pour déterminer l'articulation des instruments contenue dans chaque fichier, la nomenclature spécifiée dans [10] fut utilisée. Pour pouvoir utiliser des fenêtres d'analyses assez longues, les notes trop courtes, soit plus petite que 435 ms, ont été ignorées. De même, dans les cas où la segmentation a fait défaut, les notes trop longues, soit plus grandes que 18 secondes, ont été ignorées.

Le jeu de données résultant est énuméré au Tableau 2.2, soit 29 instruments de la famille des cordes et de la famille des vents. Le nombre de notes obtenues y est indiqué pour un total de 6 698 notes. Il est important de remarquer que la distribution des notes n'est pas uniforme à tous les instruments ce qui a eu pour effet de limiter les performances de reconnaissance de certains instruments.

Tableau 2.2 Liste du nombre de sources pour chaque instrument.
Nombre de notes obtenues par la segmentation. Seules les articulations dites « normales » sont conservées, c'est-à-dire excluant les styles de jeu tels glissando, pizzicato ou vibrato.

Instrument	**Nombre de notes**	**Instrument**	**Nombre de notes**
Accordéon	282	Guitare acoustique	463
Saxophone Alto	198	Banjo	208
Saxophone Baryton	198	Basson	240
Violoncelle	377	Clarinette	240
Cornet	62	Basse électrique	676
Guitare électrique	468	Cors anglais	60
Flûte traversière	148	Cors français	218
Harmonica	168	Mandoline	283
Hautbois	132	Flûte de pan	74
Piccolo	200	Orgue	56
Flûte à bec	150	Saxophone soprano	198
Saxophone ténor	196	Trombone	194
Trompette	141	Tuba	180
Ukulélé	144	Violon alto	360
Violon	384		
total : 6 698			

2.3.2 Structure générale du système de reconnaissance

La Figure 2.5 montre la structure générale du système de reconnaissance utilisé dans les simulations de ce mémoire ainsi que les étapes encourues. En premier lieu, depuis le nom du fichier source, on prélève le nom de l'instrument selon la codification présentée par [13]. Par la

suite, l'extraction des descripteurs est effectuée créant ainsi un ensemble de vecteurs d'observation qui serviront à alimenter l'apprentissage du classificateur en aval. Une fois la liste de modèles de référence construite, le même procédé est réitéré en redirigeant cette fois la liste de vecteurs vers une fonction qui déterminera la classe (famille ou instrument) du signal sonore contenu dans le fichier. La qualité de la classification est finalement évaluée par comparaison avec le libellé originellement obtenu au début du processus.

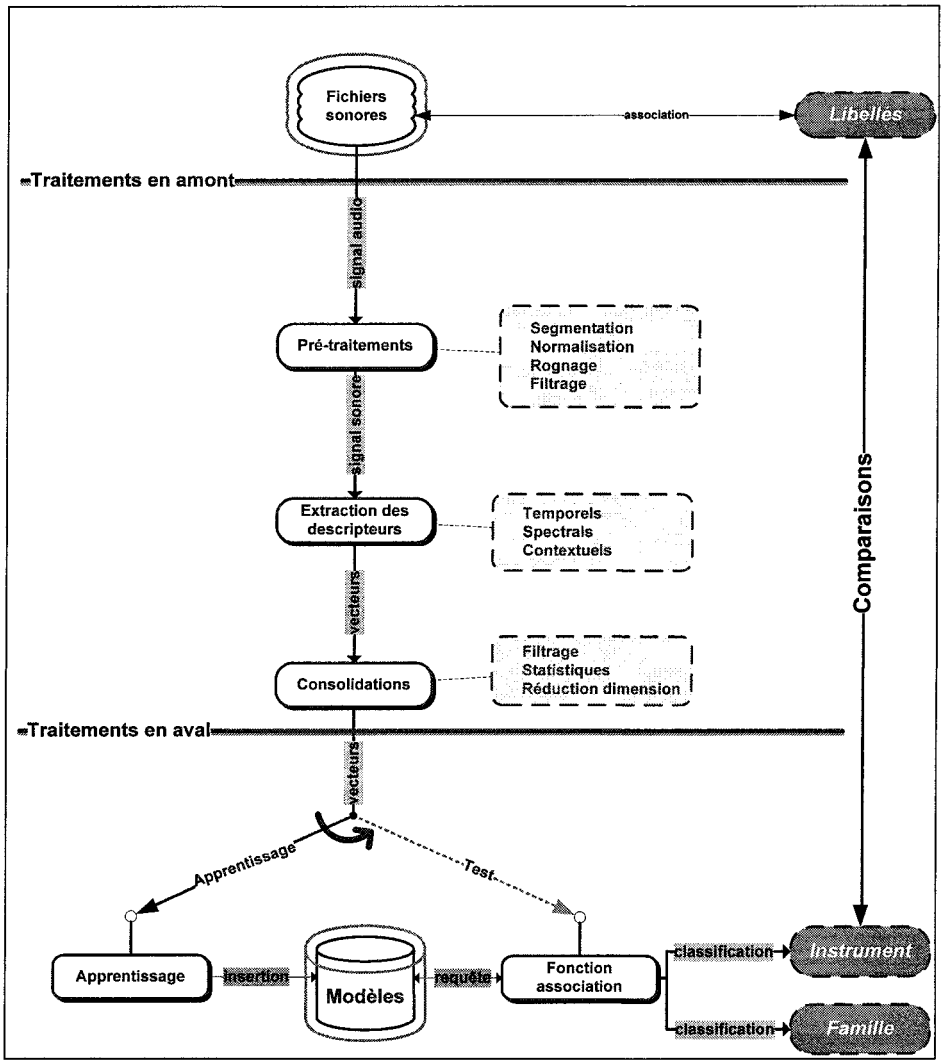

Figure 2.5 Structure de la classification des instruments de musique.

Les données sonores sont, en premier lieu, traitées de façon à les normaliser par filtrage, rognage, etc. Les descripteurs sont ensuite extraits des données normalisées et les vecteurs d'observation ainsi obtenus sont consolidés (normalisés, réduits et agrégés en une seule trame lorsque nécessaire). Un modèle numérique est construit pour chaque instrument lors de la phase d'apprentissage. Tous les modèles obtenus par apprentissage sont utilisés pour identifier l'instrument joué dans chacun des fichiers sonores lors de la phase de test.

2.3.2.1 Traitements en amont : extraction des descripteurs

Un descripteur sonore est caractérisé par son étalement temporel. Certains descripteurs sont globaux et font référence à la sonorité dans son ensemble. Par exemple l'enveloppe sonore et le temps de montée, le temps de maintien, la puissance moyenne du signal sont des descripteurs globaux. D'autres sont instantanés, plus localisés, variant dans le temps et dépendent de la fenêtre d'analyse, comme les descripteurs spectraux par exemple. Il est possible que les trames se superposent ou encore qu'elles ne soient pas uniformes mais tentent plutôt de refléter la morphologie du signal, telle la montée et le maintien de l'enveloppe sonore d'une note. Par exemple, Eronen [3] extrait les coefficients cepstraux sur l'échelle MEL (MFCC) sur deux segments de la note : à partir de l'attaque de la note et à partir du maintien de la note. Une ou plusieurs fonctions d'agrégation (moyenne, médiane, écart type, etc.) permettent, lorsque nécessaire, de condenser l'information en un seul vecteur d'observation par note.

Les paramètres peuvent être calculés directement à partir du signal temporel, comme le taux de passage par zéro (ZCR pour « Zero Crossing Rate »), ou encore calculés après une

transformation du signal comme la transformée de Fourier ou bien les Coefficients Cepstraux sur l'échelle MEL (MFCC).

2.3.2.2 Traitements en aval : classificateurs

Un classificateur peut être considéré comme un système qui, en fonction des données en entrées, fournira une décision en sortie. Dans un contexte de reconnaissance des instruments de musique, les entrées du classificateur sont des vecteurs d'observation construit à partir des descripteurs sonores ; la décision en sortie est déterminée par une catégorisation à des classes. La réalisation de cette étape s'effectue avec deux techniques : l'apprentissage supervisé et l'apprentissage non-supervisé.

L'apprentissage supervisé permet, possédant un échantillon x d'une population X et l'image de l'échantillon par une fonction inconnue *f*, d'obtenir un estimateur *h* de *f* qui pourra prédire le comportement de *f* pour des données non encore connues de la population (Figure 2.6). Plus formellement, une paire $(x, f(x)), x \in X$ est dite un *exemple* et *h* une *hypothèse* ou encore une *fonction de prédiction.* Contrairement à un problème de régression, l'ensemble des valeurs de sortie n'est pas continue mais discret; on associe à l'entrée une classe sous le format d'une étiquette.

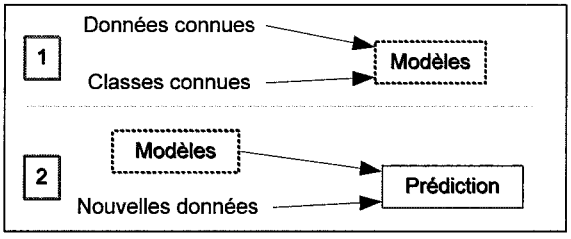

Figure 2.6 Principe de fonctionnement de l'apprentissage supervisé

L'apprentissage non-supervisé permet, entre autre, de déterminer les paramètres de la loi de probabilité décrivant les observations. L'apprentissage non-supervisé permet également de partitionner les données en groupes ou sous-groupes à l'aide d'une fonction de distance, construite sur des critères de proximité spécifiques. Par exemple, dans le contexte d'une classification hiérarchique, l'apprentissage non-supervisé permet de regrouper les instruments selon une taxonomie automatique et formelle [4], sans utiliser la taxonomie naturelle par familles d'instruments décrite dans le section 2.1. La fonction doit pouvoir minimiser les différences intra-classes et maximiser les différences interclasses pour optimiser le partitionnement. Ces deux méthodes ont la propriété de faire émerger les dissimilitudes entre groupes et les similitudes entre éléments d'un même groupe.

2.4 Travaux actuels

Les expérimentations effectuées sur la reconnaissance automatique d'instruments de musique sont abordées avec trois approches différentes : la reconnaissance avec des notes isolées, la reconnaissance de phrases musicales monophoniques et la reconnaissance avec de la musique multi-instrumentale. Chacune de ces approches propose un niveau de difficulté différent mais toutes s'appuient sur les études psycho-acoustiques et computationnelles concernant la caractérisation et la perception du timbre.

En général, les études portent sur la sélection des descripteurs en utilisant des algorithmes de réduction de la dimension par sélection et extraction des paramètres. Le Tableau 2.3 présente certains travaux effectués sur la reconnaissance des instruments de musique regroupés par bases

de données utilisées, type d'information extraite, algorithme de classification utilisé et algorithme de réduction de la dimension utilisé.

Tableau 2.3 Travaux effectués sur la reconnaissance des instruments de musique.

Auteurs	Base de données	Information musicale	Réduction de la dimension	Classificateurs
Martin [2, 14]	MUMS	Notes isolées	FMDA	*k*-NN
Eronen et Klapuri [15, 16]	MUMS	Notes isolées	SFS, SBS	*k*-NN, GMM
Agostini *et al.* [17]	MUMS	Notes isolées	Aucune	*k*-NN, QDA, CDA, SVM
Livshin et al.[18]	MUMS	Notes isolées	GDE	Gaussienne Multidimensionnelle, *k*-NN, LVQ
Kitahara et al. [19]	RWC	Notes isolées	PCA suivie de LDA	*k*-NN, Règle de décision de Bayes
Essid [4, 20]	Collection musicale personnelle sur CD	Phrases polyphoniques Notes isolées	PCA	GMM, SVM
De Poli et Prandoni [21, 22]	Inconnue	Timbre	PCA	SOM
Feiten et Günzel [23]	Inconnue	Timbre	Aucune	SOM

2.4.1 Génération d'espaces timbre

Il existe une relation entre le timbre et l'identification des instruments de musique. L'enjeu est donc de déterminer les descripteurs perceptuels multidimensionnels qui caractérisent le timbre; contrairement à la hauteur qui dépend de la fréquence fondamentale et de l'intensité qui dépend de l'intensité du niveau acoustique, le timbre dépend de paramètres acoustiques variés qui restent à définir. Du point de vue du psycho-acousticien, le timbre peut être esquissé comme une

construction géométrique perçue à travers les côtes de similarités d'un groupe d'auditeurs. La méthode *mulditimensional scaling* est généralement utilisée pour aider à trouver les descripteurs du son qui corrèlent le mieux avec les dimensions perceptuelles (brillance, finesse, compacité, etc.) [24-26]. Partant du même principe, la recherche en reconnaissance des instruments de musique débute par la construction d'un espace vectoriel décrivant le timbre, ou encore un *espace timbre*. L'idée principale est de réduire la dimension des vecteurs d'observations tout en conservant la topologie naturelle du timbre des instruments; une interprétation plus pratique devrait ainsi émerger.

De Poli et Prandoni [21, 22] utilisèrent six coefficients MFCC comme vecteurs d'entrée (fenêtres de Hamming de 32ms espacées de 4 ms chacune) à une carte auto-organisatrice de Konohen (SOM) pour construire des espaces timbre. Ils utilisèrent également la réduction de dimension à l'aide de l'analyse en composantes principales (PCA). La métrique utilisée fut la distance euclidienne. L'ensemble des données de tests fut constitué d'une note par instrument, toutes à la même hauteur. Aucun résultat sur les performances de reconnaissance des instruments ne fut cependant dévoilé. Les résultats mirent en évidence malgré tout la séparation spatiale des instruments dans l'espace timbre et le regroupement par familles des instruments de par leur proximité relative dans ce même espace.

Feiten et Günzel [23] utilisèrent également une carte SOM pour caractériser le timbre; le son des instruments fut vectorisé à partir de caractéristiques spectrales (les fréquences furent exprimées en terme de bandes critiques, en barks). Comme pour De Poli et Prandoni [21, 22], c'est la trajectoire des neurones activés dans la carte SOM sur la séquence des vecteurs d'entrées qui

détermine la forme du timbre (la Figure 2.7 montre un exemple de trajectoires). Toutefois, les trajectoires des instruments furent utilisées comme nouveaux vecteurs d'entrée dans une autre carte SOM, qu'ils nommèrent « SOM dynamique ». Les métriques utilisées furent la distance de Minkowski pour les distances spectrales et *city-block* pour les distances entre les trajectoires. L'ensemble des données de simulations fut constitué de 98 notes produites à l'aide d'un synthétiseur. Encore une fois, les performances de reconnaissance ne furent pas divulguées et les résultats dépendent nécessairement fortement des paramètres de synthèse de l'instrument.

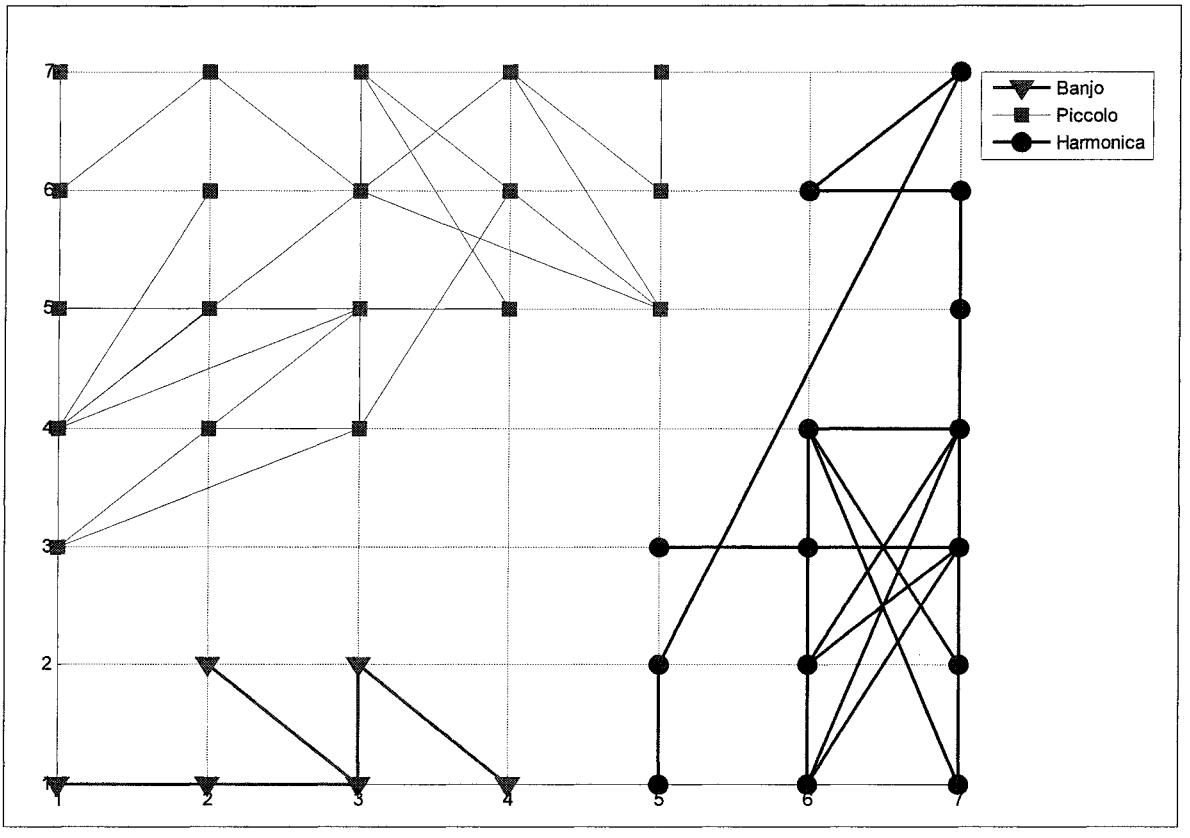

Figure 2.7 Exemple de trajectoires de neurones pour trois instruments.
Banjo (rouge), Piccolo (bleu) et Harmonica (vert).

2.4.2 Reconnaissance de notes isolées

L'utilisation de notes isolées, c'est-à-dire une seule note à la fois, présente des avantages significatifs au niveau de l'extraction des descripteurs et de la réalisation des expérimentations. Autrement dit, des descripteurs acoustiques sophistiqués sont difficilement calculables à partir d'un flux continu de notes, pouvant vraisemblablement se superposer. Également plusieurs bases de données contenant des notes isolées sont disponibles [7, 10].

Martin et Kim [14] utilisèrent 1 023 notes de 15 instruments différents contenus dans la base de donnée MUMS. Les descripteurs sonores furent calculées à partir d'un corrélogramme au lieu de l'utilisation directe de la transformée de Fourier locale, relaxant ainsi l'hypothèse sur la périodicité du signal. Au total 31 paramètres furent extraits et combinés pour former le vecteur d'observation. Pour réduire la dimension du vecteur d'observation, une analyse discriminante multiple de Fisher fut utilisée. Une classification hiérarchique par famille d'instruments et une classification non hiérarchique furent choisies et le classificateur employé fut *k*-NN. Martin et Kim [14] démontrèrent notamment que la classification hiérarchique fonctionne mieux qu'une classification directe.

Eronen et Klapuri [15, 16] reprirent la stratégie de classification hiérarchique proposée par Martin [2] avec l'objectif d'utiliser les propriétés temporelles, les propriétés de modulation, les irrégularités, la résonnance, la brillance et la synchronicité spectrale du son au lieu des descripteurs issus du corrélogramme. Un total de 1 498 notes recouvrant la plage de hauteurs de 30 instruments de la base de données MUMS fut utilisées comme données de simulations dans [15] et 5 286 notes provenant de cinq sources différentes, incluant MUMS, furent utilisées dans

[16] pour un total de 29 instruments (dont seulement 16 utilisés pour les tests de validation). Tout comme Martin [2], une stratégie de validation croisée avec des partitions de 70% de données d'apprentissage et 30% de données de tests fut utilisée. L'expérimentation démontra des résultats similaires à ceux de Martin [2] pour la stratégie de classification hiérarchique mais réussit à obtenir de meilleurs résultats avec la classification directe. La différence du nombre d'instruments entre les expérimentations rend malheureusement la comparaison difficile.

Le Tableau 2.4 montre quelques travaux effectués dans le cadre de la reconnaissance d'instruments de musique avec des notes isolées et leurs résultats respectifs. Seuls Kitahara *et al* [19] utilisèrent la même base de données que les simulations effectués dans ce mémoire. Le nombre de notes utilisé est équivalent au nombre de notes utilisé dans ce mémoire sauf que le nombre d'instruments diffère considérablement (19 pour Kitahara contre 29 pour les simulations de ce mémoire). On remarque également qu'à part Kitahara *et al.* [19], les performances sont en deçà de la barre des 80% de réussite pour une classification directe.

Tableau 2.4 Travaux effectués avec des notes isolées.

Auteur	Nombre d'instances	Nombre d'instruments	Descripteurs	Classificateurs	Taux de classification (%) (instruments)	Taux de classification (%) (familles)
Eronen [3]	5286 (MUMS, Iowa, SOL, Roland XP30, enregistrements personels)	29	MFCC, F0, SC, AM, Crest Factor, ATT	*k*-NN	35 (30)	77 (75)
Livshin et al. [18]	4381 (SOL, Iowa, MUMS, Microsoft MI, Prosonus, Vitous)	16	SC, ATT, temporal decrease, TRI, HD, SKW, KUR, SV, SS, MFCC, noisiness	*k*-NN, LDA	47-69	62-92
Eronen [27]	5895 (MUMS, Iowa, SOL, Martin, enregistrements personels)	7	MFCC, delta-MFCC + ICA	HMM	68	Inconnues
Kitahara et al. [19]	6247 (RWC)	19	SC, OER, F0 relative energy, KUR, SKW, FM, amplitude envelope slope, onset energy	Bayes (*k*-NN after PCA & LDA)	80	91
Kostek [28]	Nombre inconnu. (CMIS, MUMS)	12	Wavelet-based energy bands, MPEG-7 features	MLP	71	Inconnues
Park, Cook [29]	829 (commercial instrument sample CDs)	12	SS, SC, harmonic slope, LPC noise, harmonic expansion / contraction, spectral jitter and shimmer, spectral flux, TC, ZCR	MLP	71	88

CHAPITRE 3

DESCRIPTEURS DE SIGNAUX SONORES

L'objectif de ce chapitre est d'introduire le son et ses propriétés physiques et psycho-acoustiques sur les plans dynamique, harmonique et mélodique. Pour démontrer la pertinence des descripteurs utilisés dans les travaux de ce mémoire, les échelles de mesures psycho-acoustiques classiques sont décrites et les descripteurs psycho-acoustiques principaux utiles à l'entendement sont associés à leur phénomène physique équivalent. Certains paramètres, plus ou moins représentatifs des qualités perceptives, sont ainsi formalisés. En plus, ce chapitre décrit de façon détaillée les techniques utilisées pour extraire les paramètres numériques des descripteurs sonores des instruments de musique utilisés dans les simulations de ce mémoire. Deux types de normalisations sont également définis pour normaliser les paramètres du vecteur d'observation : la normalisation mu-sigma et la normalisation min-max.

Le choix des descripteurs est justifié par leur popularité et leurs performances dans la tâche d'identification des instruments de musique, soit : les coefficients cepstraux sur l'échelle MEL (MFCC), les coefficients de prédiction linéaire (LPC), les statistiques spectrales et d'enveloppe et le taux de passages par zéro. Un total de 38 paramètres classiques a été ainsi sélectionné. Inspiré du domaine de traitement d'image, une nouvelle représentation du timbre est proposée, le chromatimbre, et est paramétrée à l'aide de 7 moments invariants d'image.

3.1 Caractéristiques des sons instrumentaux

Les techniques et les échelles de mesure des grandeurs de l'acoustique physique sont bien définies. Par exemple la fréquence est mesurée en Hertz, l'intensité en Décibels, les pressions acoustique en Pascals, la durée en secondes, la longueur d'onde en mètres. La mesure d'une sensation est en contrepartie moins aisée, la définition d'une échelle de valeurs n'est pas donnée d'emblée. On attribue quatre caractéristiques principales aux notes de musique : durée, hauteur, intensité et timbre qui sont associées aux grandeurs physiques du temps, de la fréquence et de l'amplitude respectivement, le timbre étant un descripteur purement psychologique. Tandis que les trois premières caractéristiques ont une définition tangible, aucun consensus n'établie de façon clair quels sont les phénomènes auditifs que devraient inclure une définition du timbre. L'*American National Standard Institute* (ANSI) définit le timbre comme étant « l'attribut de la sensation auditive avec lequel un auditeur peut juger que deux sons ayant la même hauteur et la même intensité sont dissimilaires » [30]. En d'autres termes, le timbre peut admettre tout ensemble d'attributs acoustiques autre que la hauteur et l'intensité. Cette définition ne nous dit malheureusement pas ce que le timbre est mais plutôt ce que le timbre n'est pas.

On peut néanmoins schématiser une phrase musicale à l'intérieur d'un volume (Figure 3.1) où les trois dimensions proviennent des caractéristiques d'une note de musique précédemment citées (hormis le timbre). Sur le plan dynamique, on retrouve l'enveloppe sonore qu'est l'évolution de l'intensité en fonction du temps. Sur le plan mélodique, on retrouve la partition qu'est l'évolution des hauteurs des notes et accords en fonction du temps. Sur le plan harmonique, on

retrouve le spectre qu'est l'évolution de l'intensité en fonction de la fréquence. On distingue ainsi certains attributs de l'instrument de musique associé à leurs caractéristiques acoustiques [31] :

- L'échelle musicale (instrument tempéré vs instrument naturel)
- La dynamique (associée à l'intensité)
- Le timbre
- L'enveloppe sonore
- La radiation acoustique de l'instrument

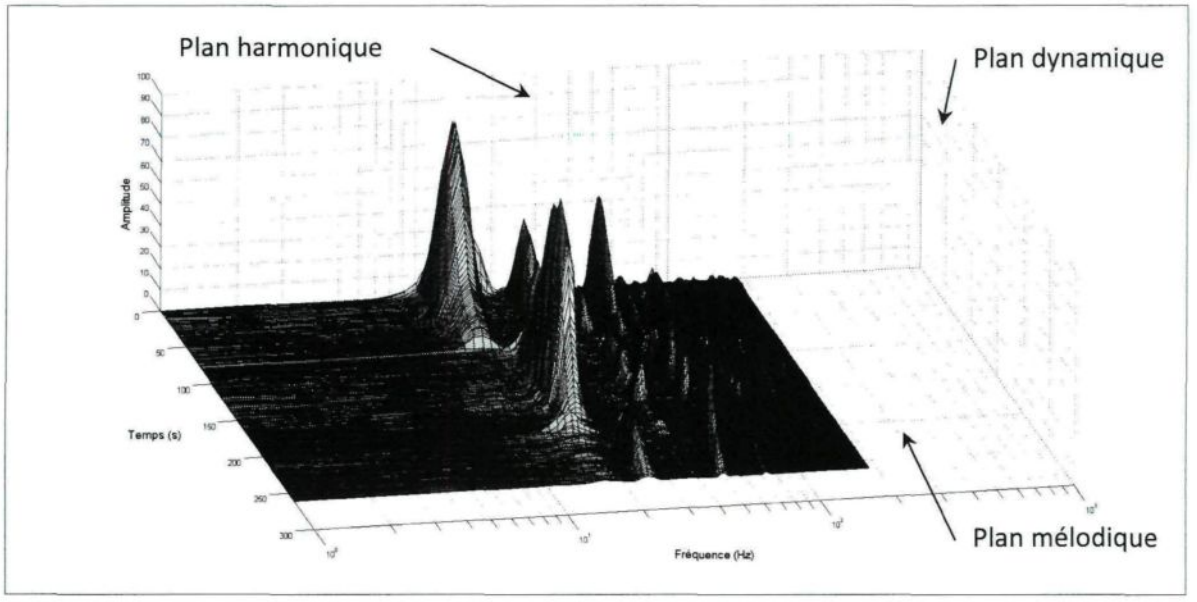

Figure 3.1 Spectrogramme d'une phrase musicale représenté en trois dimensions. Le spectrogramme contient trois notes de violoncelle jouées successivement; on y aperçoit les trois plans musicaux que sont les plans dynamique, mélodique et harmonique.

3.1.1 L'enveloppe sonore

L'enveloppe sonore de l'instrument de musique, observée sur le plan dynamique, est représentée en quatre phases distinctes de l'évolution de son amplitude :

1. la transitoire d'attaque, est le segment ascendant de la note;
2. le déclin, est le segment entre l'attaque et le maintien;
3. le maintien, est le segment constant de la note;
4. la chute, est le segment de relâche de la note.

Ces phases sont représentées à la Figure 3.2. On y fera référence comme étant l'enveloppe « ADSR » (*Attack, Decay, Sustain, Release*). L'enveloppe ADSR d'une note de guitare classique est présentée dans la Figure 3.3.

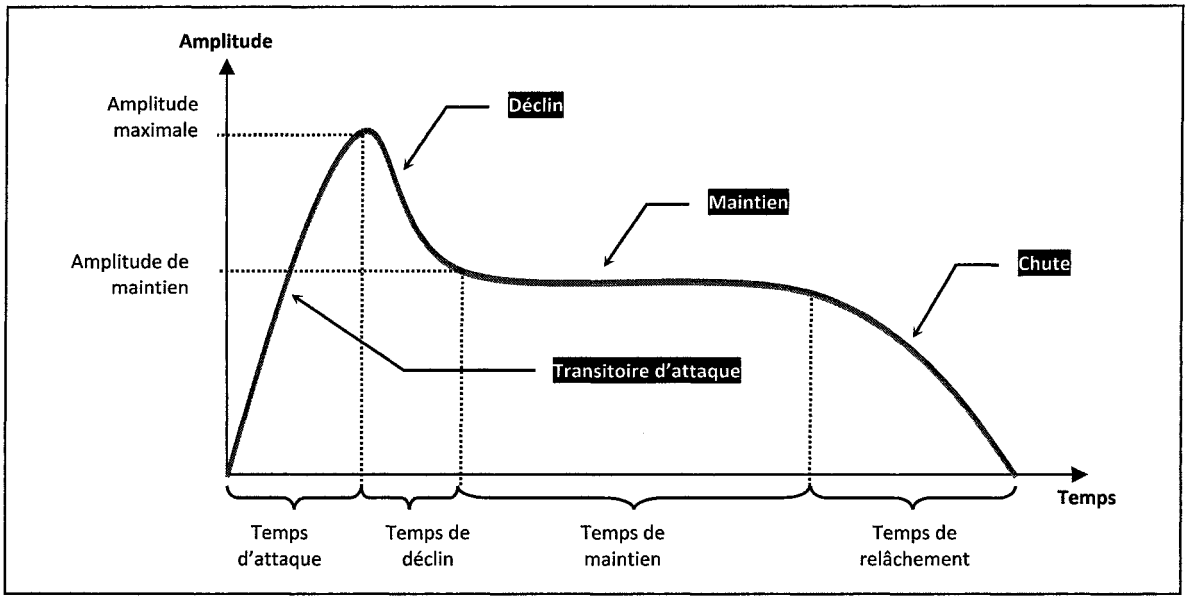

Figure 3.2 Détails de l'enveloppe ADSR (Attack, Decay, Sustain, Release).
On peut y décerner les quatre phases d'amplitude du signal : la transitoire d'attaque est le temps requis à la note pour atteindre son amplitude maximale, le déclin est le temps requis à la note pour atteindre son régime permanent, le maintien est le temps pendant lequel l'amplitude conserve un plateau permanent et la chute est le segment de relâchement ou d'extinction de la note.

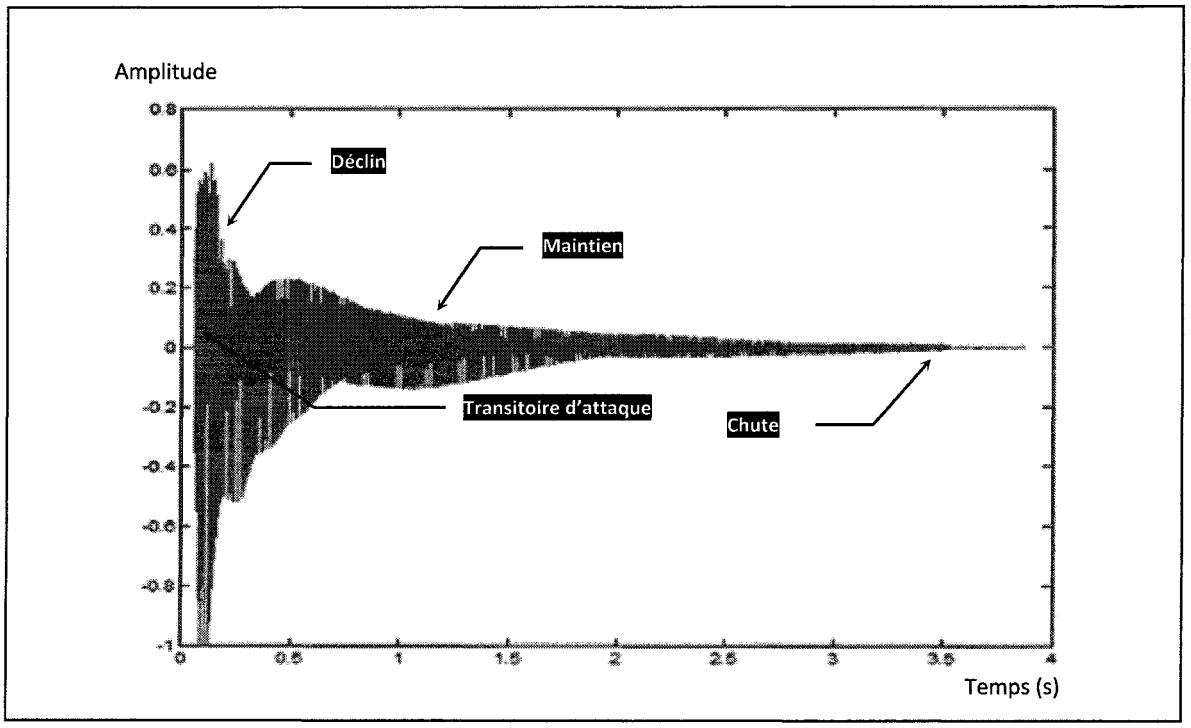

Figure 3.3 Enveloppe ADSR (Attack, Decay, Sustain, Release) d'une note de guitare classique.

3.1.2 La hauteur

La hauteur d'une note de musique, observée sur le plan mélodique et sur le plan harmonique, est définie par l'*American National Standards Institute* (ANSI) comme étant « l'attribut d'une sensation auditive pour lequel un son peut être ordonné sur une échelle de grave à aigu » [30].

Ces fréquences sont tantôt harmoniques (fréquences de multiples entiers de la fréquence fondamentale), tantôt inharmoniques. L'harmonicité (le rapport de fréquences qui ne sont pas à des fréquences multiples entières du son fondamental) d'un instrument de musique tend à imposer la qualité de la perception de la hauteur de la note jouée.

L'oreille humaine ne perçoit pas le changement d'octave comme un doublement de la fréquence au delà de 500 Hz; l'échelle de MEL, exprimée en MELS (dérivée de *melody*) a été proposée par Stevens, Volkman et Newman en 1937 à l'aide de données empiriques provenant de sujets humains. L'alignement de l'échelle de MEL par rapport aux fréquences en hertz est arbitrairement positionné pour que 1000 MELS équivaillent à 1000 Hz. Ainsi une hauteur à 1000 MELS sera « perçue » au double de cette hauteur à 2000 MELS et à la moitié de cette hauteur à 500 MELS. Les résultats ont conduit à la formule de conversion des fréquences en hertz vers une hauteur en MELS donnée dans l'équation (3.1). Sa réciproque est formulée dans l'équation (3.2). La courbe de relation entre les fréquences en hertz et la hauteur en MELS est présentée graphiquement à la Figure 3.4.

$$mel = 2595 \cdot log_{10}\left(1 + \frac{f}{700}\right) \tag{3.1}$$

$$f = 700 \cdot \left(10^{mel/2595} - 1\right) \tag{3.2}$$

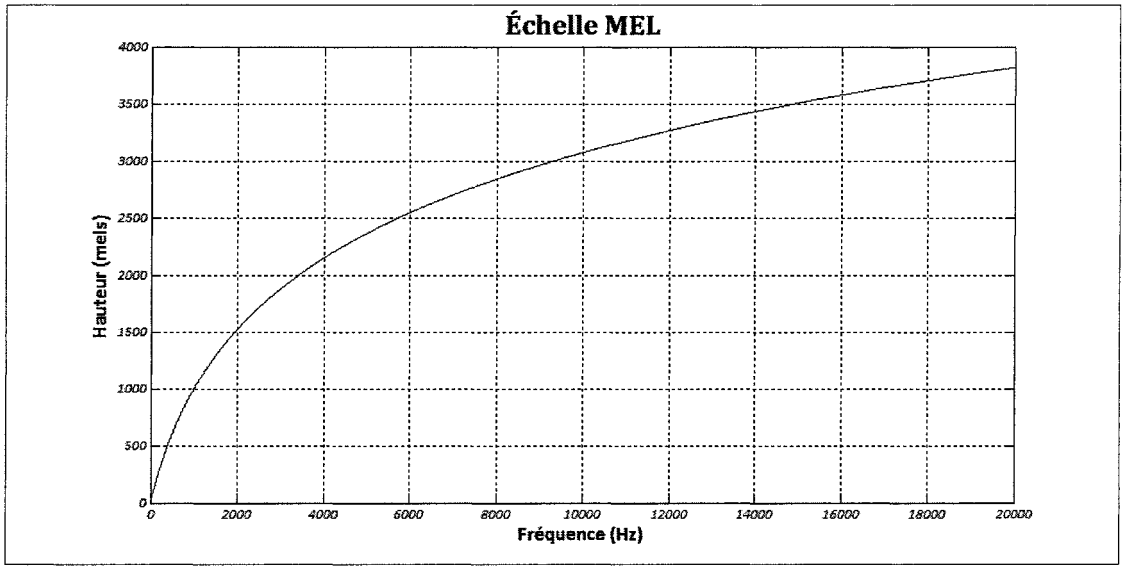

Figure 3.4 Courbe de l'échelle MEL.
Transformation des fréquences en Hertz vers des fréquences en MELS selon l'équation (3.1).

3.1.3 L'intensité

L'intensité d'une note de musique est définie par l'*American National Standards Institute* (ANSI) comme étant « l'attribut d'une sensation auditive pour lequel un son peut être ordonné sur une échelle de faible à fort » [30]. Le musicien utilise cette caractéristique pour exprimer la dynamique du son en termes de nuances (en comparaison à la mélodie liée aux intervalles de hauteur). En termes scientifiques on la compare à l'intensité acoustique mesurée en watts par mètres carrés (W/m^2) et au niveau d'intensité acoustique en décibels (dB) (en anglais on utilise l'acronyme « SPL » pour *Sound Pressure Level*). En réalité, c'est une quantité subjective, moins complexe que la hauteur néanmoins non triviale.

L'échelle d'évaluation des intensités a comme valeur unitaire le *Sone*, proposé par Stanley Smith Stevens en 1936. Cette unité caractérise l'isosonie qu'est l'équivalence en perception d'intensité de deux sons de fréquences différentes. Par convention, on accorde la valeur de un sone comme l'intensité perçue d'un son de 1 kHz d'une intensité acoustique de 40 dB SPL en champ libre. Ainsi un son de 2 sones sera perçu comme ayant une intensité double d'un son de 1 sone.

3.2 Représentation paramétrique

Les qualités perceptives d'un son instrumental ont un dual physique tel la hauteur par rapport à la fréquence et l'intensité par rapport à l'intensité acoustique. La relation entre la représentation paramétrique d'une grandeur physique et sa description perceptive n'est pas toujours connue. Certaines de ces relations sont néanmoins connues : le barycentre spectral est associé à la

brillance, les fluctuations périodiques de la hauteur sont associées au *vibrato,* les fluctuations périodiques de l'intensité sont associées au *tremolo* et la nasalité est associée à la position des *formants,* sont des exemples bien connus (Tableau 3.1). Plusieurs études psycho-acoustiques ont proposé des descripteurs pour la définition du timbre et pour la reconnaissance des instruments de musique [22, 24-26, 32].

Tableau 3.1 Description perceptive du son en lien avec certains paramètres physiques. Certaines relations entre la représentation paramétrique d'une grandeur physique et sa description perceptive sont bien connues.

Description perceptive	**Grandeur physique**
Hauteur	Fréquence
Intensité	Intensité acoustique
Dynamique	Enveloppe
Brillance	Barycentre spectral
Tonique par rapport au bruit	Platitude spectrale
Vibrato	Fluctuations périodiques de la hauteur
Tremolo	Fluctuations périodiques de l'intensité
Synchronicité	Délai entre hautes fréquences et la fréquence fondamentale pendant l'attaque

3.2.1 Coefficients cepstraux sur l'échelle MEL

La littérature actuelle présente les caractéristiques spectrales comme étant celles qui donnent les meilleurs résultats, en particulier les coefficients cepstraux sur l'échelle de MEL (MFCC), comme en témoignent [33] et [3]. Le cepstre est défini comme étant la transformée de Fourier inverse appliquée au logarithme naturel de la transformée de Fourier du signal [34]. Ainsi, plus formellement :

$$C(x(t)) = \mathcal{F}^{-1}\{ln|\mathcal{F}\{x(t)\}|\} \quad \textbf{(3.3)}$$

est l'équation du cepstre réel d'un signal $x(t)$. Pour rappeler le fait que l'on effectue une transformation inverse à partir du domaine fréquentiel, les dénominations dans le domaine cepstral sont des anagrammes de celles utilisées en fréquentiel. Ainsi le spectre devient le **cepstre**, la fréquence une **quéfrence**, un filtrage un **liftrage** et la phase une **saphe**.

Le cepstre peut être considéré comme étant la mesure du taux de changement dans l'enveloppe spectrale. En traitement homomorphique du signal, en particulier le filtrage homomorphique, le cepstre est utilisé pour séparer la combinaison du conduit vocal (considéré comme filtre) de la source d'excitation du signal sonore dans le modèle de production source-filtre (la convolution des deux systèmes est simplifiée par une simple addition dans le domaine log-spectral). Également, le cepstre permet d'estimer la fréquence fondamentale (la hauteur d'une note) d'un signal, comme en démontre la Figure 3.5. Enfin, dans cette même figure, on remarque que le cepstre est utilisé pour obtenir une estimation du spectre lissé du signal.

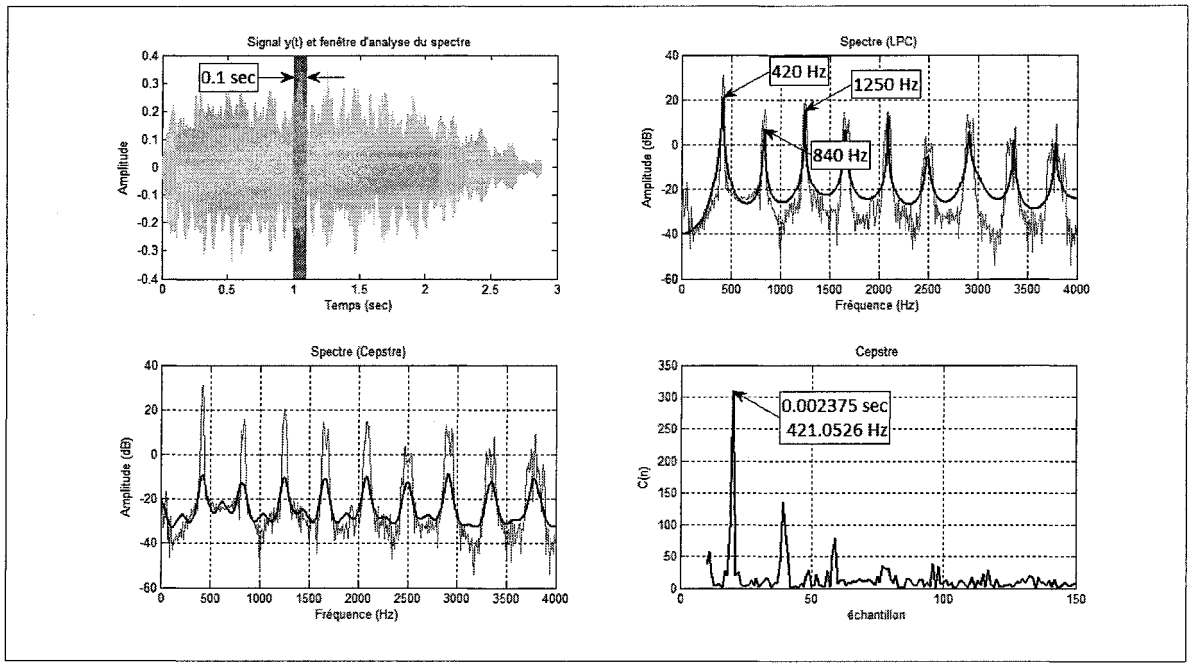

Figure 3.5 Exemple de l'analyse cepstrale d'une note d'accordéon.
Signal de la note d'accordéon et fenêtre d'analyse (100ms) (en haut à gauche).
Spectre de la transformée de Fourier discrète (TFD) (en gris pâle en bas et en haut à droite).
Spectre lissé obtenu par coefficients de prédiction linéaire (LPC) (en haut à droite).
Spectre lissé et obtenu par utilisation du cepstre (en bas à gauche).
Détermination de la hauteur de la note (en bas à droite).

Les coefficients cepstraux sur l'échelle de fréquences de MEL (*Mel Frequency Cepstral Coefficients* ou *MFCC*) est une représentation du signal acoustique introduite en 1980 par Davis et Mermelstein [35] dans une étude portant sur les techniques d'encodages de mots monosyllabiques. Cette représentation provient du calcul du cepstre sur l'échelle des fréquences de MEL et est devenue une des techniques les plus populaires dans les systèmes de reconnaissance de la parole. De Poli et al. [21] ont montré qu'il est également possible d'utiliser les coefficients MFCC dans la reconnaissance des instruments de musique.

Le calcul du cepstre en appliquant directement la formule de l'équation (3.3) est couteux puisque deux transformées de Fourier rapide (FFT pour *Fast Fourier Transform*) doivent être utilisées. Une méthode conventionnelle pour l'obtention des coefficients cepstraux est l'utilisation des coefficients de prédiction linéaire par la méthode d'autocorrélation. Une autre technique est basée sur l'utilisation d'un banc de filtres distribué sur l'échelle MEL.

Les coefficients c_i sont calculés en utilisant la transformée en cosinus discrète :

$$c_i = \sum_{k=1}^{N} X_k cos\left[i\left(k - \frac{1}{2}\right)\frac{\pi}{N}\right] \qquad \textbf{(3.4)}$$

pour M coefficients MFCC où $i = 1,2,\ldots,M$, et $X_1, X_2, \ldots, X_N$ est l'énergie à la sortie d'un banc de N filtres distribué sur l'échelle MEL. Les filtres sont triangulaires et ont un facteur de qualité Q constant.

L'algorithme est présenté dans la Figure 3.6 et se réduit ainsi :

1. Calculer la transformée de Fourier discrète du signal (ou de la trame);
2. Transposer le spectre ainsi obtenu sur l'échelle MEL, c'est-à-dire pondérer le spectre d'amplitude par un banc de filtres triangulaires espacés selon l'échelle MEL;
3. Compresser la dynamique sonore en utilisant le logarithme à la sortie des filtres;
4. Calculer la transformée de cosinus discrète en considérant les logarithmes comme étant un signal.

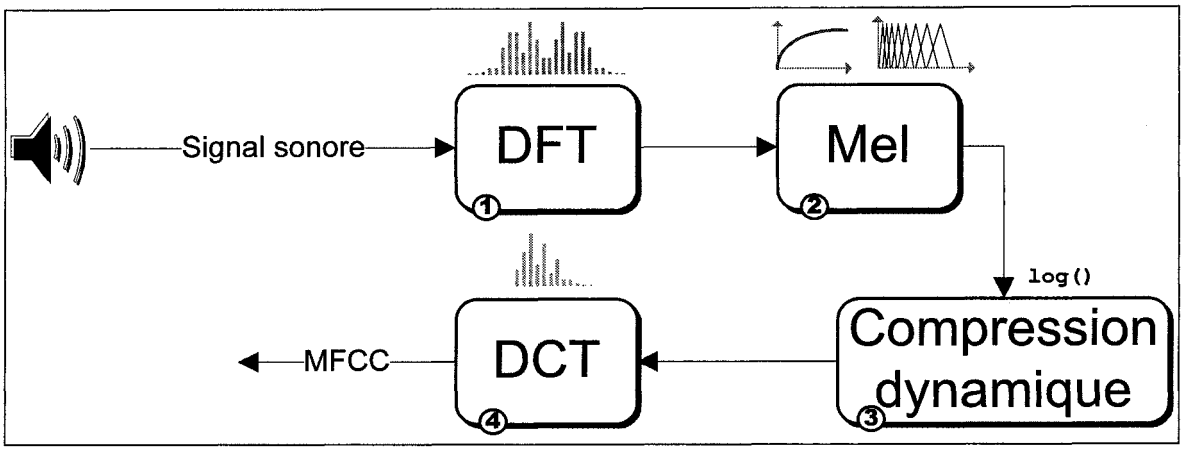

Figure 3.6 Diagramme bloc de l'extraction des coefficients MFCC.

Le coefficient C_0 détermine l'énergie moyenne du banc de filtres ($C_0 = \sum_{k=1}^{N} X_k$) et est normalement rejeté puisqu'il est fonction du gain de chacun des filtres. Rejeter le coefficient C_0 permet en outre d'obtenir efficacement une représentation normalisée du signal. L'efficacité des coefficients MFCC est principalement due à la distribution des filtres sur l'échelle MEL et à la compression dynamique des sorties de ces filtres, mimiques simplifiées des étages inférieurs du système auditif (la tonotopie de la cochlée entre autre).

3.2.2 Descripteurs spectraux

3.2.2.1 Moments spectraux

Les moments spectraux permettent de caractériser les statistiques spectrales d'un signal. Ils sont fréquemment utilisés dans la reconnaissance des instruments de musique [3, 4, 15, 16, 20]. Un moment mathématique est une mesure quantitative décrivant la forme des données aléatoires. On définit le moment d'ordre *n* d'une variable aléatoire X par la quantité $\mu_n = \mathbb{E}[X^n]$. Cette quantité est définie par l'équation (3.5), pour les variables continues

$$\mu_n = \int_I x^n f(x)\, dx \tag{3.5}$$

et par l'équation (3.6) pour les variables discrètes :

$$\mu_n = \sum x_k^n p_k \tag{3.6}$$

L'équation (3.5) est dite le $n^{ième}$ moment d'une fonction f continue sur un intervalle I et son équivalent (3.6) pour une variable aléatoire discrète avec probabilités p_k.

On pose le spectre S comme étant un histogramme. La probabilité d'obtenir la fréquence f_k sur K composantes fréquentielles est alors donnée par son amplitude $a_k = |S(f_k)|$:

$$p_k = \frac{a_k}{\sum_{i=0}^{K-1} a_i} \tag{3.7}$$

et donc le $n^{ième}$ moment spectral est donné par l'équation :

$$\mu_n = \frac{\sum_{k=0}^{K-1} f_k^n\, a_k}{\sum_{k=0}^{K-1} a_k} \tag{3.8}$$

pour K composantes fréquentielles normalisées, f_k et a_k sont la fréquence et l'amplitude de la $k^{ième}$ composante respectivement, c'est-à-dire que la $k^{ième}$ composante fréquentielle normalisée est donnée par $f_k = \frac{k}{N}, k = 1 \dots K$ et est d'amplitude a_k calculée avec une TFD sur $2N$ échantillons. La composante fréquentielle non-normalisée dépend de la fréquence d'échantillonnage $f_{éch.}$ et est donnée par $f'_k = \frac{f_{éch.} k}{N}, k = 1 \dots K$.

Barycentre spectral

Le barycentre spectral (SC pour *spectral centroid*) est une mesure caractérisant le « centre de gravité » (Figure 3.7) du spectre sonore du signal sous analyse, donnant ainsi la fréquence centrale. Le barycentre spectral est l'équivalent d'une moyenne pondérée des fréquences présentes dans le signal. On le calcule à l'aide du moment spectral d'ordre 1 :

$$SC = \mu_1 = \frac{\sum_{k=0}^{K-1} f_k \, a_k}{\sum_{k=0}^{K-1} a_k} \qquad \textbf{(3.9)}$$

D'un point de vue perceptif le barycentre spectral permet de quantifier la « brillance » du son; la sensation de netteté est accrue d'autant que la valeur du barycentre est élevée.

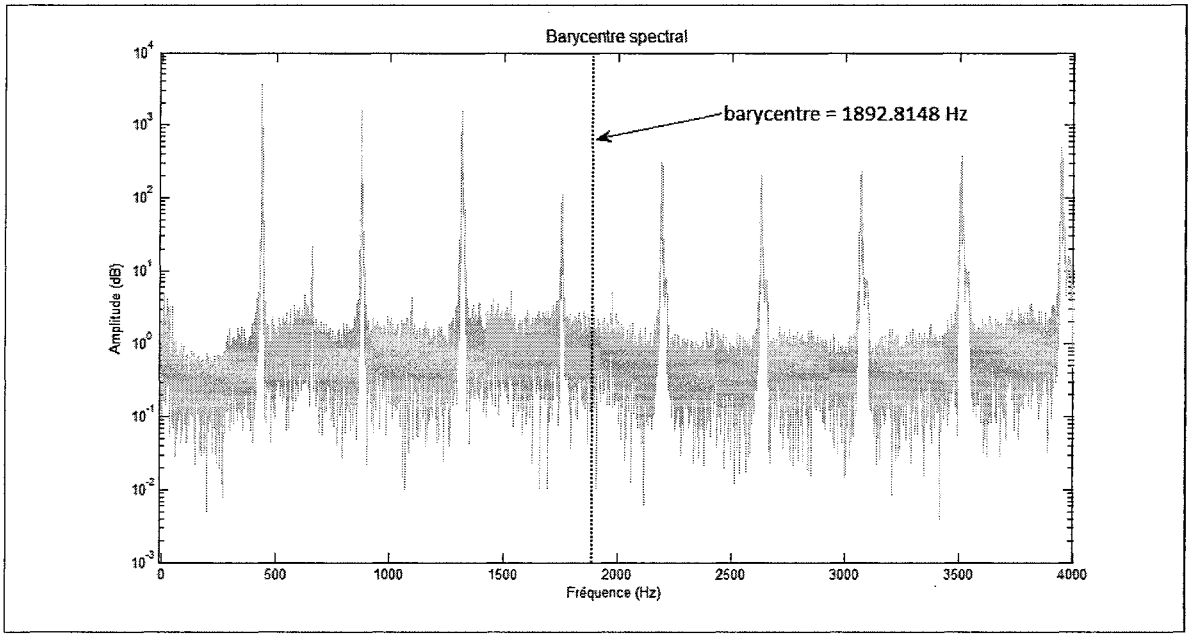

Figure 3.7 Barycentre spectral d'une note d'harmonica.
Fréquence d'échantillonnage de 8 kHz.

Largeur spectrale

La largeur spectrale (SW pour *spectral width*) décrit l'étendue du spectre autour de sa moyenne $\mu = \mathbb{E}[X]$ et est calculée à partir de la variance $\sigma^2 = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$. C'est l'équivalent de calculer le 2ième moment de la variable centrée $X \mapsto X - \mu$:

$$SW = \sqrt{\mu_2 - \mu_1^2} \tag{3.10}$$

avec les moments μ_2 et μ_1 obtenus de l'équation (3.8).

Asymétrie spectrale

L'asymétrie spectrale (SS pour *spectral skewness*) permet de représenter la symétrie du spectre autour de sa moyenne; c'est le 3ième moment de la variable centrée-réduite $X \mapsto \frac{X-\mu}{\sigma}$:

$$SS = \frac{2\mu_1^3 - 3\mu_1\mu_2 + \mu_3}{SW^3} \tag{3.11}$$

avec les moments μ_3, μ_2 et μ_1 obtenus de l'équation (3.8) et *SW* obtenu de l'équation (3.10).

Platitude spectrale

La platitude spectrale (SK pour *spectral kurtosis*) correspond à une mesure de l'aplatissement du spectre autour de sa moyenne (Figure 3.8) ; c'est le 4ième moment de la variable centrée-réduite $X \mapsto \frac{X-\mu}{\sigma}$:

$$SK = \frac{-3\mu_1^4 + 6\mu_1\mu_2 - 4\mu_1\mu_3 + \mu_4}{SW^4} - 3 \tag{3.12}$$

avec les moments μ_4, μ_3, μ_2 et μ_1 obtenus de l'équation (3.8) et *SW* obtenu de l'équation (3.10).

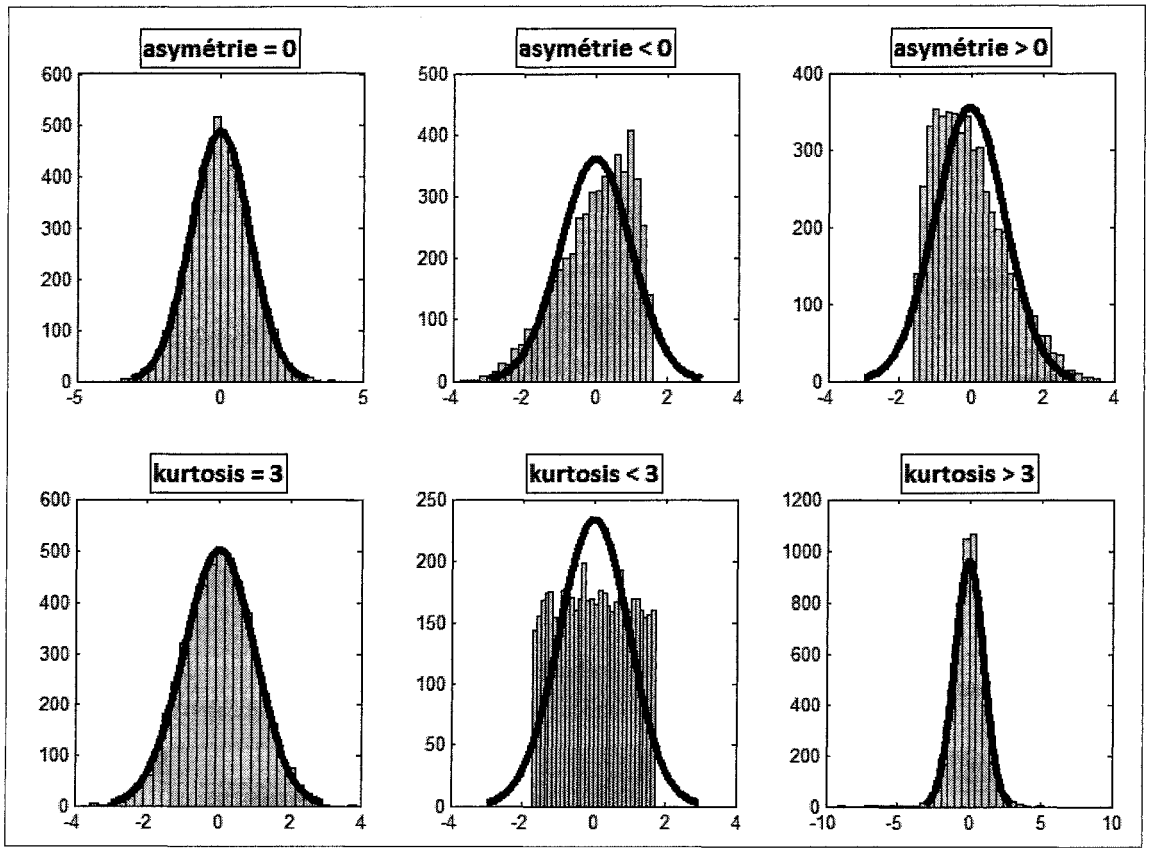

Figure 3.8 Effets de l'asymétrie et de la platitude sur la forme des données.

3.2.2.2 Coefficients de prédiction linéaire

La prédiction linéaire est un outil permettant d'obtenir une approximation « lisse » du spectre du signal; les coefficients de prédiction servent alors à décrire l'enveloppe spectrale du signal sonore. Les coefficients de prédiction linéaire (LPC pour *Linear Prediction Coefficients*) sont utilisés en particulier pour le traitement de la parole mais également pour la reconnaissance d'instruments de musique.

Un système linéaire invariant dans le temps (LTI pour *Linear Time-Invariant*) est décrit par la combinaison linéaire d'un signal d'entrée *x* et du signal de sortie *y* décalés (Figure 3.9), c'est-à-dire un modèle autorégressif et moyenne mobile (ARMA pour *Autoregressive Moving-Average*) :

$$y(n) - \sum_{k=1}^{p} a_k\, y(n-k) = \sum_{j=0}^{q} b_j\, x(n-j) \qquad \textbf{(3.13)}$$

La transformée en *z* du système de l'équation (3.13) nous donne la fonction de transfert du système :

$$H(z) = \frac{Y(z)}{X(z)} = \frac{\sum_{j=0}^{q} b_j\, z^{-j}}{1 - \sum_{k=1}^{p} a_k\, z^{-k}} \qquad \textbf{(3.14)}$$

En traitement de la parole, le signal *x(n)* peut représenter la source dans le modèle source-filtre et est alors modélisé par un train d'impulsions unitaire $G \cdot u(n)$ avec un gain constant *G* [36]. On tente d'obtenir les paramètres du filtre $H(z)$ décrivant le canal vocal comme éléments représentatifs du locuteur à l'aide d'un système tout-pôles. Dans ce cas particulier, le système LTI (*Linear Time-Invariant*) :

$$y(n) - \sum_{k=1}^{p} a_k\, y(n-k) = Gu(n) \qquad \textbf{(3.15)}$$

est purement autorégressif et la fonction de transfert du filtre est donnée par :

$$H(z) = \frac{Y(z)}{GU(z)} = \frac{1}{1 - \sum_{k=1}^{p} a_k\, z^{-k}} = \frac{1}{A(z)} \qquad \textbf{(3.16)}$$

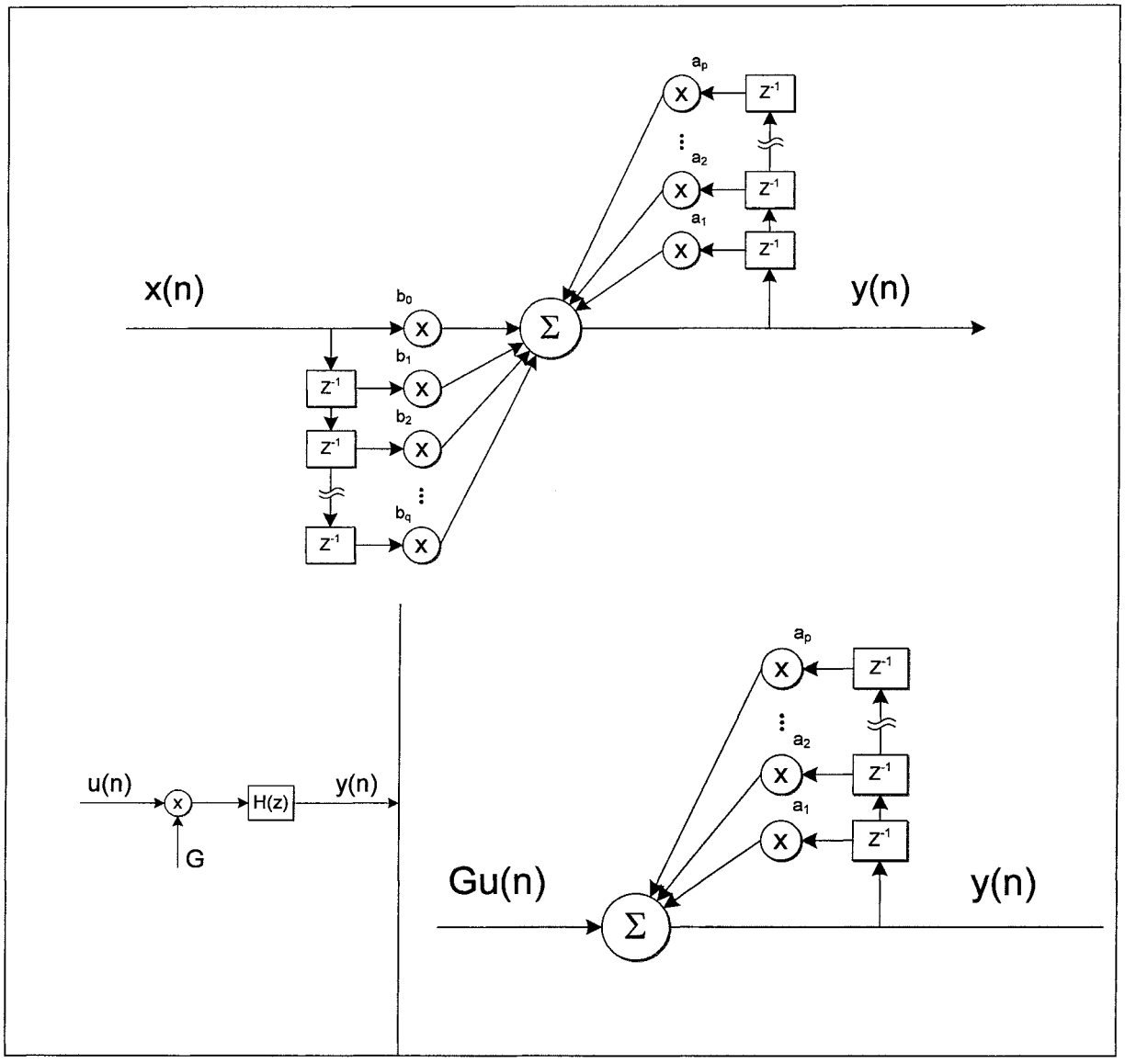

Figure 3.9 Schémas d'un système LTI.
Système LTI composé d'un modèle ARMA (autorégressif et moyenne mobile) (en haut); représentation du modèle source-filtre de la parole (à gauche); système LTI obtenu du modèle source-filtre (à droite).

Pour déterminer les coefficients $\{a_k\}$ du filtre de l'équation (3.16), on définit l'opération de prédiction d'ordre *p* qui consiste à estimer les valeurs du signal au temps *n* à partir de *p* valeurs passées [36-38] :

$$\hat{x}(n) = \sum_{k=1}^{p} a_k \, x(n-k) \tag{3.17}$$

où $\hat{x}(n)$ est la valeur du signal à prédire au temps *n*. On pose alors l'erreur de prédiction par :

$$\begin{aligned} e(n) &= x(n) - \hat{x}(n) \\ &= x(n) - \sum_{k=1}^{p} a_k \, x(n-k) \end{aligned} \tag{3.18}$$

avec l'objectif de minimiser l'erreur quadratique (MSE pour *Mean Squared Error*) $e_{MSE} = \mathbb{E}[e^2]$. Par le principe d'orthogonalité on sait que le minimum de l'erreur quadratique est atteint lorsque l'erreur est orthogonale au signal [37]. Ce système se résout par les méthodes de covariance ou d'autocorrélation [38]. Cette dernière est donnée par *p* équations linéaires :

$$\sum_{k=1}^{p} R(|i-k|) \, a_k = R(i), \qquad 1 \leq i \leq p \tag{3.19}$$

tel que $R(k)$ est le $k^{ième}$ coefficient d'autocorrélation d'une fenêtre de *N* échantillons défini par :

$$R(k) = \sum_{n=0}^{N-1-k} x(n) \, x(n+k) \tag{3.20}$$

L'ensemble des *p* équations à *p* inconnues de l'équation (3.19) se nomme les équations de Yule-Walker et se réécrit sous forme matricielle :

$$\mathbf{a} = \mathbf{R}^{-1}\mathbf{p} \tag{3.21}$$

Puisque **R** est une matrice de Toeplitz, la récursion de Levinson-Durbin permet de résoudre l'équation (3.21) efficacement (en $O(n^2)$ au lieu de $O(n^3)$ avec la méthode Gauss-Jordan).

3.2.3 Descripteurs temporels

3.2.3.1 Taux de passage par zéro

Le taux de passage par zéro (*ZCR* pour *zero crossing rate*) est une mesure de la fréquence de passage de l'amplitude par l'origine. Elle tend à révéler les signaux bruités (taux élevé) par rapport aux signaux périodiques (taux faible) [4] et est une mesure approximative de la fréquence fondamentale. Le ZCR est notamment utilisé en traitement de la parole pour reconnaitre les sons voisés et les sons non-voisés.

$$zcr = \frac{1}{T-1}\sum_{t=1}^{T} \varphi(x_t x_{t-1} < 0) \tag{3.22}$$

avec $\varphi(A) = \begin{cases} 1, & si\ la\ proposition\ A\ est\ vraie \\ 0, & sinon \end{cases}$

3.2.4 Paramètres proposés : moments invariants du chromatimbre

3.2.4.1 Introduction

Le spectrogramme du signal contient de l'information sur une partie du timbre de l'instrument tant sur le plan spectral que temporel. En effet, par définition, le spectrogramme est la représentation du spectre ponctuel du signal calculé sur des fenêtres successives. Par extension, le chromagramme contient également cette information avec l'avantage de condenser le spectre sur un nombre limité de composantes fréquentielles, nommée « bins ». De plus, le chromagramme est une représentation importante pour les applications MIR (*Music Information Retreival*) notamment comme descripteur dans la reconnaissance de clé musicale [39, 40].

Une nouvelle représentation du timbre à partir du chromagramme est proposée, le *chromatimbre*, permettant ainsi de construire une empreinte d'instrument de musique. Cette nouvelle empreinte transpose conséquemment un problème de traitements de signal vers un problème de traitements d'image. Les *moments invariants d'image* furent de ce fait choisis comme descripteur du chromatimbre dans une perspective d'identification d'images.

3.2.4.2 Chromagramme

Le chromagramme est obtenu en distribuant les composantes fréquentielles du spectrogramme parmi plusieurs classes de hauteurs; il décrit donc l'évolution temporelle des hauteurs musicales. Ces classes ou « bins » sont inscrites à l'intérieur d'une échelle nommée *échelle musicale*, plus particulièrement sur l'*échelle chromatique*. Autrement dit, le

chromagramme représente l'intensité de chacune des 12 notes de la gamme chromatique présente dans le signal sonore à un instant donné.

L'échelle chromatique est utilisée dans la musique tonale depuis la Renaissance. Elle est séparée en 12 degrés, soit les sept degrés de l'échelle diatonique plus cinq notes intermédiaires. Ces nouvelles notes sont obtenues par altérations et divisent chacun des cinq tons de l'échelle diatonique en deux demi-tons (pas nécessairement identiques). Cependant, dans le tempérament égal, l'octave est divisée en 12 demi-tons rigoureusement égaux. Les bins de l'échelle chromatique tempérée sont alors :

$$f_k = \left(2^{1/B}\right)^k f_{min} \tag{3.23}$$

où f_k varie de f_{min} à un fréquence plus haute f_{max}, B est le nombre de degrés, 12 en l'occurrence. Dans les faits, on pose le diapason à LA_4 = 440 Hz ce qui positionne le DO_4 à $f_{min} = 261{,}63\ Hz$.

Plusieurs variations d'algorithmes sont disponibles pour extraire le chromagramme d'un son mais suivent généralement l'algorithme suivant [40]:

1. On calcule la transformée de fourrier discrète $X(k)$
2. On calcule la transformée *constant-Q*, $X_{CQ}(k)$, à partir de $X(k)$, les composantes étant distribuées selon l'échelle chromatique
3. On obtient un vecteur décrivant les bins de chacun des B degrés :

$$CH(b) = \sum_{m=0}^{M-1} \left| X_{CQ}(b + mB) \right| \tag{3.24}$$

où $b = 1,2,\ldots,B$ est l'index des bins du chromagramme, M est le nombre d'octaves de la transformée *constant-Q*.

3.2.4.3 Chromatimbre

Par construction, le chromagramme est représentable à l'aide d'une image, c'est-à-dire qu'à une intensité donnée est associée une couleur. On obtient ainsi une image bitmap dont l'abscisse représente le temps et l'ordonnée le chroma. Le chromagramme décrit également une dénivellation où à chaque point du plan temps-chroma est associé une élévation. Si on coupe la dénivellation à une hauteur donnée, on obtient le contour du chromagramme (schéma du haut, Figure 3.10). On trace ainsi plusieurs contours du chromagramme en ne gardant par la suite que le chroma principal, c'est-à-dire le chroma possédant le maximum d'énergie (schéma du bas, Figure 3.10). C'est ce qu'on appelle le *chromatimbre*, qui est indépendant de la note jouée. Le travail s'effectuera alors sur un ensemble de points du contour au lieu de la matrice du chromagramme. Actuellement, aucune hypothèse n'est faite quant à la façon d'obtenir les contours; puisqu'une fonction de la plateforme de développement est disponible et fonctionnelle cette spécificité est laissée à une étude ultérieure.

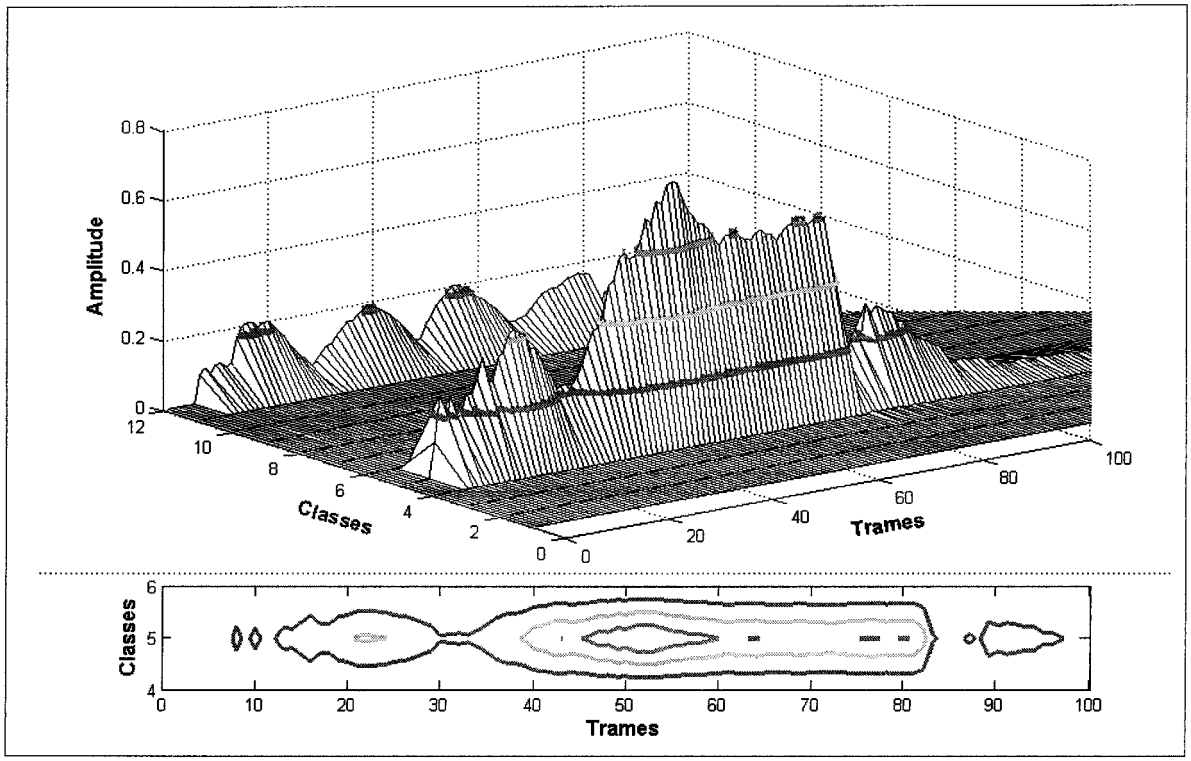

Figure 3.10 Traçage du chromatimbre d'une note d'orgue.
Les contours du chromagramme sont obtenus par des coupes du chromagramme (en haut).
Le chromatimbre est déterminé par les contours du chroma ayant le maximum d'énergie (en bas).

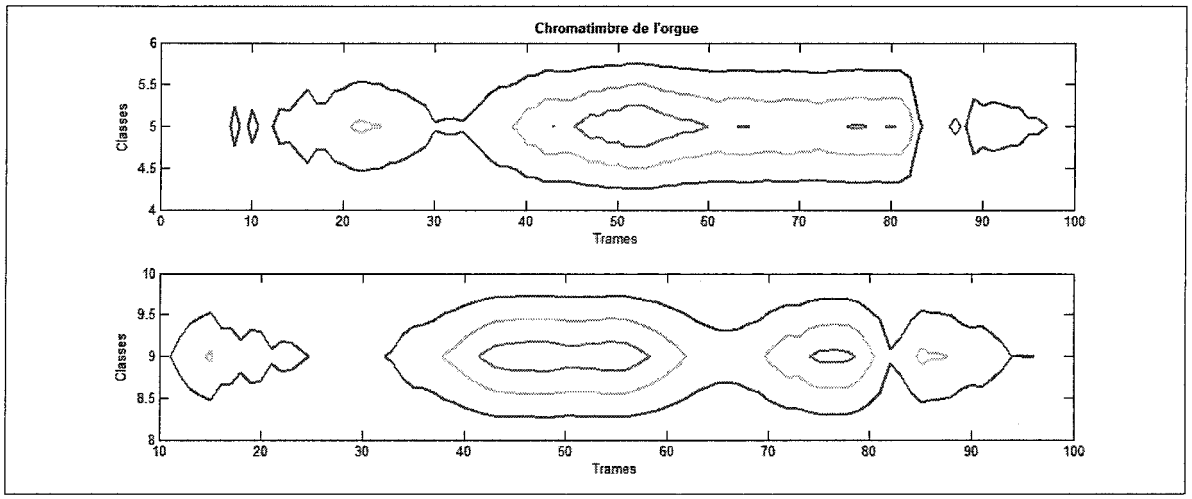

Figure 3.11 Traçage du chromatimbre de deux notes d'orgue.

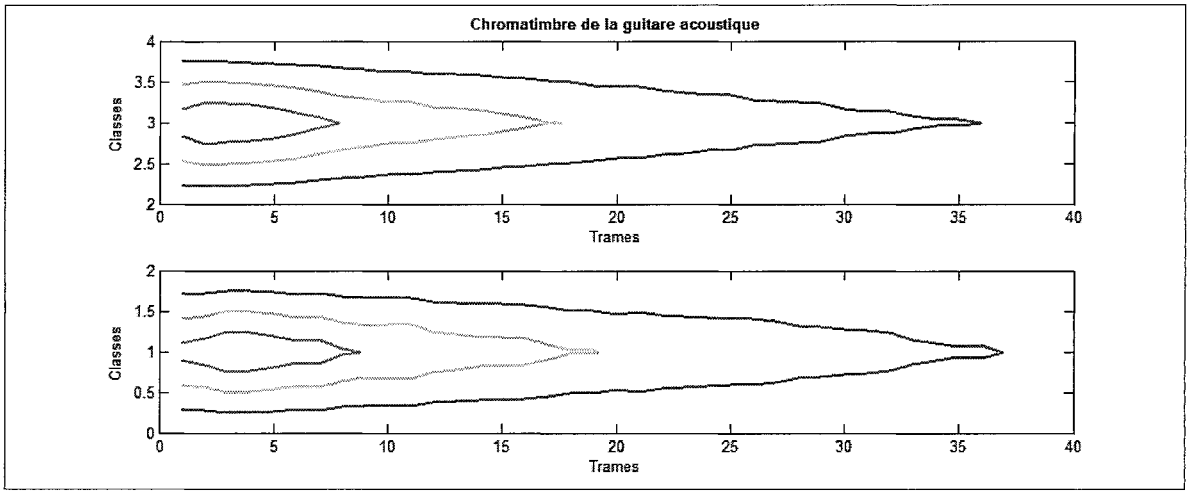

Figure 3.12 Traçage du chromatimbre de deux notes de guitare acoustique.

3.2.4.4 Moments invariants du chromatimbre

On cherche à caractériser le chromatimbre en utilisant une représentation statistique qui sera :

- Invariante par translation
- Invariante par rotation
- Invariante par homothétie

On définit pour ce faire les moments statistiques d'ordre (*p,q*)

$$M_{pq} = \sum_x \sum_y x^p y^q \, I(x,y) \qquad \textbf{(3.25)}$$

avec *x* et *y* les coordonnées du contour et $I(x,y)$ l'intensité du contour. On définit les moments centraux des variables centrées $x = x - \bar{x}$ et $y = y - \bar{y}$:

$$\mu_{pq} = \sum_x \sum_y (x - \bar{x})^p (y - \bar{y})^q I(x, y) \quad \textbf{(3.26)}$$

où $\bar{x} = \frac{M_{10}}{M_{00}}$ et $\bar{y} = \frac{M_{01}}{M_{00}}$ est le barycentre du contour. Il est démontrable que les moments centraux sont invariants par translation [41]. On construit des moments invariants par homothétie et par translation avec :

$$\eta_{ij} = \frac{\mu_{ij}}{\mu_{00}^{\left(1+\frac{i+j}{2}\right)}} \quad \textbf{(3.27)}$$

On obtient finalement sept (7) moments invariants par translation, homothétie et rotation [41, 42] :

$$\begin{aligned}
M_1 &= \eta_{20} + \eta_{02} \\
M_2 &= (\eta_{20} - \eta_{02})^2 + (2\eta_{11})^2 \\
M_3 &= (\eta_{30} - 3\eta_{12})^2 + (3\eta_{12} - \eta_{03})^2 \\
M_4 &= (\eta_{30} + \eta_{12})^2 + (\eta_{21} + \eta_{03})^2 \\
M_5 &= (\eta_{30} - 3\eta_{12})(\eta_{30} + \eta_{12})[(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2] \\
&\quad + (3\eta_{21} - \eta_{03})(\eta_{21} + \eta_{03})[3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2] \\
M_6 &= (\eta_{20} - \eta_{02})[(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2] \\
&\quad + 4\eta_{11}(\eta_{30} + \eta_{12})(\eta_{21} + \eta_{03}) \\
M_7 &= (3\eta_{21} - \eta_{03})(\eta_{30} + \eta_{12})[(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2] \\
&\quad - (\eta_{30} - 3\eta_{12})(\eta_{21} + \eta_{03})[3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2]
\end{aligned} \quad \textbf{(3.28)}$$

3.3 Normalisation des paramètres du vecteur d'observation

Généralement une normalisation des paramètres du vecteur d'observation permet d'obtenir de meilleurs résultats en réduisant la plage dans lequel s'étendent les valeurs des paramètres du vecteur d'observation. Pour cette raison, les simulations ont étés effectuées avec ou sans les normalisations suivantes : mu-sigma et min-max (ces normalisations sont décrites plus bas). Lors

de toutes les simulations, les données d'apprentissage servirent de données d'inférences aux valeurs statistiques (moyenne, variance, minimum et maximum) dans les calculs de normalisation; les données de tests furent donc normalisées avec les statistiques obtenues à partir des données d'apprentissage.

La normalisation mu-sigma est obtenue en réduisant chacun des paramètres par sa variance et en les centrant par leur moyenne de la façon suivante :

$$\tilde{x}_{ij} = \frac{x_{ij} - \bar{x}_j}{\sigma_j} \qquad \textbf{(3.29)}$$

avec $x_i = (x_{i1}, x_{i2}, \dots, x_{ij}, \dots, x_{in})$ le i^{ème} vecteur d'observation, $\bar{x}_j$ et σ_j la moyenne et la variance du j^{ième} paramètre et $\tilde{x}_i = (\tilde{x}_{i1}, \tilde{x}_{i2}, \dots, \tilde{x}_{ij}, \dots, \tilde{x}_{in})$ le vecteur d'observation centré-réduit obtenu de la normalisation.

La normalisation min-max est obtenue en réduisant les paramètres entre les valeurs $[-1,1]$ par :

$$\tilde{x}_{ij} = \frac{x_{ij}}{\max(|x_{ij}|)} \qquad \textbf{(3.30)}$$

CHAPITRE 4

RECONNAISSANCE ET CLASSIFICATION AUTOMATIQUE

L'objectif de ce chapitre est de présenter les méthodes et techniques de reconnaissance automatisée utilisées dans les simulations du présent mémoire pour la réalisation d'un système d'identification d'instruments de musique. Pour permettre au lecteur de se familiariser avec le vocabulaire lié au domaine de l'intelligence artificielle, les notions théoriques des classificateurs en général sont introduites. Les algorithmes de classification qui ont été utilisés dans les simulations, soit le modèle de mélange de gaussiennes (GMM) et les k plus proches voisins (*k*-NN), sont exposés en détails. De plus, deux algorithmes de réduction de la dimension ont été utilisés dans les expérimentations : l'analyse en composante principale (PCA) et la sélection séquentielle des descripteurs (SBS et SFS). Leur fonctionnement est également expliqué en détail.

Le choix d'utiliser les classificateurs GMM et *k*-NN est principalement justifié par leur popularité et en pratique par leur robustesse en reconnaissance de formes. De plus, l'implantation de classificateurs de natures différentes permet d'estimer l'influence de l'étape de classification dans un système de reconnaissance des instruments de musique. La loi de distribution des données extraites du timbre n'étant pas encore connue *a priori*, l'utilisation d'un classificateur sans paramètres statistiques, comme le *k*-NN, permet d'éviter de postuler sur un modèle statistique particulier.

4.1 Introduction

Un algorithme de classification est un algorithme permettant de reconnaître des motifs par un ensemble de techniques et de méthodes à partir de données brutes. C'est une branche de l'intelligence artificielle qui fait largement appel aux techniques d'apprentissage automatique et aux modèles statistiques. Les motifs à reconnaître sont variés tels le contenu visuel (reconnaissance du visage, d'empreintes digitales ou de l'écriture manuscrite) ou sonore (reconnaissance de la parole, du locuteur ou d'instruments de musique). Ainsi, indépendamment de la nature des motifs, les données brutes proviennent toujours de mesures du monde réel. La connaissance du phénomène mesuré est donc essentielle étant donné le processus de reconnaissance déterminé par les grandeurs physiques et les descripteurs du phénomène. Ces descripteurs (*features* en anglais) obtenus sous formes de paramètres numériques, caractérisent les propriétés du phénomène mesuré et servent à révéler les motifs qui appartiennent à une classe donnée. L'ensemble des n paramètres numériques pour l'ensemble des descripteurs calculés sont regroupés en n-uplet formant ainsi un vecteur dit *d'observation*.

Le phénomène physique associé aux instruments de musique est déterminée par les vibrations de leur corps résonnant se propageant dans l'air ambiant. La pression acoustique générée est captée et traduite en signaux électriques à l'aide de microphones et convertie en données numériques avec des circuits électroniques de conversion analogique/numérique. Les techniques de réalisation de ce processus sont matures et ne nécessitent que peu d'intérêt dans la recherche actuelle mis à part la qualité sonore exigée par l'étude. Par exemple, pour un rapport signal-sur-bruit faible, la reconnaissance est difficile mais le contexte plus près de la réalité. Il faut également

voir à ce que la prise de données soit uniforme pour tous les enregistrements, ce qui implique d'utiliser une stratégie formelle et un équipement adéquat lors de l'enregistrement des instruments de musique. Autrement dit, s'assurer que la réverbération de la salle d'enregistrement soit constante pour tous les instruments si applicable, éviter de changer de type de micro pour des instruments similaires, mais également voir à utiliser un micro adéquat pour des instruments ayant des spécificités particulières (la dynamique des instruments, l'échelle musicale et le rayonnement acoustique nécessitent de sélectionner le bon type de micro, c'est-à-dire omnidirectionnel, unidirectionnel, compensé, largeur de bande étroite et de prendre des mesures tant ambiophoniques qu'à distance prédéterminée dépendamment de la dispersion de l'onde sonore à proximité de l'instrument).

On tente ensuite de trouver les points d'intérêts les plus susceptibles de contenir une information pertinente. Par analogie à la reconnaissance des visages où la distance entre les orbites des yeux, la grandeur de la bouche, etc. sont des descripteurs potentiels pour décrire un sous-ensemble du tableau général, un signal sonore est modélisé mathématiquement avec plusieurs descripteurs représentatifs comme l'amplitude, la fréquence fondamentale et les harmoniques. Une sélection judicieuse des descripteurs les plus significatifs permet de réduire considérablement la quantité de données nécessaires pour représenter l'information.

Finalement, la classification s'obtient à partir d'algorithmes d'intelligence artificielle et d'étude statistique sur la dispersion des données. Indépendamment des méthodes choisies, les modèles sont construits par « apprentissage », c'est-à-dire par une inspection des données d'observations qui mène à l'acquisition des paramètres des modèles. Il existe deux types d'algorithmes

d'apprentissage pour la classification : l'apprentissage supervisé et l'apprentissage non-supervisé. Un apprentissage supervisé survient lorsque l'on possède un ensemble d'exemples (des données dont on connait les classes associées) pour effectuer l'apprentissage. Lorsqu'aucun exemple ne peut être construit par absence d'information sur la classe d'appartenance et que seules les données d'un échantillon sont disponibles, les algorithmes d'apprentissage non-supervisés permettent de regrouper les données en groupes hétérogènes pour en extraire l'organisation naturelle.

4.2 Modèle de mélange de gaussiennes

Cette méthode de reconnaissance de motifs est un algorithme d'apprentissage supervisé, introduite par Reynolds [43] pour la reconnaissance du locuteur. Le modèle de mélange de gaussiennes (GMM pour *Gaussian Mixture Model*) est un modèle statistique exprimé selon un mélange de densités, c'est-à-dire une combinaison convexe de plusieurs densités de probabilité gaussiennes. Pour un mélange quelconque on pose :

$$\Theta_k = (\mu_k, \Sigma_k) \tag{4.1}$$

où Θ_k est le $k^{\text{ième}}$ paramètre de la loi normale multidimensionnelle de vecteur moyenne μ_k et de matrice de variance-covariance Σ_k. La densité de probabilité pour N composantes est alors définie par :

$$f = \sum_{k=1}^{N} w_k \cdot f_{\Theta_k} \tag{4.2}$$

où f_{Θ_k} est une *composante du mélange*, soit la $k^{\text{ième}}$ densité de probabilité gaussienne paramétrée par Θ_k. Pour respecter le deuxième axiome de Kolmogorov on vérifie la convexité de la combinaison linéaire soit $\sum_{k=1}^{N} w_k = 1$. Enfin, la loi normale multidimensionnelle paramétrée par $\Theta_{\mathrm{k}} = (\mu_k, \Sigma_{\mathrm{k}})$ de vecteur moyenne μ_k et de matrice de variance-covariance Σ_{k} est fournie par l'équation :

$$f_{\Theta_{\mathrm{k}}} = \frac{1}{(2\pi)^{\frac{N}{2}} |\Sigma_{\mathrm{k}}|^{\frac{1}{2}}} exp\left[-\frac{1}{2}(x-\mu_k)^T \Sigma_{\mathrm{k}}^{-1}(x-\mu_k)\right] \quad \textbf{(4.3)}$$

Les paramètres du modèle associé à la $i^{\text{ème}}$ classe, notés $\Omega_i = \{(w_{i,k}, \mu_{i,k}, \Sigma_{i,k}) | k = 1, \dots, N\}$, sont estimés en utilisant l'algorithme espérance-maximisation (EM pour *Expectation Maximisation*). Cet algorithme est un algorithme itératif monotone croissant permettant d'estimer le maximum de vraisemblance du modèle Ω avec $\widetilde{\Omega} = \{(\widetilde{w}_k, \tilde{\mu}_k, \tilde{\Sigma}_k) | k = 1, \dots, N\}$ à partir de M données d'observations $X = \{x_1, x_2, \dots, x_M\}$ tel qu'à l'itération *j+1* la probabilité soit augmentée par rapport à l'itération *j* :

$$\mathrm{p}(X|\widetilde{\Omega}_{j+1}) \geq \mathrm{p}(X|\widetilde{\Omega}_j) \quad \textbf{(4.4)}$$

4.2.1.1 Sélection, initialisation et restriction des paramètres

Plusieurs facteurs sont à considérer dans la construction du modèle GMM et de l'utilisation de l'algorithme EM pour l'inférence des paramètres. Le nombre de composantes du mélange doit être assez large pour rendre compte de la distribution des données sans toutefois être trop large pour l'ensemble d'exemples disponible à l'apprentissage. En pratique, le choix de l'ordre du modèle est normalement obtenu à l'aide de résultats empiriques [3, 43].

L'initialisation des paramètres est également importante pour éviter que l'algorithme EM converge vers un maximum local sous-optimal. Reynolds [43] ne conclut pas en faveur de méthodes élaborées par rapport à une initialisation aléatoire des paramètres dans le cadre de l'identification du locuteur. Essid [4] initialise les poids avec $w_k = \frac{1}{N}$, les centroïdes des régions de Voronoï servent de moyennes $\tilde{\mu}_k$ et les matrices de covariances sont initialisées avec des estimations empiriques de la variance des données dans chaque région de Voronoï. Eronen [3] initialise les moyennes à partir d'une sélection aléatoire parmi les données d'apprentissage suivi d'une unique itération de l'algorithme de *k*-moyennes pour initialiser les moyennes, les pondérations et les matrices de variances des composantes du mélange.

La variance étant sensible au bruit et à la quantité d'observations disponibles pour l'apprentissage, causant des singularités dans la fonction de vraisemblance du modèle, une contrainte restrictive doit être appliquée à chaque itération de l'algorithme EM. Ce dernier ayant tendance à réagir au bruit par de faibles valeurs de variance, la contrainte est formulée de la façon suivante :

$$\tilde{\sigma}^2 = \begin{cases} \sigma^2 & \text{si } \sigma^2 > \sigma_{min}^2 \\ \sigma_{min}^2 & \text{sinon} \end{cases} \qquad \textbf{(4.5)}$$

4.2.1.2 Règle de classification

L'objectif de la classification avec le modèle de mélange de gaussiennes est de trouver la classe C_i parmi M classes ayant la probabilité maximale *a posteriori* pour une séquence X d'observations données :

$$i = \arg \max_{1 \le j \le M} [\mathrm{p}(\Omega_\mathrm{j}|X)] \tag{4.6}$$

En vertu de la formule de Bayes :

$$\mathrm{p}(\Omega_\mathrm{j}|X) = \frac{\mathrm{p}(\Omega_\mathrm{j})\mathrm{p}(X|\Omega_\mathrm{j})}{\mathrm{p}(X)} \tag{4.7}$$

on réécrit l'équation (4.6) comme :

$$i = \arg \max_{1 \le j \le M} \left[\frac{\mathrm{p}(\Omega_\mathrm{j})\mathrm{p}(X|\Omega_\mathrm{j})}{\mathrm{p}(X)} \right] \tag{4.8}$$

On fait l'hypothèse que les classes C_i sont équiprobables, c'est-à-dire que $\mathrm{p}(\Omega_\mathrm{j}) = \frac{1}{M}$ est constante, et puisque $\mathrm{p}(X)$ est constante pour toutes les classes, (4.8) se simplifie :

$$i = \arg \max_{1 \le j \le M} \mathrm{p}(X|\Omega_\mathrm{j}) \tag{4.9}$$

Puisque X est un ensemble d'observations supposées indépendantes, l'indice de la classe cherchée est donné par :

$$i = \arg \max_{1 \le j \le M} \prod_{x \in X} \mathrm{p}(x|\Omega_\mathrm{j}) \tag{4.10}$$

Finalement, pour éviter les problèmes numériques de pertes de précision reliés à l'opération répétée de multiplications, on préfère utiliser le logarithme de l'expression (4.10) ce qui donne lieu à la sommation :

$$i = \arg \max_{1 \le j \le M} \sum_{x \in X} \log \mathrm{p}(x|\Omega_j) \qquad \textbf{(4.11)}$$

4.3 Les k plus proches voisins

Cette méthode de reconnaissance de motifs est un algorithme d'apprentissage supervisé. Son principal avantage provient du fait que la loi régissant les densités de probabilités des données n'a pas besoin d'être connue pour réaliser la classification. C'est donc une méthode de classification sans paramètres statistiques puisqu'aucune estimation de paramètres n'est nécessaire, comme c'est le cas avec les GMM ou la régression linéaire. En effet, la méthode des k plus proches voisins (*k*-NN : *k-Nearest Neighbours*) associe une observation *x* à la classe la plus fréquente *y* parmi les *k* plus proches voisins de *x* (Figure 4.1). Ainsi seuls deux paramètres sont nécessaires à la construction d'une fonction de prédiction *h* : la métrique *D* décrivant la proximité entre les observations et le nombre *k* d'observations considérés dans le voisinage engendré par *D*. L'algorithme se réduit ainsi :

1. Trouver les *k* plus proches observations de *x ;*
2. Utiliser une règle de décision à la majorité pour classifier une nouvelle observation.

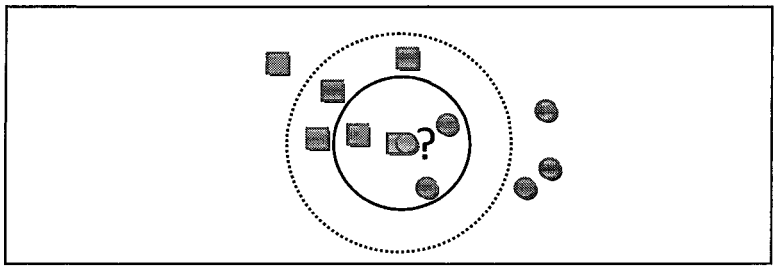

Figure 4.1 Principe de la classification avec les *k*-NN.
L'observation de test (en vert) est marqué comme appartenant à la classe des cercles si *k*=3 tandis qu'il est marqué comme appartenant à la classe des carrés si *k*=6.

L'algorithme du *k*-NN est simple d'implantation; pour un large ensemble de données il faut cependant disposer d'une mécanique efficace pour obtenir les *k* plus proches voisins de *x* puisque le calcul des distances entre chaque paire de données (x, x_i) prendrait un effort de calcul considérable. Pour palier à ce problème on peut :

- diminuer l'effort de calcul de la distance en réduisant la dimension de l'espace vectoriel (en utilisant une analyse en composante principale par exemple) ;
- utiliser une structure de données de type index spatial pour effectuer des requêtes de proximité ;
- éditer les données d'apprentissage pour éliminer les observations redondantes. De temps à autre aux alentours d'une observation, on retrouve des observations similaires qui appartiennent à la même classe.

Une haute dimension de l'espace vectoriel est également problématique puisque les observations sont dispersées et donc distantes les unes des autres. Avec un ensemble de *N* données d'apprentissage de dimension *d*, pour qu'un *k*-voisinage d'une observation *x* soit circonscrit dans un hypercube dont les côtés sont de longueur *b*, on a que $b = (k/N)^{1/d}$ [44]. En outre, pour qu'un 10-voisinage d'une observation construite de 100 paramètres soit circonscrit au maximum à un peu plus de 10% de l'espace vectoriel, un ensemble de $1\text{x}10^{100}$ données d'apprentissage sont nécessaires.

Les performances sont ainsi liées aux choix de la métrique et de *k*. Une sélection de la valeur de *k* élevée produit un lissage qui réduit le risque de sur-apprentissage dû au bruit dans les données ; une valeur de *k* trop petite résulte dans une classification sujette aux variabilités locales des

données. Par contre, une valeur de k trop grande résulte dans une classification qui est la même partout ; ultimement on sélectionne $k=N$ et l'unique classe retenue est celle majoritaire dans les données d'apprentissage.

4.4 Réduction de la dimension

La détermination des paramètres les plus représentatifs aux problèmes de classification est un enjeu primordial. La quantité de descripteurs potentiels des instruments musicaux, comme dans bien d'autres domaines de l'intelligence artificielle, ne cesse de croître dans la littérature des dernières années, imposant la nécessité de recourir à des méthodes systématiques de sélection des descripteurs les plus significatifs afin de réduire la dimension de l'espace de ces descripteurs. Une dimension trop élevée réduit les performances en généralisation, comme c'est le cas avec la méthode des k-NN, puisque l'espace des descripteurs devient difficile à modéliser d'autant plus qu'il s'étend et que les données deviennent clairsemés. Ce problème est connu depuis longtemps ; on le référence sous l'appelation *curse of dimensionality*, établie par Richard Bellman en 1961, qui signifie carrément *fléau* ou *malédiction* de la dimension. Deux aspects du problème de dimension élevée ont étés soulevés dans de nombreux ouvrages soient la concentration des distances et les « hubs »[45]. La concentration des distances est la tendance des points d'un espace de haute dimension à devenir équidistants. Les « hubs » sont des points populaires dont la fréquence d'appartenance à plusieurs k-voisinages est anormalement élevée. Tous deux biaisent les résultats de classification basées sur une fonction distance.

Les techniques de réduction de la dimension sont regroupées en deux classes : les algorithmes de *sélection des descripteurs* (*feature selection*) et les algorithmes d'*extraction des descripteurs*

(*feature extraction*) [5]. La sélection des descripteurs permet de sélectionner parmi un ensemble de paramètres de descripteurs ceux qui sont les plus susceptibles d'être significatifs; ce sont des algorithmes *ad hoc*. On s'en sert notamment pour faciliter la visualisation des données, réduire l'effort d'apprentissage et de tests et supprimer l'effet *curse of dimensionality*. Dans un autre ordre d'idée, l'extraction des descripteurs s'appuie sur des notions statistiques universelles et se compose de transformations linéaires et non-linéaires.

La désignation des descripteurs optimaux requièrent une recherche exhaustive des meilleurs paramètres des descripteurs parmi tous les sous-ensembles possibles. D'un point de vue combinatoire, l'ensemble des parties d'un ensemble de *d* éléments est de cardinalité 2^d qui est de complexité exponentielle. Au lieu de chercher l'arrangement optimal, on utilise des algorithmes de sélection des descripteurs. Les *filters* affectent un rang sur un sous-ensemble des paramètres des descripteurs en se basant sur une métrique; ils sont en conséquence indépendants de la fonction de prédiction. Les *wrappers* sélectionnent un sous-ensemble des paramètres des descripteurs en relation avec leurs performances à une fonction de prédiction donnée. Enfin, les *embedders* intègrent en un seul processus l'optimisation conjointe des *wrappers* et du classificateur [46].

En statistique, le problème de réduction de la dimension se réduit à un problème de réduction du nombre de variables aléatoires sous observation. Ainsi, mathématiquement, le problème s'énonce de la façon suivante : ayant une variable aléatoire $x = (x_1, x_2, \dots, x_d)^T$ de dimension *d* donnée, trouver une représentation simplifiée de dimension inférieure $x' = f(x) = (x'_1, x'_2, \dots, x'_k)^T$ avec $k \leq d$. Les descripteurs ne conservent pas nécessairement leur signification, le nouveau vecteur étant construit de la combinaison (linéaire ou non-linéaire) des

éléments du vecteur source. Les techniques sont nombreuses et incluent les nuages de points, les transformations linéaires (PCA/SVD, LDA), les transformations spectrales (Fourier, Hadamard) et les transformations en ondelettes [46].

4.4.1 Analyse en Composante Principale

L'analyse en composante principale (PCA pour *Principal Component Analysis*) est basée sur le théorème de décomposition en valeurs singulières et utilise la matrice de covariance des variables [47]. En essence, PCA tente de réduire la dimension des données en cherchant les combinaisons linéaires orthogonales ayant les plus hautes variances, ignorant les autres. Géométriquement, l'algorithme du PCA définit une base orthonormée de l'espace vectoriel dont la direction de chacune des coordonnées coïncide avec le maximum de variance. Les coordonnées sont ainsi ordonnées de manière à refléter l'ordre d'importance des variances.

C'est un algorithme d'extraction des descripteurs par combinaison linéaire et sa transformation est donnée par :

$$x' = \mathrm{W}x \tag{4.12}$$

où W est la matrice de transformation linéaire cherchée, x est un vecteur d'observation quelconque. En considérant la matrice de covariance empirique centrée suivante :

$$\Sigma = \frac{1}{n}\mathrm{X}\mathrm{X}^t \tag{4.13}$$

où X est la matrice d'observation de *n* lignes et *d* colonnes construite de *n* vecteurs d'apprentissage de *d* paramètres de descripteurs, on effectue une décomposition en valeurs propres de la matrice de covariance :

$$\Sigma = \mathrm{U}\Lambda\mathrm{U}^{t} \tag{4.14}$$

où $\Lambda = diag(\lambda_1, \dots, \lambda_d)$ est la matrice diagonale ordonnée des valeurs propres λ_i de $\sum$ en imposant $\lambda_1 \geq \cdots \geq \lambda_d$ et U est la matrice des vecteurs propres de $\sum$. Puisque les vecteurs propres de U sont des maximas (d'après le théorème des multiplicateurs de Lagrange), ils représentent conséquemment les directions des plus grandes variations de l'ensemble des observations. Finalement, la projection de chacune des observations sur U est donnée par :

$$x' = \mathrm{U}^{\mathrm{t}}x \tag{4.15}$$

et la transformation cherchée selon l'équation (4.12) est donnée par $\mathrm{W} = \mathrm{U}^{\mathrm{t}}$. En tronquant le vecteur x' à la $i^{ème}$ coordonnée $i = \min\{j | \lambda_j \leq \alpha\}$ pour un seuil α donné, on ne conserve que les $k = d - i$ coordonnées les plus significatives et on diminue ainsi la dimension du vecteur d'observation.

4.4.2 Sélection séquentielle des descripteurs

Cette méthode permet de sélectionner les paramètres des descripteurs selon un score obtenu d'une fonction objectif. Elle permet, contrairement au PCA, de dégrossir l'ensemble des paramètres des descripteurs disponibles en ne conservant que les plus significatifs au problème. Ceci a pour avantage de faire émerger les descripteurs les plus susceptibles d'apporter une

influence sur le système et d'apporter par le fait même une meilleure compréhension du phénomène. D'un point de vue architectural, la sélection séquentielle des descripteurs permet de sélectionner un sous-ensemble des paramètres des descripteurs et éviter ainsi de calculer tous les descripteurs comme l'exigerait un algorithme d'extraction des descripteurs. Son implantation est simple cependant l'algorithme peut converger vers une solution sous-optimale, comme en démontre l'exemple à la Figure 4.2.

La séquence de parcours de l'algorithme peut s'effectuer en ajoutant des paramètres, dans quel cas l'algorithme est nommé sélection séquentielle en avant (*Sequential Forward Selection* ou *SFS*), ou bien à l'inverse, en retirant des paramètres, dans quel cas l'algorithme est nommé sélection séquentielle en arrière (*Sequential Backward Selection* ou *SBS*) [48]. L'itération en SFS débute avec un ensemble vide de paramètres (ou bien la totalité des paramètres dans le cas SBS). À chaque nouvelle itération, parmi tous les paramètres non-sélectionnés, on détermine celui qui, combiné aux paramètres précédemment sélectionnés, obtiendra le meilleur score au regard de la fonction objectif. Plus formellement, l'algorithme SFS se résume ainsi :

$$\begin{aligned} Y_0 &= \{\emptyset\} \\ Y_{i+1} &= Y_i \cup \left\{\arg\max_{x \notin Y_i} J[Y_i \cup \{x\}]\right\} \end{aligned} \tag{4.16}$$

pour un ensemble de paramètres $x \epsilon X$ et une fonction objectif J donnée. La dernière itération est soit fixée à k paramètres auquel cas $|Y_k| = k$ ou jusqu'à ce que $J(Y_{i+1}) - J(Y_i) \leq \varepsilon$, c'est-à-dire qu'une nouvelle itération n'apporte qu'un gain ε non-significatif ou encore une perte d'information.

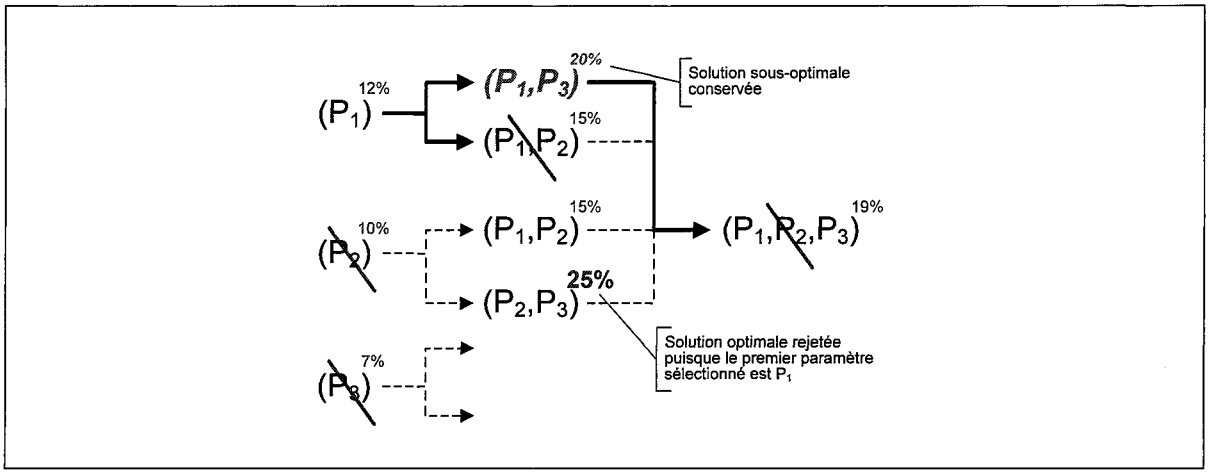

Figure 4.2 Représentation schématique de la sélection séquentielle en avant SFS. La séquence de parcours de l'algorithme ajoute le paramètre donnant un taux de reconnaissance maximum à chaque itération. La solution peur être sous-optimale, comme en démontre l'exemple ci-dessus : le premier paramètre sélectionné P_1 donne un taux de reconnaissance plus élevé que P_2 et P_3 prient individuellement. Le taux de reconnaissance est cependant mais moins élevé lorsque P_2 et P_3 sont combinés comparativement à la seconde itération qui combine P_1 et P_2 après avoir sélectionné P_1.

4.4.3 Réduction de la dimension dans une classification hiérarchique

Lors de la classification hiérarchique, la sélection des paramètres à chaque nœud de la classification hiérarchique est déterminée de façon dynamique à l'aide des algorithmes décrits précédemment. Ceci permet d'adapter le vecteur d'observation à chaque nœud de la hiérarchie. La réduction de la dimension s'effectue à l'aide d'un prétraitement avant la validation croisée comme dans le cas d'une classification directe sans hiérarchie. Mais contrairement à la réduction de la dimension dans le contexte d'une classification directe, la réduction de la dimension dans une classification hiérarchique est appliquée successivement à chaque nœud de la hiérarchie. Ainsi les paramètres sélectionnés, dans le cas de la sélection séquentielle des descripteurs (SBS et SFS) par exemple, sont conservés dans un arbre hiérarchique ayant la même structure que la taxonomie utilisée dans la classification hiérarchique. Lors de la classification hiérarchique, le

vecteur d'observation est réduit en considérant les paramètres présélectionnés pour le nœud en cours de classification.

CHAPITRE 5

EXPÉRIMENTATIONS ET RÉSULTATS

L'objectif de ce chapitre est de comparer les différents descripteurs des instruments de musique sélectionnés et présentés dans le Chapitre 3. L'approche proposée est d'analyser, à l'aide de la classification hiérarchique présentée dans la section 2.1.1, les paramètres efficaces pour la caractérisation du timbre. Une sélection séquentielle en avant « *Sequential Forward Selection* » (SFS), une sélection séquentielle en arrière « *Sequential Backward Selection* » (SBS) ainsi qu'une réduction par analyse en composante principale (PCA) ont été appliquées à tous les nœuds de la hiérarchie, permettant ainsi de faire évoluer le vecteur d'observation de façon dynamique et de faire émerger les descripteurs significatifs pour chaque classe de la taxonomie. Les simulations réalisées sont divisées en quatre parties différentes :

- La première partie concerne l'optimisation des paramètres des algorithmes d'extraction et de classification; en outre les paramètres du chromatimbre dans une classification directe sont visités en détail.
- La seconde partie concerne la reconnaissance des instruments dans une classification directe (sans hiérarchie).
- La troisième partie concerne la reconnaissance des instruments dans une classification hiérarchique utilisant les taxonomies proposées dans la section 2.1.

- La dernière partie propose une analyse psycho-visuelle du chromatimbre et permet d'apprécier les capacités de ségrégation de cette nouvelle représentation.

Les taux présentés dans les résultats sont définis comme étant les taux de réussite à la classification des instruments (ou des familles d'instruments), c'est-à-dire le ratio entre le nombre d'instruments correctement classifiés et le nombre total des instruments à classifier. Les termes « taux de reconnaissance », « taux de classification » et « score » sont utilisés conjointement pour signifier le taux de réussite à la classification. La totalité des simulations a été effectuée avec une validation croisée de 10% de données de test et 90% de données d'apprentissage. Ce rapport permet une bonne analyse sans trop biaiser les résultats et permet d'effectuer une validation qui n'est pas excessivement onéreuse [49]. Les données utilisées lors des simulations sont celles énumérées dans le Tableau 2.2, c'est-à-dire 6 698 notes isolées provenant de 29 instruments.

5.1 Sélection des paramètres des algorithmes

5.1.1 Algorithmes de classification

Dans toutes les simulations, la métrique utilisée pour le classificateur *k*-NN est la distance euclidienne avec un nombre de voisinage fixé à 4, valeur déterminée par des tests empiriques. Eronen [3] ne spécifie pas la valeur du voisinage qu'il a utilisé tandis qu'Essid [4] utilise une règle d'or en fixant $k = \sqrt{nombre\ données\ apprentissage}$, lequel ne donne pas de résultat satisfaisant à notre égard. Le nombre de modèles pour le classificateur GMM fut également fixé à 4, valeur déterminée par des tests empiriques (Figure 5.1). Eronen [3] et Essid [4] fixèrent leur nombre de gaussiennes à 8; cependant Eronen ne justifie pas son choix et les tests d'Essid démontrent qu'au-delà de cette valeur, les performances se détériorent.

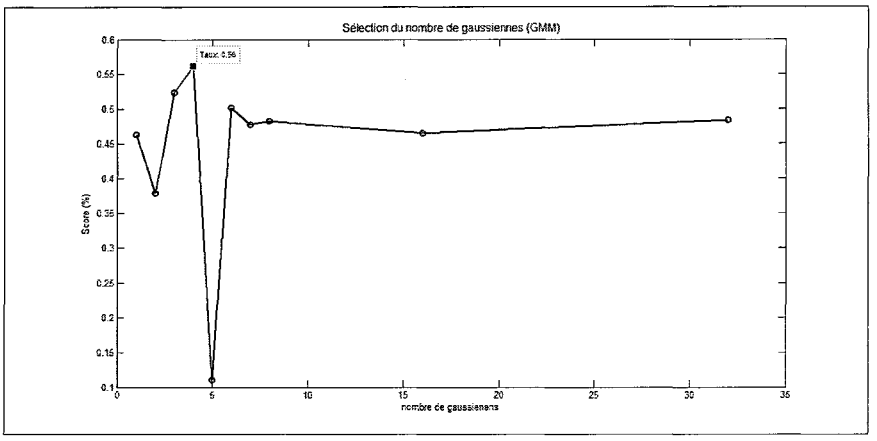

Figure 5.1 Sélection du nombre de mélange de l'algorithme GMM. (Classification directe, moyenne des trames et normalisation mu-sigma). Le score maximal est obtenu avec 4 gaussiennes.

5.1.2 Algorithmes d'extraction des descripteurs

5.1.2.1 Paramètres du vecteur d'observation

Les descripteurs ont été regroupés en cinq groupes pour un total de 45 paramètres, soient 13 coefficients MFCC, 14 coefficients LPC, 4 moments spectraux, 7 moments invariants du chromagramme et 7 paramètres de l'enveloppe. Les paramètres de l'enveloppe incluent le taux de passage par zéro (ZCR), le temps d'attaque, la pente de l'enveloppe (obtenue par régression linéaire) et 4 moments statistiques appliqués à l'enveloppe c'est-à-dire son barycentre, sa largeur, son asymétrie et son kurtosis (au même titre que les moments spectraux).

Pour permettre de différencier l'apport du chromatimbre à la tâche de reconnaissance d'instrument de musique, les simulations ont été effectuées alternativement avec des vecteurs d'observation constitués de tous les paramètres des descripteurs avec chromatimbre et de tous les paramètres des descripteurs hormis le chromatimbre. Un vecteur d'observation constitué de tous les paramètres des descripteurs est nommé « vecteur entier » (pour un total de 45

paramètres) tandis qu'un vecteur d'observation dont le chromatimbre est absent est nommé « vecteur sans chromatimbre » (pour un total de 38 paramètres).

5.1.2.2 Extraction de l'enveloppe

L'extraction de l'enveloppe utilise le pseudo-histogramme du spectrogramme, inspiré des résultats du chromatimbre (voir section 5.1.4). Concrètement, le pseudo-histogramme du spectrogramme est la moyenne du spectre sur chacune de ses trames. Toutefois, pour « lisser » le spectrogramme, un filtre spatial 2D, dont le noyau est une moyenne mobile d'ordre 20 X 100, fut appliqué à l'image du spectrogramme avant de calculer la moyenne (Figure 5.2). Le spectrogramme fut calculé sur des fenêtres de « hamming » de 10ms, soit de 441 échantillons, entrelacées de 50%; la fréquence maximale fut limitée à 4kHz.

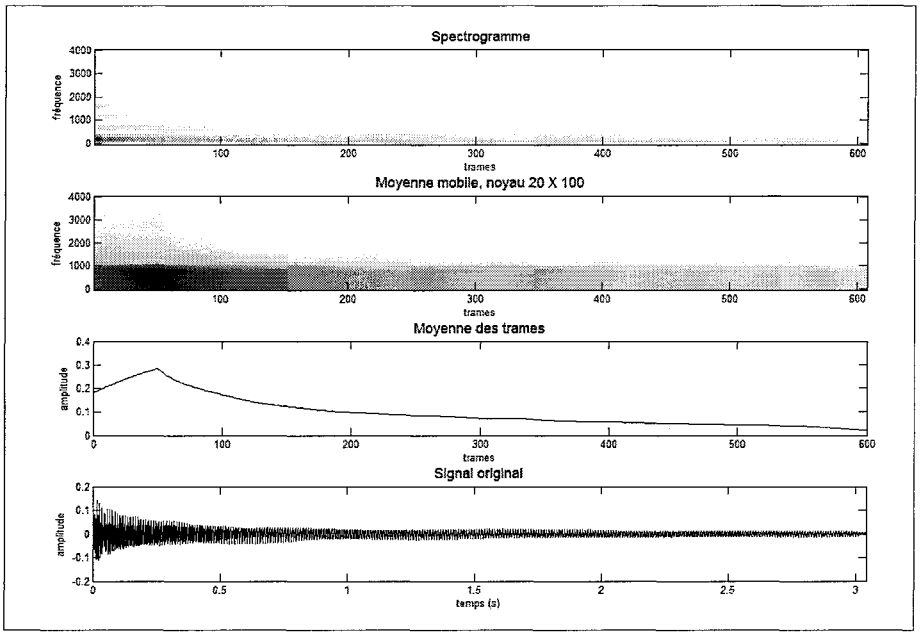

Figure 5.2 Pseudo-histogramme d'une note de guitare acoustique.
Extraction de l'enveloppe d'une note de guitare acoustique à l'aide du pseudo-histogramme du spectrogramme. Une moyenne mobile ayant un noyau de 20×100 a permis de « lisser » le spectrogramme.

Pour comparer les performances de la méthode du pseudo-histogramme, deux autres algorithmes d'extraction de l'enveloppe utilisés directement sur le signal ont été implantés : une moyenne mobile et un filtre passe-bas. La moyenne mobile fut fixée à 10 000 échantillons pour un signal échantillonné à 44,1 kHz, soit des fenêtres de 226,7ms. Le filtre passe-bas est un filtre de type RII (réponse impulsionnelle infinie) du premier ordre ayant une constante de temps fixée à 20ms (donc une fréquence de coupure d'environ 7.96 Hz). La méthode du pseudo-histogramme du spectrogramme obtient le meilleur score de classification en comparaison avec un filtre passe-bas et une moyenne mobile (Figure 5.3). Une simulation avec des vecteurs d'observation construits à partir de la combinaison des trois méthodes d'extraction fut également appliquée.

Un fait intéressant apparait dans les résultats de la Figure 5.3 concernant la combinaison des trois algorithmes d'extraction de l'enveloppe. Un taux d'un peu plus de 60% (par rapport à 67% pour les coefficients MFCC, Figure 5.8) est obtenu en n'utilisant que l'enveloppe comme descripteur et différentes manières de le calculer. On peut donc spéculer que l'enveloppe joue un rôle primordial dans le timbre de l'instrument. Également, on peut suggérer que le fait d'utiliser différentes approches d'extraction pour un même descripteur permet d'extraire une information plus riche sur l'empreinte de l'instrument de musique.

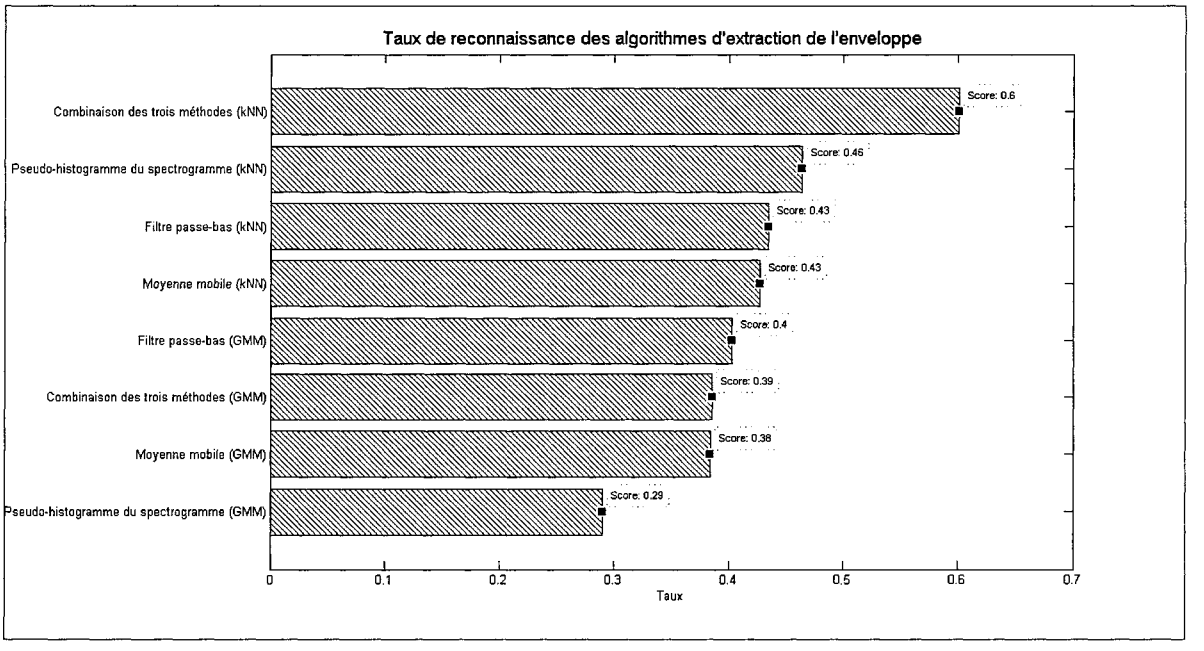

Figure 5.3 Sélection de l'algorithme d'extraction de l'enveloppe.
Trois méthodes d'extraction de l'enveloppe sont comparées : filtre passe-bas, moyenne mobile et pseudo-histogramme du spectrogramme. Les paramètres utilisés pour la classification sont la pente, l'asymétrie, le barycentre, le kurtosis, la largeur et le temps de montée. Deux classificateurs (*k*-NN et GMM) sont utilisés en classification directe non -hiérarchique. Un vecteur d'observation combinant les trois méthodes est également construit et utilisé dans la classification.

5.1.2.3 Coefficients MFCC et LPC

Le nombre de coefficients MFCC et LPC fut déterminé en s'inspirant de l'analyse effectuée dans [3]. Le nombre de coefficients MFCC fut fixé à 13 coefficients et le nombre de coefficients LPC fut fixé à 14 coefficients.

5.1.2.4 Agrégation des trames

Deux types de descripteurs sont identifiables dans le vecteur d'observation : les descripteurs globaux et les descripteurs locaux. Les descripteurs globaux sont calculés sur l'ensemble de la note : ils apportent essentiellement une information globale (moyenne) que se soit au niveau

temporel ou spectral. Les descripteurs locaux sont quant à eux obtenus sur des fenêtres courtes. Ainsi l'information globale est distribuée en plusieurs trames : ils apportent une information ponctuelle sur l'évolution de la fréquence par rapport au temps.

Des descripteurs incompatibles, c'est-à-dire ayant plusieurs trames par note (descripteurs locaux) ou une seule trame par note (descripteurs globaux), doivent en conséquence être combinés pour construire les vecteurs d'observation. Deux scénarios de combinaison ont été comparés. Le premier scénario permet d'obtenir une liste agrégée de vecteurs d'observations ayant un seul vecteur d'observation par note. Une fonction d'agrégation, en l'occurrence la moyenne des trames, fut utilisée pour réduire les descripteurs locaux à une seule trame par note. Les descripteurs globaux sont utilisés tels quels. À l'opposé, le second scénario permet d'obtenir une liste de vecteurs d'observations ayant plusieurs vecteurs d'observation par note. Pour ce faire, les descripteurs globaux ont été dupliqués pour se conformer au nombre de trames des descripteurs locaux.

Les descripteurs locaux (spectrogramme, MFCC, LPC) ont été extraits de la façon suivante : fenêtres de « Hamming » de 50ms entrelacés à 50%.

5.1.3 Algorithme de réduction de la dimension

Le nombre de paramètres du vecteur d'observation à conserver dans l'algorithme de réduction de la dimension par analyse en composantes principales (PCA) fut déterminé par des tests empiriques et fut fixé à 34 paramètres (Figure 5.4). Le vecteur d'observation original fut le vecteur entier constitué de 45 paramètres (tels que définit dans la section 5.1.2.1).

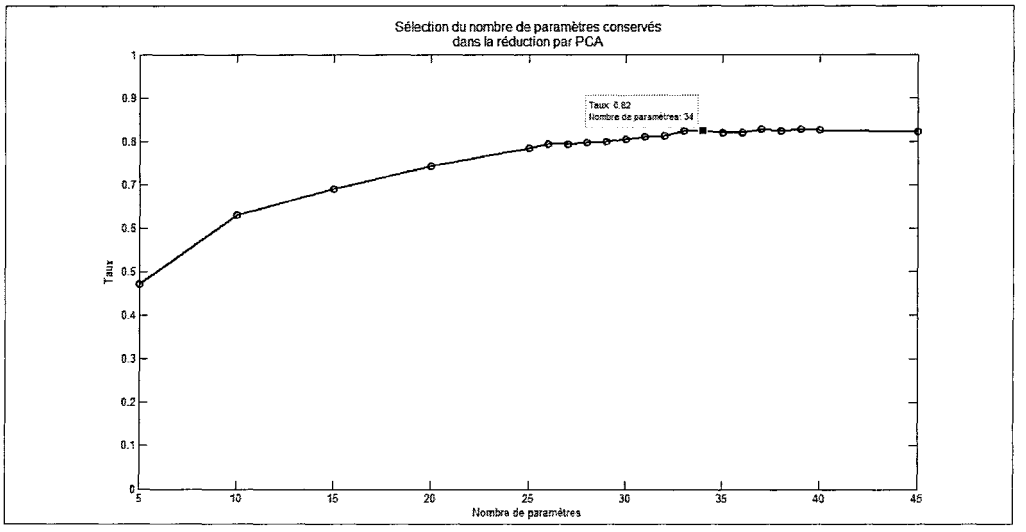

Figure 5.4 Sélection du nombre de paramètres conservées dans l'algorithme PCA. (Classification directe, moyenne des trames et normalisation mu-sigma). La dimension sélectionnée est de 34 paramètres sur un total de 45 paramètres disponibles, soit une réduction de 11 paramètres.

5.1.4 Variations des paramètres du chromatimbre

Le chromatimbre est obtenu à partir des contours du chromagramme (voir les sections 3.2.4.2 et 3.2.4.3) et est paramétré par les moments invariants d'image (voir la section 3.2.4.4). Ainsi, la forme du chromatimbre dépendra forcément des paramètres d'extraction du chromagramme, plus particulièrement du nombre de bins, mais aussi du nombre de contours constituant le chromatimbre. Pour cette raison, on s'est intéressé à étudier l'impact des paramètres d'extraction du chromatimbre sur la tâche d'identification des instruments de musique; plusieurs scénarios combinant différents paramètres d'extraction du chromatimbre sont construits. Par exemple, le nombre de contours considéré fut varié entre *automatique* (déterminé par la fonction de la plateforme), 1, 5 et 10 contours par notes. Pour lisser les contours, un filtre médian est appliqué au chromatimbre. Pour restreindre les effets de la durée des notes, une normalisation du chromatimbre est appliquée verticalement et horizontalement en redimensionnant l'image vers

une image rectangulaire de dimension fixe. Le Tableau 5.1 présente les différentes combinaisons de paramètres constituant les scénarios, pour un total de 160 combinaisons.

La meilleure performance est obtenue en utilisant 10 contours normalisés d'un chromagramme à 24 bins et un noyau de filtre médian de 1×5, pour un taux de reconnaissance de 25,6%, telle que présentée dans le Tableau 5.2. Cependant, on découvre à partir de ce tableau que les 17 meilleurs résultats sont attribuables au nombre de contours et à la normalisation uniquement. Les cinq meilleurs résultats ont tous 24 bins; on en déduit que le filtre médian n'a que peu d'effet sur les résultats.

Tableau 5.1 Variation des paramètres du chromatimbre
Différentes valeurs de paramètres du chromatimbre ont été combinées pour étudier l'impact des paramètres d'extraction du chromatimbre.

Paramètre	**Valeurs**
Nombre de contours	Automatique
	1
	5
	10
Filtre médian	Aucun
	1×2
	1×5
	1×10
	1×30
Normalisation	Sans normalisation
	Avec normalisation
Nombre de bins	12
	24
	48
	96

La Figure 5.5 montre le taux de reconnaissance des instruments par rapport à la variation du nombre de contours avec normalisation mais sans filtre médian. On remarque que le nombre de

bins influence peu les performances, c'est-à-dire moins de 10% entre le meilleur et le pire score. Les valeurs des paramètres considérés dans les simulations des sections 5.2 et 5.3 ont été sélectionnées par conséquent de la façon suivante : 10 contours normalisés d'un chromagramme à 24 bins.

Tableau 5.2 Taux de reconnaissance pour l'ensemble des scénarios du chromatimbre. (Classification directe sans hiérarchie). Le filtre médian n'y apporte que peu d'influence. Le filtre médian n'apporte que peu d'influence sur les 5 meilleurs scénarios. Les 17 meilleurs scénarios sont déterminés uniquement par la normalisation et le nombre de contours.

	normalisation	nombre de composantes	nombre de contours	filtre médian	%
1.	*oui*	*24*	*10*	1x5	25,6
2.	*oui*	*24*	*10*	1x10	25,08
3.	*oui*	*24*	*10*	1x2	24,93
4.	*oui*	*24*	*10*	1x30	24,69
5.	*oui*	*24*	*10*	aucun	24,5
6.	*oui*	48	*10*	aucun	22,04
7.	*oui*	48	*10*	1x5	21,66
8.	*oui*	96	*10*	1x30	21,56
9.	*oui*	48	*10*	1x30	21,25
10.	*oui*	12	*10*	1x30	21,25
11.	*oui*	12	*10*	aucun	21,19
12.	*oui*	12	*10*	1x10	21,13
13.	*oui*	12	*10*	1x2	21,11
14.	*oui*	96	*10*	1x10	21,01
15.	*oui*	48	*10*	1x2	20,99
16.	*oui*	48	*10*	1x10	20,93
17.	*oui*	12	*10*	1x5	20,86
18.	non	24	5	1x10	20,75
19.	oui	24	5	1x5	20,74
20.	non	24	5	1x5	20,65
...	...	...	...	...	...
160.	oui	12	1	1x5	8,85

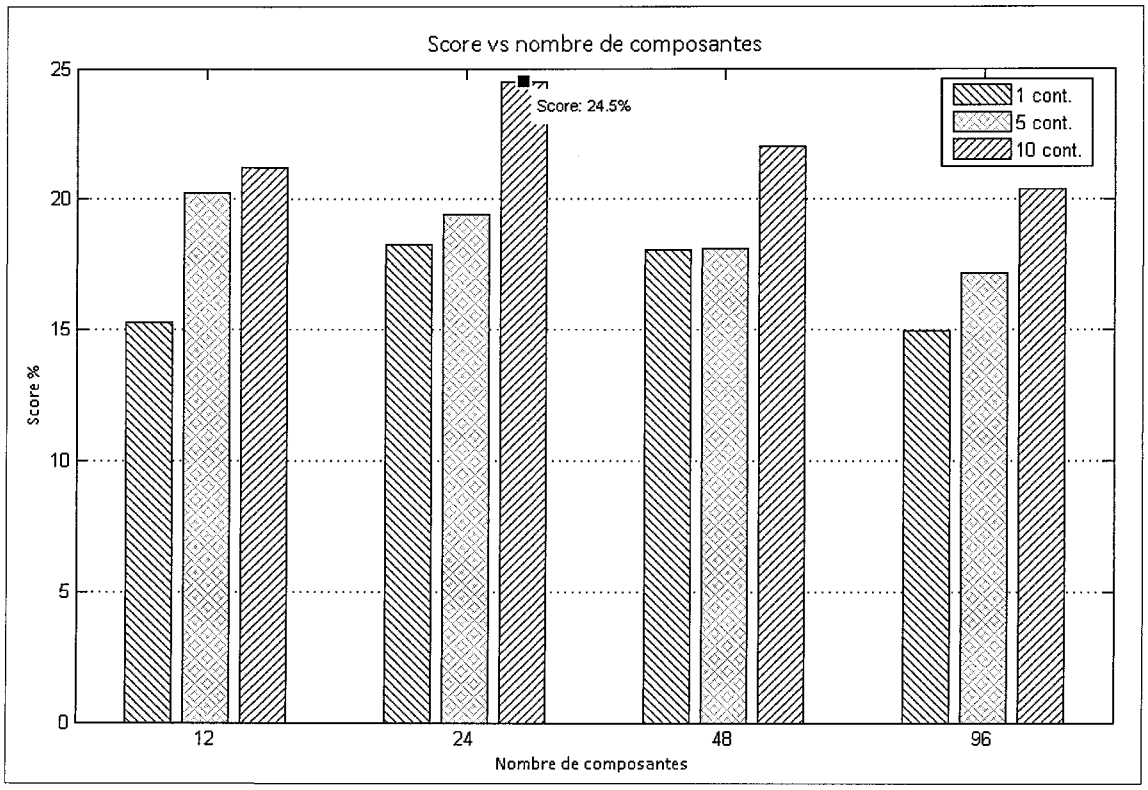

Figure 5.5 Score pour différentes valeurs de paramètres du chromatimbre.
Le nombre de composantes fréquentielles fut varié entre 12, 24, 48 et 96 composantes. Le nombre de contours fut varié entre 1, 5 et 10 contours. Classification directe sans hiérarchique avec classificateur *k*-NN, utilisation de la normalisation et sans utilisation du filtre médian.

5.1.4.1 Discussion

Un chromatimbre extrait de 10 contours normalisés d'un chromagramme de 24 bins donne le taux de reconnaissance le plus élevé lorsque paramétré avec les moments invariants d'image (Figure 5.5). Ce score est nettement insuffisant et semble indiquer que la représentation n'est pas appropriée à la tâche d'identification des instruments de musique. Dans les faits, c'est le processus de paramétrisation du chromatimbre par les moments invariants d'image qui n'est pas adéquat. Ce dernier ne tient pas compte de la dynamique temporelle et des variations d'une même note

d'un instrument. Cette affirmation est confirmée avec les simulations psycho-visuelles de la section 5.4.

Cependant, il est intéressant de remarquer que les résultats présentés à la Figure 5.6 font ressortir le pouvoir discriminant de ce descripteur par rapport au niveau hiérarchique « articulation », c'est-à-dire par rapport aux instruments *pizzicato* versus les instruments *soutenus* (voir la taxonomie présentée à la Figure 2.1). En effet, le groupe d'instruments *pizzicato* (guitare acoustique, banjo, guitare électrique, ukulélé, mandoline et basse électrique) est clairement distingué de la classe des instruments *soutenus*, l'erreur intra-classe étant plus élevée pour chacun de ces deux groupes. Puisque par définition l'enveloppe temporelle entre ces deux classes d'instruments est distincte, on en déduit qu'une information pertinente sur l'enveloppe temporelle est présente dans le spectrogramme. C'est cette constatation qui a amené l'introduction de l'extraction de l'enveloppe par le pseudo-histogramme du spectrogramme. En utilisant le pseudo-histogramme du spectrogramme on obtient un gain de plus de 10% sur le taux de reconnaissance de l'articulation par rapport à l'utilisation du pseudo-histogramme du chromatimbre (Figure 5.7).

86,85%	pizzicato	soutenus
pizzicato	1861	381
soutenus	500	3956

Figure 5.6 Matrice de confusion pour le meilleur scénario du chromatimbre.
Les instruments sont groupés en deux classes, soient *pizzicato* et *soutenus*. Une classification directe sans hiérarchique avec le classificateur *k*-NN est utilisée pour la classification. Le chromatimbre est calculé avec 10 contours normalisés d'un chromagramme de 24 bins; taux de reconnaissance de l'articulation de 86,85%.

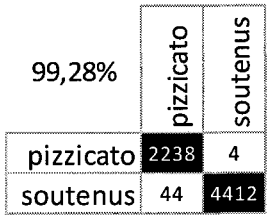

Figure 5.7 Matrice de confusion pour le pseudo-histogramme du spectrogramme. Les instruments sont groupés par articulation dans une classification directe sans hiérarchie, classificateur *k*-NN; taux de reconnaissance de l'articulation de 99,28%.

5.2 Reconnaissance dans une classification directe

Dans cette simulation, tous les descripteurs ont été utilisés; combinés entre eux ou pris individuellement. Les paramètres des algorithmes d'extraction et de classification sont ceux évalués et choisis dans la section 5.1 précédente, c'est-à-dire 10 contours normalisés d'un chromagramme de 24 bins. Une sélection séquentielle en avant SFS et une sélection séquentielle en arrière SBS, tels que décrits dans la section 4.4.2, ont permis de sélectionner les descripteurs les plus représentatifs. Une réduction de la dimension par analyse en composante principale (PCA) a également été appliquée sur le vecteur entier (voir la section 5.1.2.1 pour la définition de vecteur entier).

5.2.1 Performances des scénarios

5.2.1.1 Performance des descripteurs

Les performances des descripteurs pris ensemble et individuellement pour la classification directe, c'est-à-dire sans hiérarchie, avec classificateurs *k*-NN (normalisation mu-sigma) et GMM (normalisation min-max) sont présentées dans la Figure 5.8. Les meilleures performances sont attribuées à l'utilisation du vecteur entier pour le classificateur *k*-NN et du vecteur sans

chromatimbre pour le classificateur GMM; néanmoins les coefficients MFCC contribuent majoritairement aux résultats obtenus. On remarque que le chromatimbre, combiné aux autres descripteurs, apporte très peu de gain sur le taux de reconnaissance, d'autant plus qu'il obtient le pire score. Ce résultat est relié essentiellement à notre approche simplifiée pour paramétrer le chromatimbre avec les moments invariants d'image. La simulation psycho-visuelle de la section 5.4 confirme ce constat.

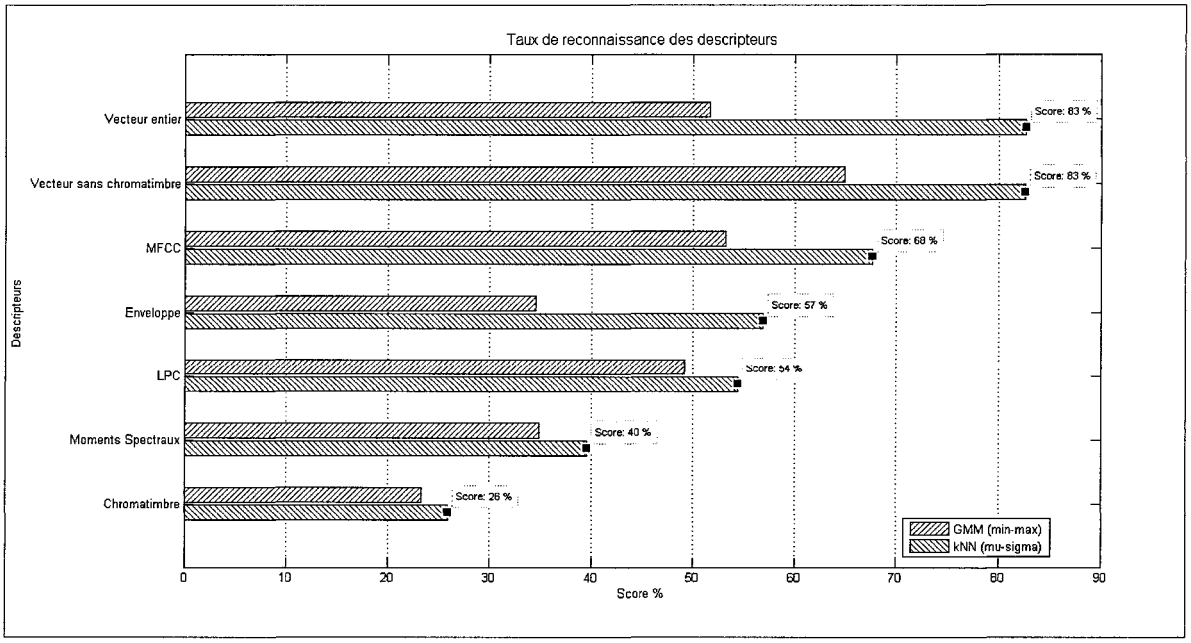

Figure 5.8 Taux de reconnaissance des descripteurs.
Les descripteurs illustrés sont : les moments spectraux, les coefficients cepstraux sur l'échelle de MEL (MFCC), les moments invariants d'image du chromatimbre, les descripteurs de l'enveloppe, les coefficients de prédiction linéaire (LPC), le vecteur d'observation entier contenant tous les descripteurs et le vecteur d'observation sans chromatimbre (voir section 5.1.2.1 pour la nomenclature). La normalisation mu-sigma est utilisée pour le classificateur *k*-NN et la normalisation min-max est utilisée pour le classificateur GMM (voir section 3.3 pour la définition des normalisations). (Moyenne des trames, classification directe).

5.2.1.2 Performance des algorithmes de sélection et de réduction

Les taux de reconnaissance des algorithmes de sélection et de réduction (voir la section 4.4 pour une description des algorithmes) pour une classification directe sans hiérarchie avec classificateur *k*-NN sont présentés dans la Figure 5.9. On constate que la sélection séquentielle en arrière SBS a un taux de reconnaissance supérieure à la réduction par PCA. Ces deux algorithmes obtiennent de meilleurs résultats qu'une classification n'utilisant pas d'algorithme de réduction de la dimension (Figure 5.8). On s'attend effectivement à ce genre de résultat puisque ces deux algorithmes ont comme propriété de retirer les paramètres dépendants et ont inévitablement des scores aux pires égaux à ceux des vecteurs non-réduits (le vecteur entier et le vecteur sans chromatimbre).

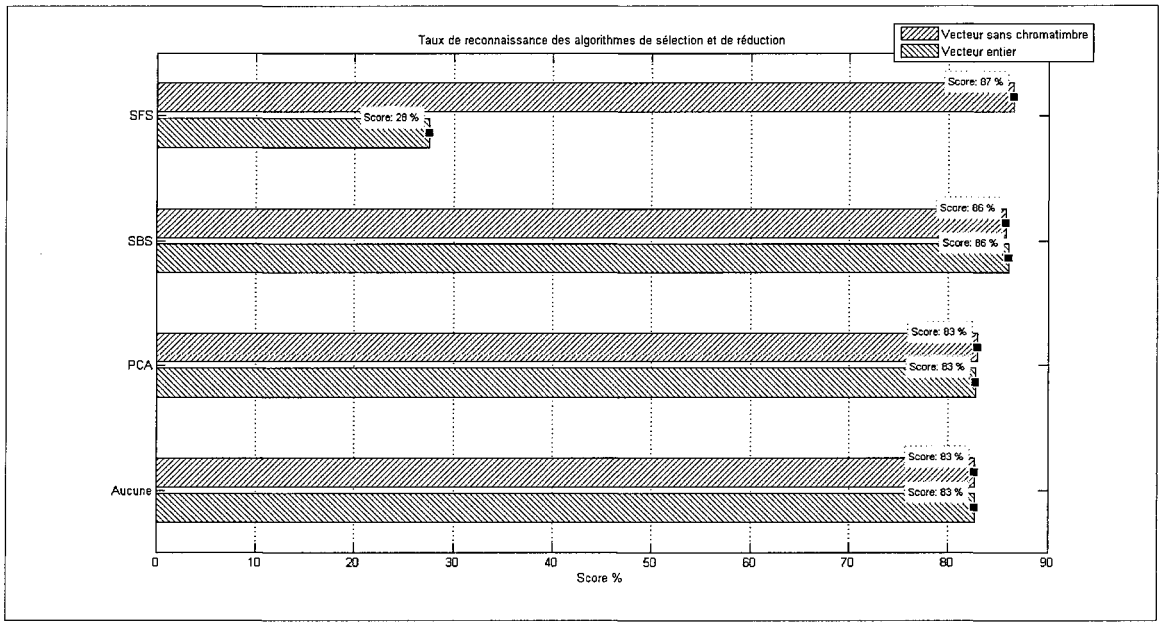

Figure 5.9 Taux de reconnaissance des algorithmes de sélection et de réduction. Les algorithmes de sélection et de réduction illustrés sont : *Sequential Forward Selection* (SFS), *Sequential Backward Selection* (SBS), analyse en composante principale (PCA) fixée à 34 paramètres et aucune réduction ni sélection. (Moyenne des trames, classificateur *k*-NN, classification directe et normalisation mu-sigma)

Cependant, la sélection SFS appliquée au vecteur entier obtient un score décevant, très en deçà des performances obtenues par tous les descripteurs. Cela provient de la forte possibilité d'atteindre un maximum local lors des premières itérations de l'algorithme. En fait, le résultat est dû principalement à la présence du chromatimbre dont les paramètres ont été normalisés; on le constate sur la Figure 5.9, l'algorithme SFS obtient le meilleur score en son absence. Utilisés individuellement, les paramètres du chromatimbre ont de meilleures performances que les autres paramètres utilisés individuellement. Le fait de normaliser les paramètres ici a un effet important sur le chromatimbre et l'algorithme SFS. La Figure 5.10 le démontre bien : en l'absence de normalisation, l'algorithme SFS donne de meilleurs résultats en itérant à partir du vecteur entier comme vecteur de départ qu'avec le vecteur sans chromatimbre comme vecteur de départ.

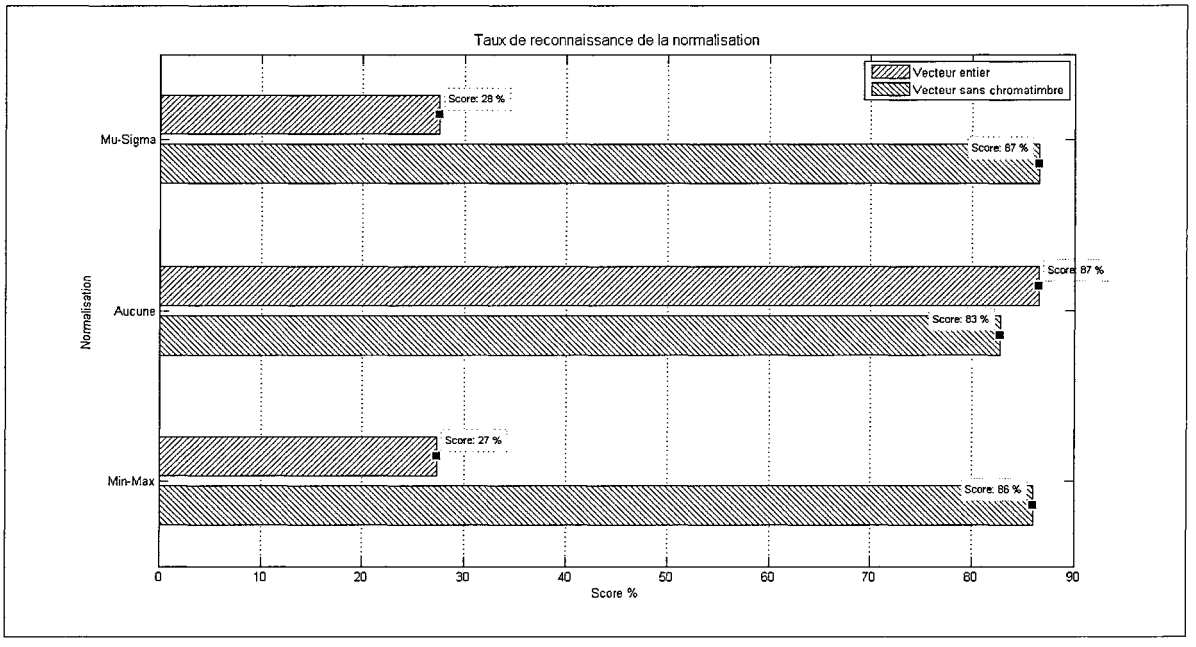

Figure 5.10 Taux de reconnaissance de la normalisation avec sélection SFS.
(Condition expérimentale : sélection séquentielle *Sequential Forward Selection* (SFS), moyenne des trames, classificateur *k*-NN, classification directe)

Les paramètres sélectionnés par l'algorithme SFS appliqué au vecteur sans chromatimbre pour chacune des catégories de descripteurs du vecteur sans chromatimbre sont listés dans le Tableau 5.3. Les paramètres ont été regroupés en 4 catégories de descripteurs soient les moments spectraux, les paramètres de l'enveloppe, les coefficients MFCC et les coefficients LPC.

Tableau 5.3 Paramètres sélectionnés par sélection SFS en classification directe. (Classification directe sans hiérarchie à partir du vecteur sans chromatimbre comme vecteur de départ et normalisation mu-sigma).

Descripteur	Paramètres
Moments spectraux	barycentre
	largeur
Enveloppe	asymétrie
	kurtosis
	barycentre
	largeur
	pente
	taux de passage par zéro
MFCC	MFCC-01
	MFCC-02
	MFCC-03
	MFCC-04
	MFCC-05
	MFCC-06
	MFCC-07
LPC	LPC-01
	LPC-03
	LPC-04
	LPC-05
	LPC-06
	LPC-07
	LPC-08
	LPC-10
	LPC-11

5.2.1.3 Effets de la normalisation

Pour les *k*-plus proches voisins, la normalisation des données n'a que peu de conséquences sur les résultats (Figure 5.11) hormis l'effet négatif à l'application de l'algorithme SFS lors de la présence du chromatimbre (Figure 5.10).

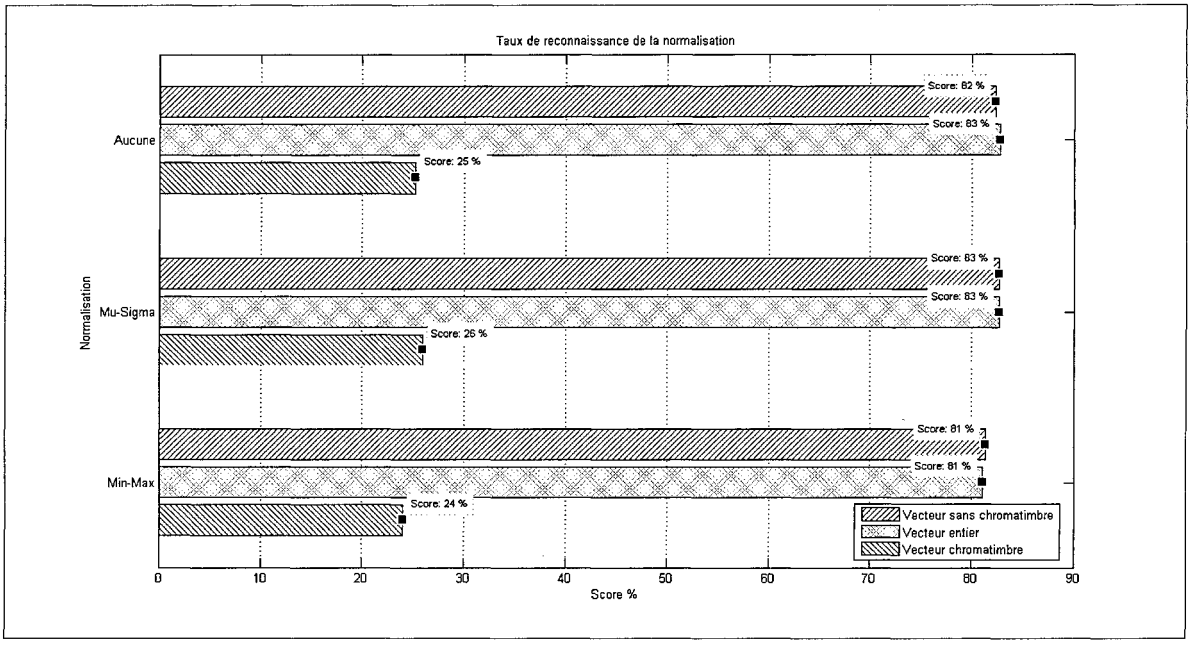

Figure 5.11 Taux de reconnaissance de la normalisation avec le classificateur *k*-NN. (Moyenne des trames et classification directe sans hiérarchie)

Le modèle de mélange de gaussiennes (GMM) étant un modèle statistique, la normalisation a un effet beaucoup plus perceptible que dans le cas des *k*-plus proches voisins. C'est la normalisation min-max qui donne le meilleur score dans le cadre d'une classification avec le vecteur sans chromatimbre (Figure 5.12). Cependant, lorsque chaque descripteur est évalué individuellement, on remarque que la nature du descripteur joue un rôle dans la sélection de la normalisation. On voit dans la Figure 5.12 que les coefficients MFCC et l'enveloppe donnent de

meilleurs taux de reconnaissance sans normalisation. Une classification qui prend en compte ce résultat devrait obtenir de meilleurs scores.

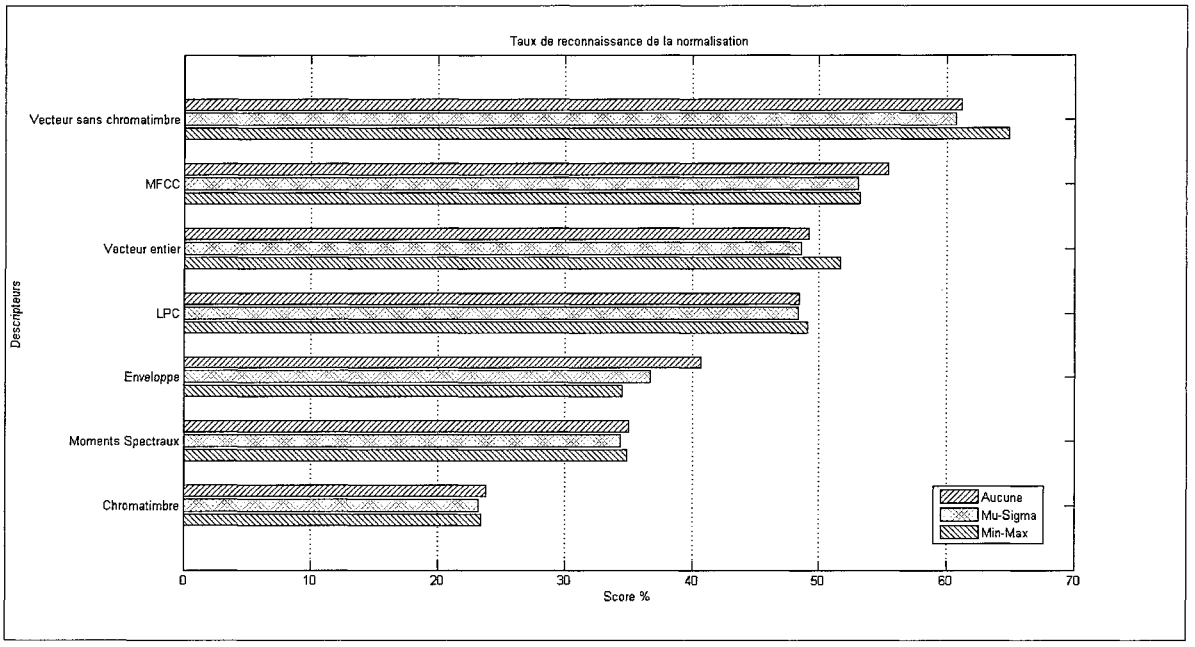

Figure 5.12 Taux de reconnaissance de la normalisation avec le classificateur GMM. (Moyenne des trames et classification directe sans hiérarchie).

5.2.1.4 Performances des classificateurs

Le classificateur qui donne les meilleurs taux de reconnaissance est *k*-NN (Figure 5.8). La classification avec les k-plus proches voisins (*k*-NN) obtient des résultats significativement supérieurs au modèle de mélange de gaussiennes (GMM). Ce résultat est différent des résultats obtenus par Eronen [3] et Essid [4], pour lesquels le classificateur GMM n'offre des taux de reconnaissance que très légèrement inférieures au classificateur *k*-NN. L'ajout du chromatimbre apporte une légère amélioration des taux de reconnaissance pour le classificateur *k*-NN. Dans le

cas des GMM, l'ajout du chromatimbre apporte une perte au taux de reconnaissance notable. Possiblement, les distributions statistiques du chromatimbre ne suivent pas des lois gaussiennes.

On remarque que le classificateur GMM donne de meilleurs scores avec les coefficients de type MFCC et LPC tandis que pour le classificateur *k*-NN, les paramètres de l'enveloppe donnent de meilleurs scores que le descripteur LPC (Figure 5.8). Un système en parallèle utilisant GMM avec les coefficients MFCC et *k*-NN avec l'enveloppe pourrait profiter de ce comportement.

5.2.1.5 Effets de l'agrégation des trames

L'utilisation de toutes les trames permet d'obtenir un jeu d'apprentissage beaucoup plus dense au détriment de calculs onéreux. Deux simulations ont été effectuées pour vérifier l'effet de l'agrégation des trames. La première consiste à réduire le nombre de trames des descripteurs locaux en une seule trame et à combiner les descripteurs globaux et les descripteurs locaux en un seul vecteur d'observation par note. La seconde consiste à dupliquer l'unique trame des vecteurs globaux en autant de trame qu'en ont les vecteurs locaux. Les deux types de descripteurs sont ensuite combinés pour former plusieurs vecteurs d'observation par note (pour une description détaillée des stratégies d'agrégation, voir la section 5.1.2.4). La classification avec l'utilisation de toutes les trames, c'est-à-dire sans agrégation, donne de meilleurs taux de reconnaissance pour les deux classificateurs (Figure 5.13). Pour être utilisable dans une application réelle, le procédé devrait s'assurer de limiter le nombre de trames par note, soit en réduisant par agrégation soit en utilisant des fenêtres plus longues, ce qui entrainerait malheureusement une perte d'information.

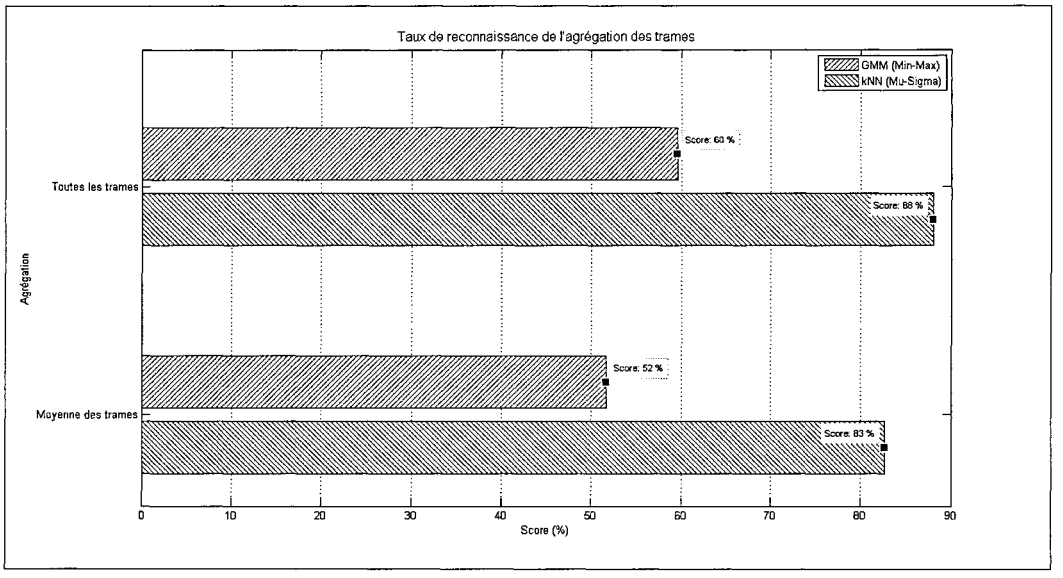

Figure 5.13 Taux de reconnaissance de l'agrégation des trames.
Comparaison entre l'effet de l'agrégation des trames par la moyenne et sans agrégation des trames sur les classificateurs *k*-NN (normalisation mu-sigma) et GMM (normalisation min-max). (Classification directe sans hiérarchie et utilisation du vecteur entier).

5.2.2 Analyse des résultats

L'utilisation de toutes les trames d'analyse, c'est-à-dire sans agrégation des trames, entraîne des calculs excessifs; les résultats de la Figure 5.13 ont par conséquent été ignorés dans la sélection du meilleur scénario. Le scénario offrant les meilleures performances d'identification des instruments en classification directe avec agrégation des trames, avec un taux de reconnaissance de 86,56%, utilise l'algorithme de sélection SFS appliqué au vecteur sans chromatimbre et utilise la normalisation mu-sigma (Figure 5.8, Figure 5.9, Figure 5.10, Figure 5.11, Figure 5.12). L'articulation est reconnue dans 99,45% des cas (Figure 5.14); la famille des instruments est reconnue dans 94,33% des cas (Figure 5.15). Les instruments sont reconnus dans 86,56% des cas (Figure 5.16). Les taux de reconnaissance de l'articulation et de l'instrument dépassent ceux d'Eronen [3]. Le taux de reconnaissance de la famille est cependant inférieur de 0,37%.

On remarque également, à partir des paramètres sélectionnés par l'algorithme SFS appliqué au vecteur sans chromatimbre comme vecteur de départ du Tableau 5.3, que les 7 premiers coefficients MFCC ont été sélectionnés, soit un peu moins de 54% du nombre de paramètres de cette catégorie de descripteurs. Également, seuls le barycentre spectral et la largeur spectrale ont été sélectionnés parmi les moments spectraux (50% du nombre de paramètres de cette catégorie de descripteurs). On s'attendait à ce résultat puisque les moments spectraux ont eu le pire score excluant le chromatimbre (Figure 5.8). Le temps d'attaque n'a pas été ajouté aux paramètres de l'enveloppe (85,71% du nombre de paramètres de cette catégorie de descripteurs) et 9 coefficients LPC ont été sélectionnés (64,29% du nombre de paramètres de cette catégorie de descripteurs).

99,45%	pizzicato	soutenus
pizzicato	2218	24
soutenus	13	4443

Figure 5.14 Matrice de confusion pour les articulations en classification directe.
(Classification directe sans hiérarchie, k-NN, sélection SFS à partir du vecteur sans chromatimbre comme vecteur de départ et normalisation mu-sigma)

94,33%	cordes	brass	flûte/piccolo	anches
cordes	3273	22	21	47
cuivres	41	1556	15	33
flûte/piccolo	9	20	537	6
anches	64	70	32	952

Figure 5.15 Matrice de confusion pour les familles en classification directe.
(Classification directe sans hiérarchie, k-NN, sélection SFS à partir du vecteur sans chromatimbre comme vecteur de départ et normalisation mu-sigma)

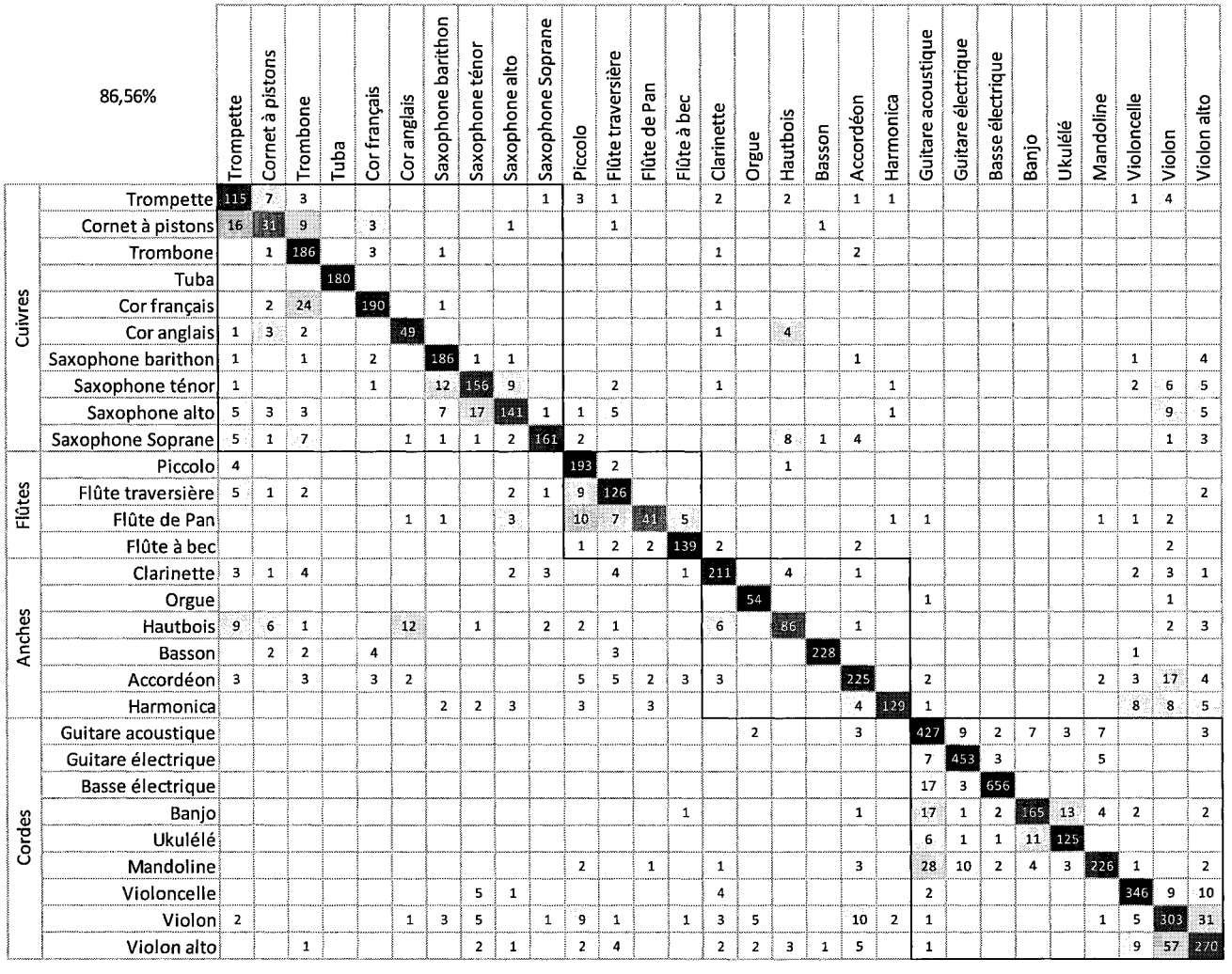

Figure 5.16 Matrice de confusion pour les instruments en classification directe.
(Classification directe sans hiérarchie, *k*-NN, sélection SFS appliquée au vecteur sans chromatimbre comme vecteur de départ et normalisation mu-sigma). Les rectangles représentent les différentes familles d'instruments.

5.3 Reconnaissance dans une classification hiérarchique

Dans cette simulation, tous les descripteurs ont été utilisés soit combinés entre eux ou pris individuellement. Les paramètres des algorithmes sont ceux évalués et choisis dans la section 5.1 précédente. Une sélection séquentielle en avant hiérarchique SFS et une sélection séquentielle en arrière hiérarchique SBS ont permis de sélectionner les paramètres les plus représentatifs pour chaque nœud de l'arbre hiérarchique. Une réduction de la dimension hiérarchique par PCA a également été appliquée à chaque nœud de l'arbre hiérarchique. La réduction de la dimension dans le contexte d'une classification hiérarchique est détaillée dans la section 4.4.3.

Une classification hiérarchique, c'est-à-dire en utilisant un arbre de classifications successives, a été utilisée dans les simulations de cette section, selon la taxonomie naturelle décrite dans la Figure 2.1. L'erreur étant propagée de la racine vers les instruments, un instrument mal classifié dans un niveau hiérarchique le sera tout au long de la classification. Toutefois, pour chaque groupe de la taxonomie, le score est propre aux instruments ayant été correctement classifiés et représente bien le taux de reconnaissance pour ce groupe. Enfin, pour chaque niveau de la hiérarchie, une sélection SFS et SBS, de même qu'une réduction par PCA, ont été appliquées pour obtenir les meilleurs taux à la reconnaissance de l'ensemble des instruments d'un même groupe.

Pour vérifier la validité de la taxonomie naturelle décrite dans la Figure 2.1, une taxonomie générée automatiquement par nuages de points (Figure 2.2) fut utilisée dans la classification hiérarchique. Les mêmes algorithmes ont été appliqués à la classification hiérarchique avec la taxonomie automatique qu'avec la taxonomie naturelle. Le principe de la taxonomie générée automatiquement par nuages de points est donné dans la section 5.3.2.

5.3.1 Performance de la taxonomie naturelle

La classification hiérarchique obtient de meilleurs taux que la classification directe avec l'utilisation d'un algorithme de sélection. Ceci est dû au fait que la sélection s'effectue à chaque niveau de la hiérarchie, déterminant ainsi les paramètres les plus influents pour chacun des groupes d'instruments (Figure 5.17). De plus, le calcul des statistiques pour la normalisation des paramètres s'effectue avec les données d'apprentissage appartenant au niveau hiérarchique uniquement, ce qui diminue le biais introduit par les données des autres groupes.

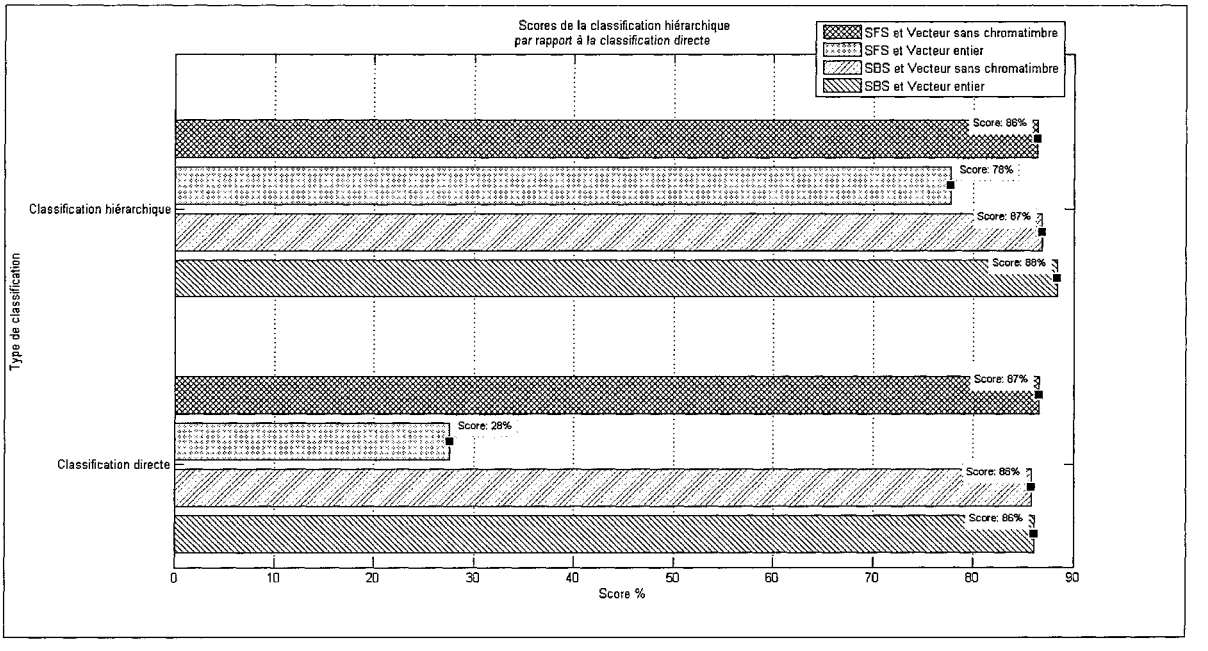

Figure 5.17 Taux de reconnaissance de la classification hiérarchique.
Classification hiérarchique utilisant la taxonomie naturelle décrite à la Figure 2.1. Deux algorithmes de sélection séquentielle sont appliqués au vecteur entier et au vecteur sans chromatimbre, soient : *Sequential Forward Selection* (SFS) et *Sequential Backward Selection* (SBS). (Classificateur *k*-NN, normalisation mu-sigma).

5.3.2 Performance de la taxonomie automatique

Pour vérifier les performances de la taxonomie naturelle proposée par Martin [2], des simulations ont été effectuées avec une taxonomie générée à l'aide d'un algorithme de classification par nuages de points. Cet algorithme de classification par nuages de points fut *k*-moyennes avec le coefficient de corrélation comme métrique de distance. Cette distance permit d'obtenir une séparation en groupes de taille hétérogène. Les résultats comparatifs sont présentés dans la Figure 5.18. La taxonomie obtenue par les *k*-moyennes est présentée dans la Figure 2.2. On remarque que les résultats sont légèrement inférieurs pour la taxonomie obtenue par nuage de points que pour la taxonomie naturelle. Ceci s'explique sans doute parce que la métrique utilisée avec *k*-NN est la distance euclidienne et celle utilisé par les *k*- moyennes est le coefficient de corrélation. L'utilisation du coefficient de corrélation comme fonction de distance pour le classificateur *k*-NN aurait sans doute donné de meilleurs résultats.

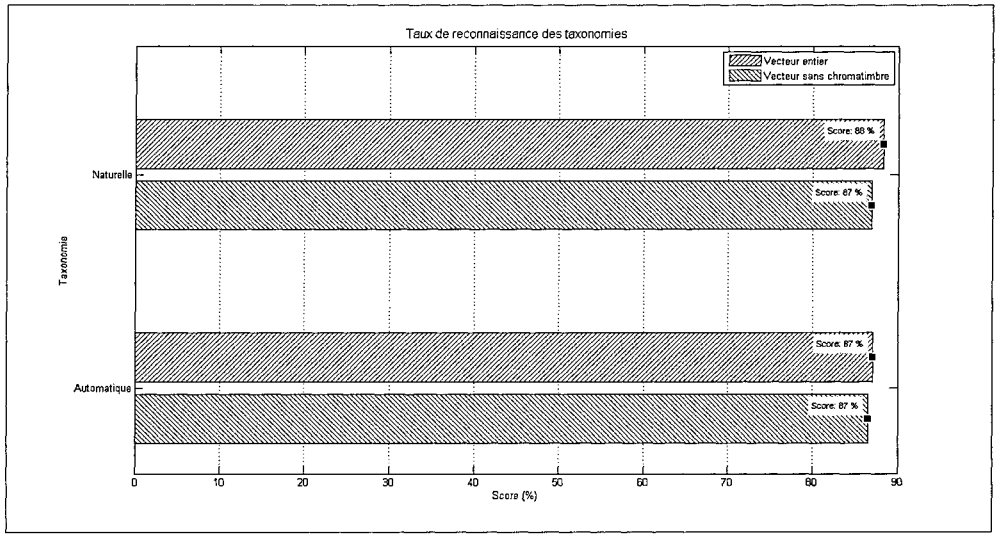

Figure 5.18 Taux de reconnaissance des taxonomies.
La taxonomie naturelle est celle décrite dans la Figure 2.1. La taxonomie automatique est obtenue par *k*-moyennes (Figure 2.2). (Classificateur *k*-NN, sélection séquentielle en arrière SBS et normalisation mu-sigma).

5.3.3 Analyse des résultats

Le scénario offrant le meilleur taux de reconnaissance pour la classification hiérarchique est construit avec le classificateur *k*-NN en utilisant le vecteur entier normalisé par la méthode mu-sigma décrite dans l'équation (3.29) et en sélectionnant les paramètres par sélection séquentielle en arrière (SBS) avec la taxonomie naturelle (Figure 5.17 et Figure 5.18). Le score obtenu est de plus de 2% au dessus de la classification directe sans hiérarchie utilisant les mêmes paramètres de classification. Contrairement à Eronen [3], la classification hiérarchique avec taxonomie naturelle obtient de meilleurs scores que la classification directe. Il faut cependant spécifier qu'Eronen avait sélectionné les paramètres manuellement pour chaque niveau hiérarchique tandis que les simulations effectués dans ce mémoire ont utilisé la sélection SFS et SBS en prétraitements pour déterminer les meilleures paramètres à conserver pour chaque niveau (voir la section 4.4.3 pour la description de la sélection des paramètres hiérarchique). Cette sélection automatique a une répercussion positive sur le score de la classification hiérarchique par rapport à la classification directe. Cependant, Essid [4] a obtenu de meilleurs scores avec une taxonomie automatique par rapport à la taxonomie naturelle. Le choix de la méthode de classification par nuages de points influence énormément les résultats et puisqu'Essid a utilisé une technique différente de celle proposée dans ce mémoire, une comparaison avec les résultats des présentes simulations est difficile.

La matrice de confusion pour la classification des familles des instruments est présentée dans la Figure 5.19. Les matrices de confusion pour chaque niveau de la classification hiérarchique sont présentées dans la Figure 5.20. La matrice de confusion pour la classification des instruments est présentée dans la Figure 5.21. On doit noter cependant que les matrices de confusions pour

chaque niveau hiérarchique de la Figure 5.20 ne peuvent être utilisées directement pour cumuler les valeurs dans les matrices de confusions des familles. En effet, au premier niveau hiérarchique, 2 233 instruments ont été correctement classifiés *pizzicato*, qui sont des instruments à cordes. À l'opposé, 1 037 instruments *soutenus* à cordes ont été correctement classifiés, ce qui donne un cumul de 3 270 instruments à cordes classifiés dans la famille des cordes. Cependant, à la Figure 5.19, un total de 3 271 instruments a été identifié comme correctement classifié dans la famille des cordes. La différence entre ces deux résultats provient d'une instance de violoncelle qui a été classifiée comme basse électrique, qui est un instrument *pizzicato* et non *soutenus*, comme on peut le constater dans la Figure 5.21. Or, cette instance n'est pas propagée vers les niveaux hiérarchiques inférieurs de la Figure 5.20 mais est plutôt conservée parmi les 8 instruments *soutenus* ayant été classifiés à tort *pizzicato*.

Les paramètres sélectionnés par sélection SBS pour chaque niveau de la hiérarchie sont présentés dans la Figure 5.22. Les paramètres ayants été sélectionnés à tous les niveaux sont : le premier moment du chromatimbre, la largeur spectrale, le kurtosis de l'enveloppe, la pente de l'enveloppe, le taux de passage par zéro, les coefficients MFCC 2, 4, 7 et 8 ainsi que les coefficients LPC 1, 7, 9 et 11.

Le niveau ayant le moins bien performé est celui contenant les instruments à archets, c'est-à-dire les instruments à cordes soutenus (violoncelle, violon et violon alto). La très forte similitude entre le violon et le violon alto entraîne une importante confusion intra-classe entre ces deux instruments. La guitare acoustique semble également être un instrument apportant de la confusion intra-classe; beaucoup d'instruments *pizzicato* lui sont confondus.

Les instruments à anches semblent également problématiques puisqu'ils sont souvent confondus avec les cuivres (presque 8% des instruments à anches sont confondus avec les cuivres). L'hautbois est notamment source de confusion inter-classe et intra-classe et aurait avantage à être classé parmi les cuivres.

94,74%	cordes	cuivres	flûte/piccolo	anches
cordes	3271	35	19	38
cuivres	22	1584	8	31
flûte/piccolo	14	20	529	9
anches	40	88	28	962

Figure 5.19 Matrice de confusion des familles de la classification hiérarchique. (Taxonomie naturelle, sélection SBS, *k*-NN et normalisation mu-sigma).

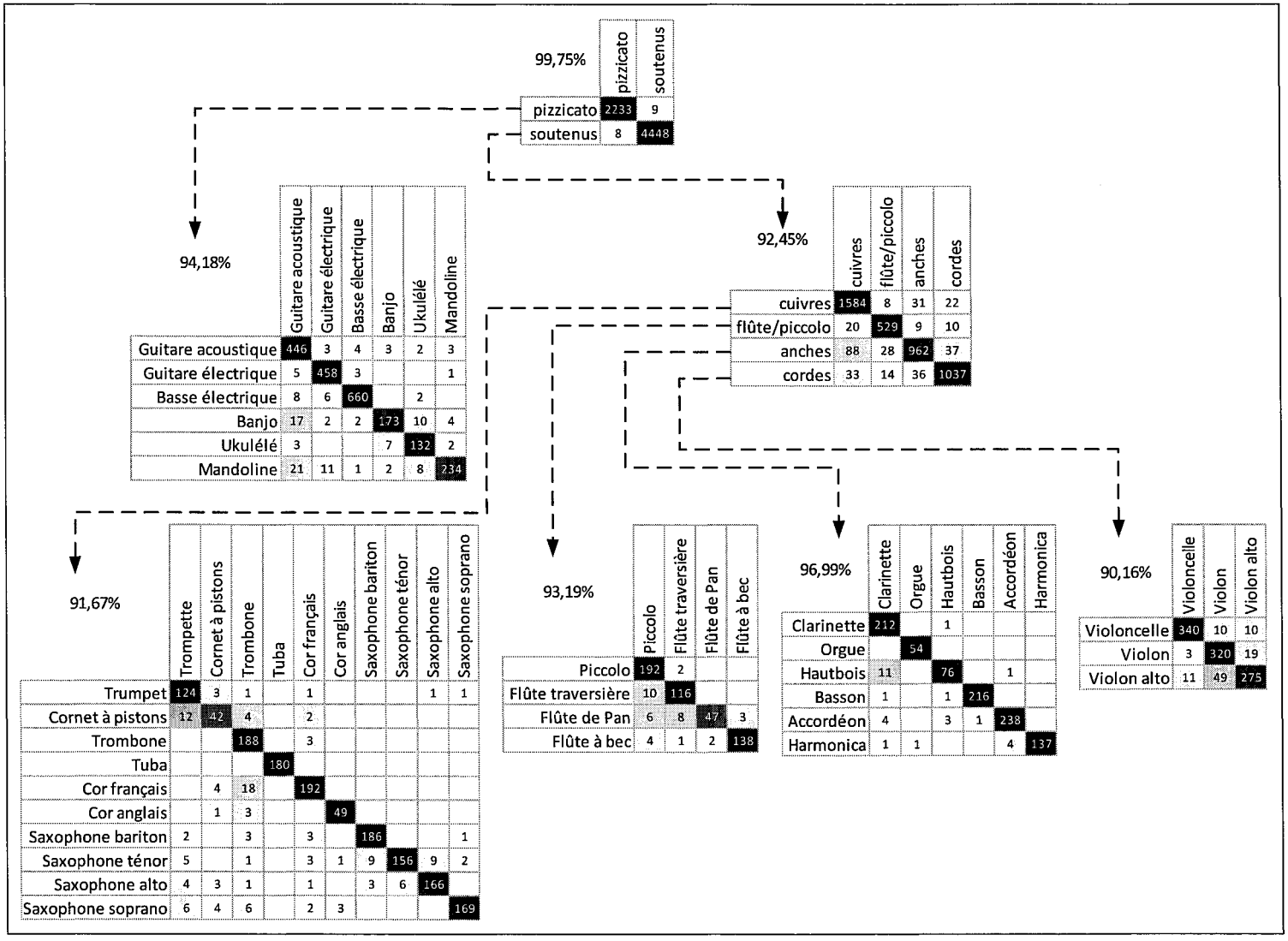

Figure 5.20 Matrices de confusion pour chaque niveau de la hiérarchie.
(Classification hiérarchique avec taxonomie naturelle, sélection séquentielle SBS, *k*-NN et normalisation mu-sigma).

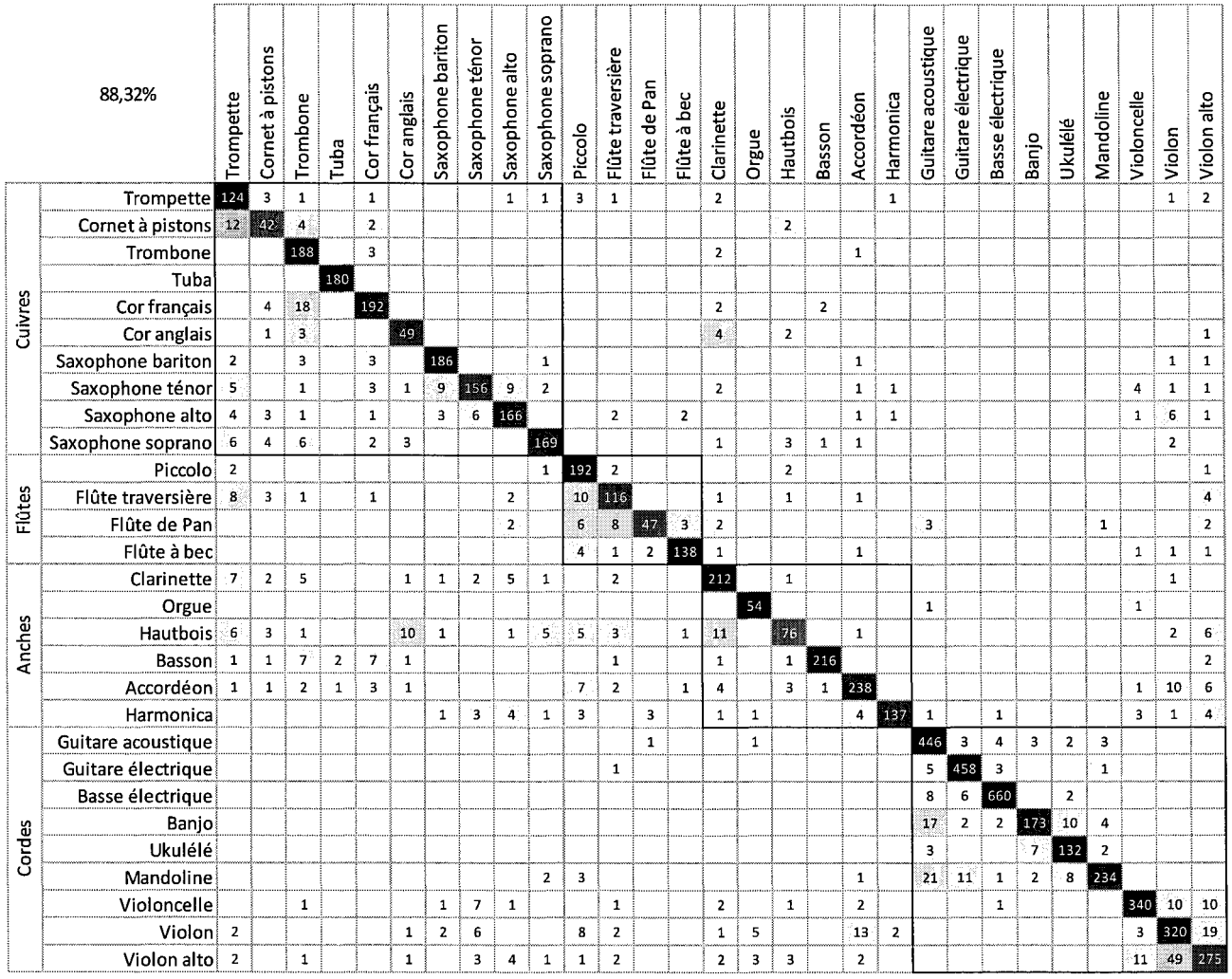

Figure 5.21 Matrice de confusion des instruments pour la classification hiérarchique.
(Taxonomie naturelle, sélection SBS, *k*-NN et normalisation mu-sigma). Les rectangles représentent les différentes familles.

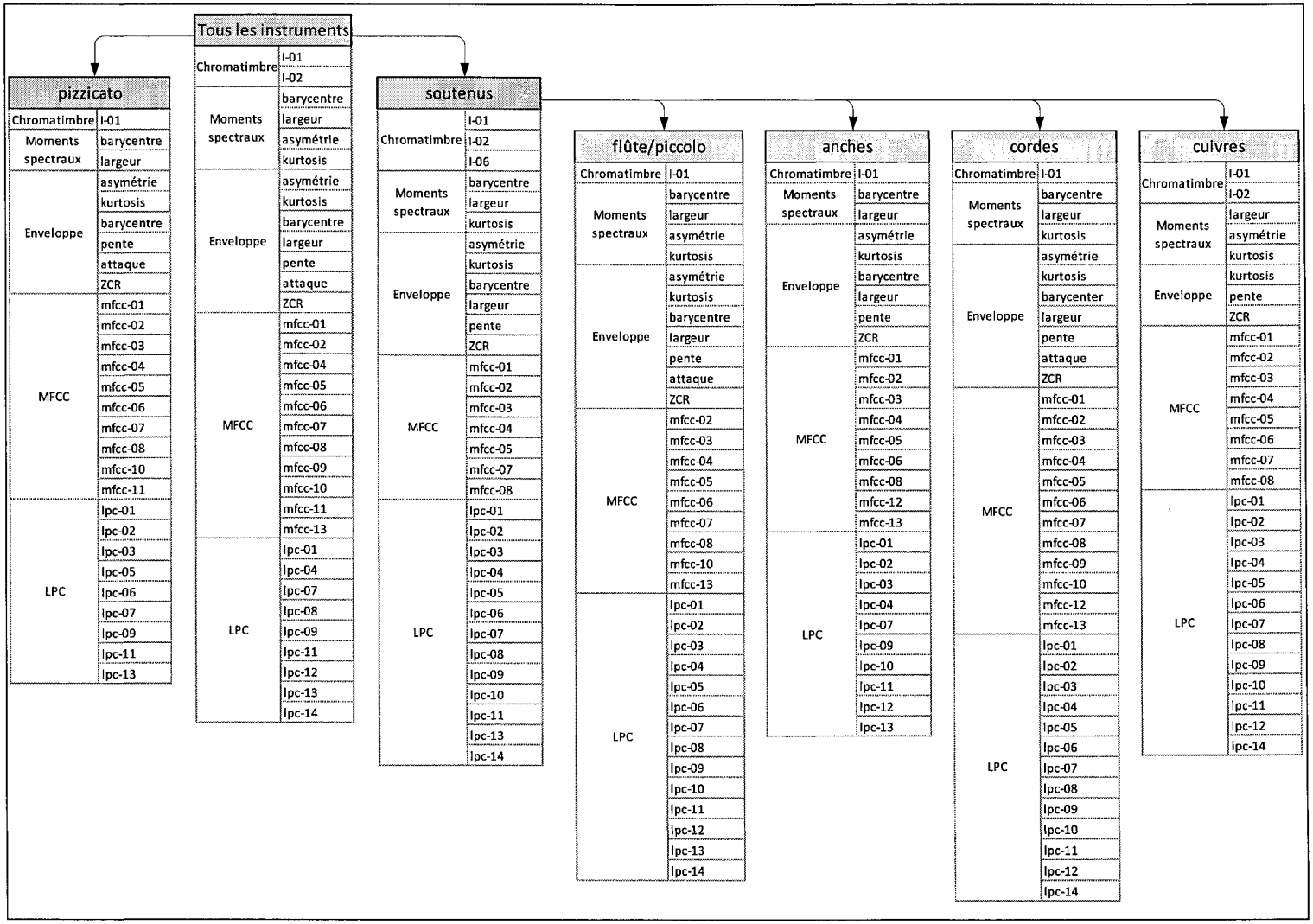

Figure 5.22 Paramètres sélectionnés par sélection séquentielle SBS hiérarchique.
(Classification hiérarchique avec taxonomie naturelle, classificateur k-NN et normalisation mu-sigma).

5.4 Reconnaissance psycho-visuelle du chromatimbre

Les moments invariants du chromatimbre ne permettent pas de représenter distinctement les instruments de musique dans un contexte de reconnaissance automatique. Néanmoins le principe de chromatimbre démontre, à notre avis, une aptitude à faire émerger l'empreinte du timbre de l'instrument. La reconnaissance automatique d'image et l'extraction de points d'intérêts visuels n'étant pas notre domaine d'expertise, une validation psycho-visuelle fut appliquée au jeu de données ; c'est-à-dire que la forme du chromatimbre d'un instrument fut évaluée et classée visuellement par un observateur humain. Cet exercice permit entre autre de se convaincre de la validité de la représentation.

5.4.1 Structure des simulations

Pour éviter d'effectuer une validation croisée manuelle avec un jeu de données imposant, un ensemble de données de tests à été construit au hasard parmi les 6 698 fichiers disponibles. Les données restantes, soit 6 598 fichiers, ont servi de données d'apprentissage et ont été présentées à l'évaluateur comme images de références. Le nombre de données de test fut donc limité à 100 fichiers. Le nombre de fichiers par instrument ainsi obtenu est présenté dans le Tableau 5.4. Pour ne pas que l'évaluateur puisse « compter » les occurrences de chaque instrument et biaiser les résultats, le nombre d'occurrences pour chaque instrument ne fut pas spécifié. Ceci a eu pour effet de ne pas nécessairement inclure tous les instruments disponibles lors de la validation. On peut constater dans le Tableau 5.4 que les instruments suivants ne sont pas disponibles parmi les fichiers de tests : piccolo, flûte à bec, basson et flûte de pan.

Tableau 5.4 Liste du nombre de notes par instrument des tests de l'analyse psycho-visuelle.

Instrument	Nombre de notes	Instrument	Nombre de notes
Accordéon	5	Guitare acoustique	6
Saxophone Alto	4	Banjo	6
Saxophone Baryton	7	Basson	0
Violoncelle	1	Clarinette	1
Cornet	2	Basse électrique	12
Guitare électrique	10	Cor anglais	1
Flûte traversière	2	Cor français	5
Harmonica	3	Mandoline	3
Hautbois	2	Flûte de pan	0
Piccolo	0	Orgue	1
Flûte à bec	0	Saxophone soprano	3
Saxophone Ténor	3	Trombone	2
Trompette	4	Tuba	1
Ukulélé	2	Violon alto	7
Violon	7		
total : 100			

Le chromagramme est obtenu d'une transformée *constant-Q* (CQT) de 96 bins et le chromatimbre y est extrait avec 5 contours. Ces valeurs permettent de lire adéquatement les détails de la forme, puisque trop ou pas assez de contours peuvent alourdir l'analyse de l'observateur. Des fenêtres de type « blackman-harris » entrelacés de 12,5% ont servis à la construction des trames du chromagramme.

5.4.2 Analyse descriptive du chromatimbre

Plusieurs descripteurs visuels peuvent être remarqués à partir de l'apparence du chromatimbre. En particulier l'articulation est parfaitement visible et il n'y a aucune ambigüité à distinguer les instruments soutenus des instruments pizzicato. Le chromatimbre de l'accordéon est spécialement facile à reconnaitre de par ses caractéristiques uniques. Cependant, les formes

du chromatimbre ne sont pas triviales et il serait difficile ici d'énumérer sur papier toutes les caractéristiques que peuvent posséder chacun des instruments. De plus, le chromatimbre peut prendre plusieurs formes pour un seul instrument. Les figures 5.23 à 5.40 présentent quelques traits remarquables utilisés lors de l'évaluation psycho-visuelle et pourrait servir comme point de départ à l'élaboration de descripteurs plus complexes.

Les figures 5.23 à 5.28 montrent un exemple de chromatimbre d'instruments pizzicato. On discerne l'amplitude de l'attaque par le noyau rouge centré à gauche de la forme. On peut aussi se rendre compte de l'effet de l'inharmonicité des fréquences à l'attaque par l'étirement de la forme vers le haut et le bas.

Étant projeté sur le plan mélodique, le chromatimbre permet d'avoir une vue d'ensemble de l'enveloppe et du spectre de la note. On peut facilement y discerner l'étalement spectral, les modulations en amplitude et en fréquence, l'attaque, le maintien et le relâchement de la note. Sur la Figure 5.34, on remarque aisément que la note de violon est jouée avec un effet de trémolo (modulation en fréquence) ; sur la Figure 5.40, on remarque la dynamique de l'intensité de l'orgue.

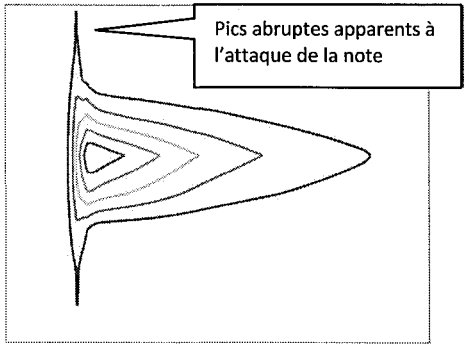

Figure 5.23 Chromatimbre type de la guitare acoustique.

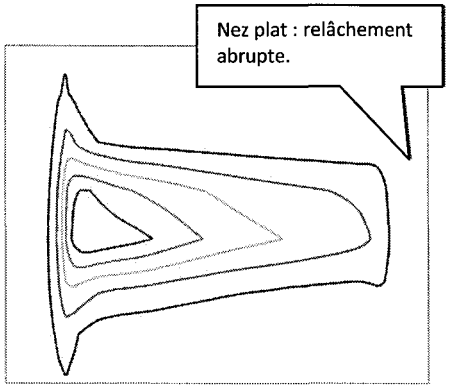

Figure 5.24 Chromatimbre type de la basse électrique.

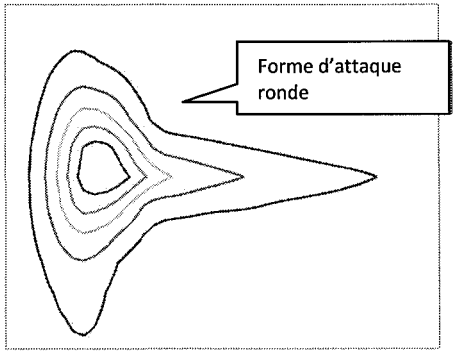

Figure 5.25 Chromatimbre type du banjo.

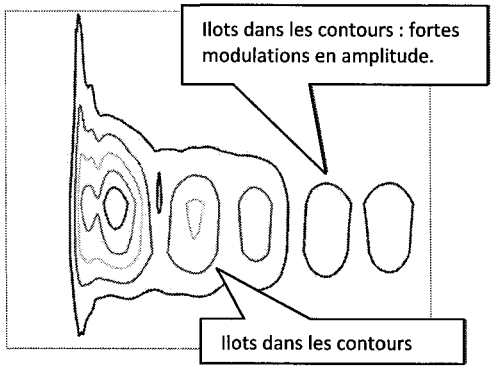

Figure 5.26 Chromatimbre type de la mandoline.

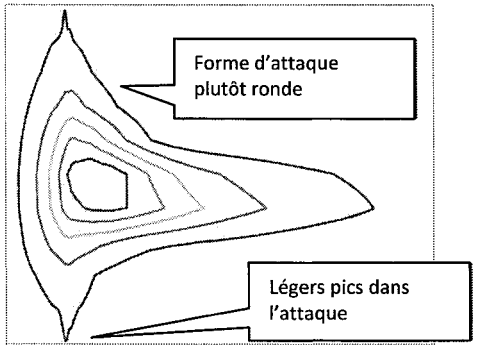

Figure 5.27 Chromatimbre type de l'ukulélé.

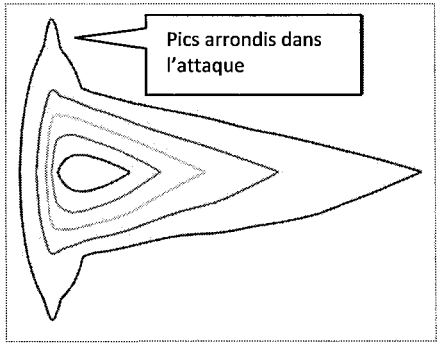

Figure 5.28 Chromatimbre type de la guitare électrique.

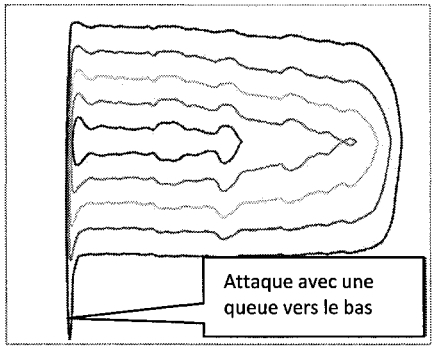

Figure 5.29 Chromatimbre type de l'harmonica.

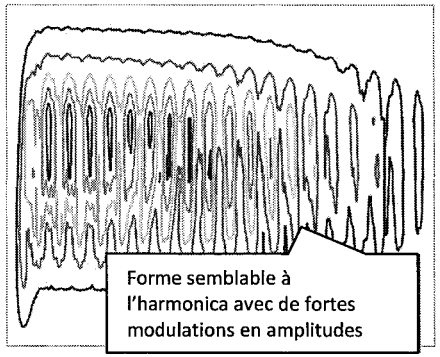

Figure 5.30 Chromatimbre type de l'accordéon.

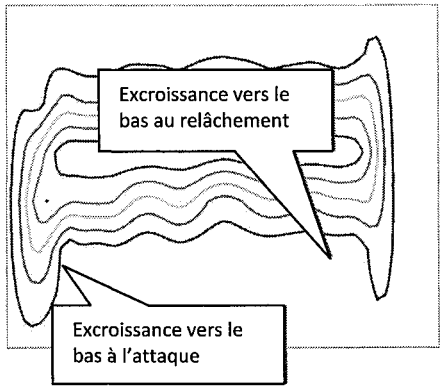

Figure 5.31 Chromatimbre type de la flute à bec.

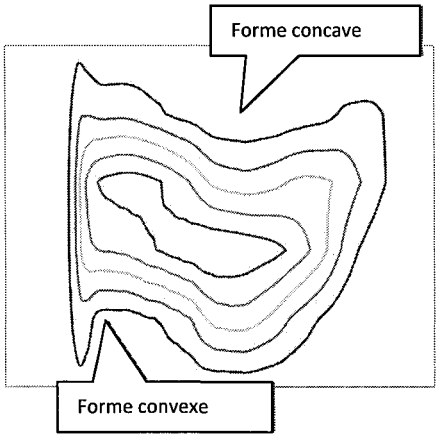

Figure 5.32 Chromatimbre type de la clarinette

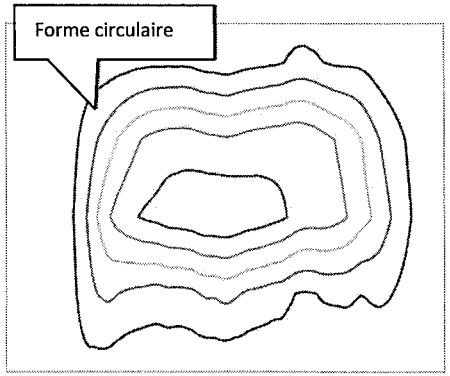

Figure 5.33 Chromatimbre type du tuba.

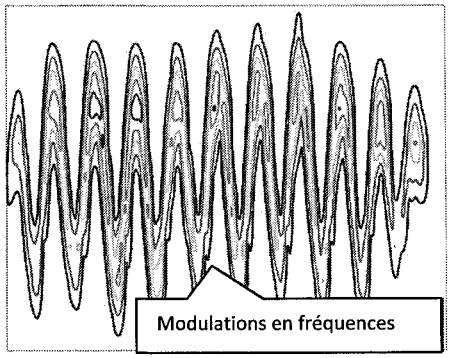

Figure 5.34 Chromatimbre type du violon.

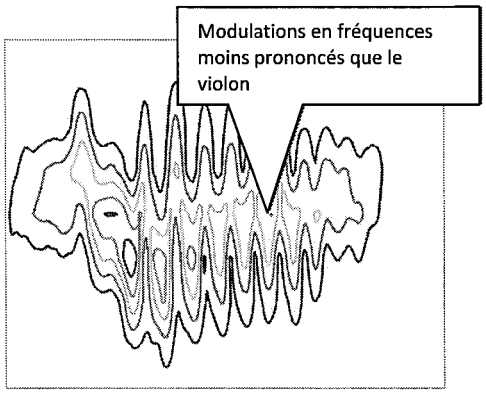

Figure 5.35 Chromatimbre type du violon alto.

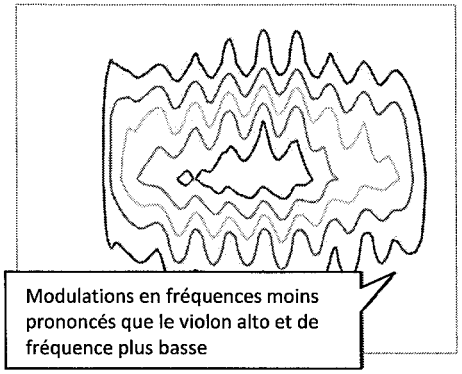

Figure 5.36 Chromatimbre type du violoncelle.

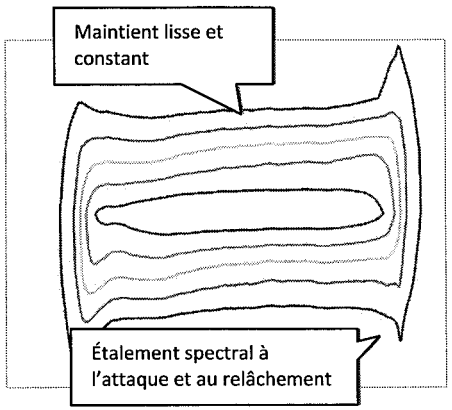

Figure 5.37 Chromatimbre type du saxophone.

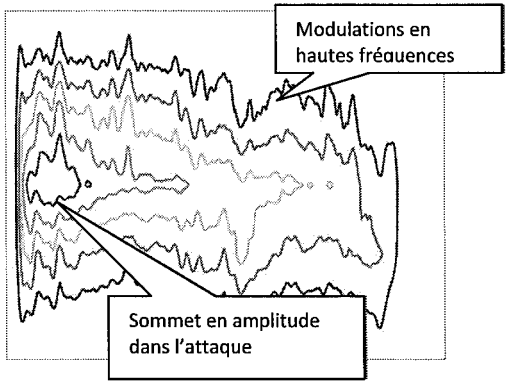

Figure 5.38 Chromatimbre type de la flute traversière.

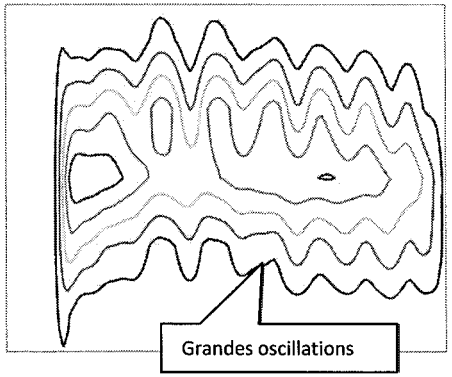

Figure 5.39 Chromatimbre type du cor anglais.

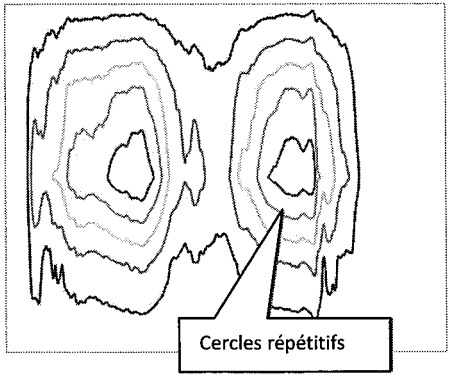

Figure 5.40 Chromatimbre type de l'orgue

5.4.3 Analyse des résultats

Les résultats de l'analyse psycho-visuelle sont très intéressants. La reconnaissance des instruments obtient un taux de reconnaissance de 74% (Figure 5.41). Le taux de reconnaissance de l'articulation est parfait (Figure 5.42) ; de même, celui de la reconnaissance de la famille est excellent avec 89% de succès (Figure 5.43). Le nombre d'instruments de musique considérés influence grandement ce résultat puisque l'évaluateur doit mémoriser beaucoup plus de caractéristiques et manipuler beaucoup plus d'exemples, ce qui n'est pas le cas avec l'articulation dont l'empreinte est trivial. Un processus automatique serait capable de construire des descripteurs amplement plus complexes à l'aide d'un jeu d'exemples plus large.

De plus, un effort devrait être apporté au développement d'une méthode d'extraction des contours pour diminuer les variabilités et rehausser les détails importants. Les coupes du chromagramme, pour tracer les contours, pourraient bénéficier d'une disposition sur une échelle logarithmique ou quelconque permettant ainsi d'aller chercher les zones plus denses en information pertinente. Le chromagramme devrait également jouir d'algorithmes plus performants pour qu'il puisse être calculé avec plus de composantes spectrales.

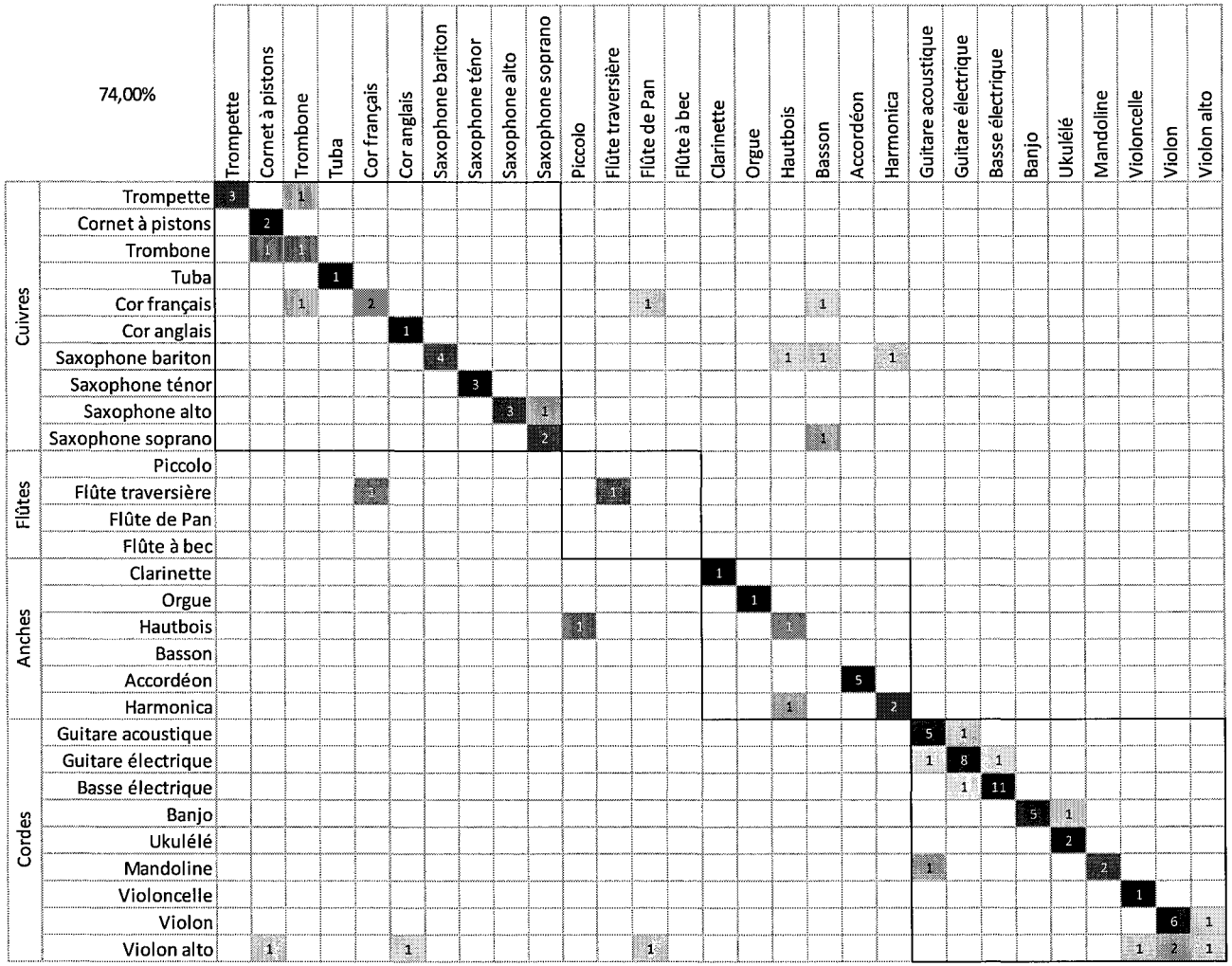

Figure 5.41 Matrice de confusion pour l'analyse psycho-visuelle du chromatimbre.
Les rectangles représentent les différentes familles d'instruments

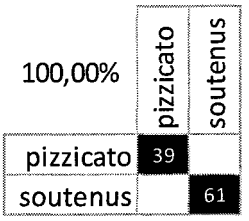

Figure 5.42 Matrice de confusion pour les articulations de l'analyse psycho-visuelle du chromatimbre.

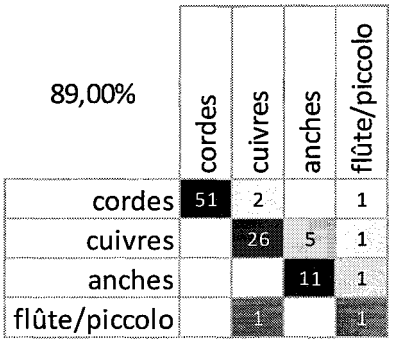

Figure 5.43 Matrice de confusion pour les familles de l'analyse psycho-visuelle du chromatimbre.

5.5 Discussions

5.5.1 Sélection des paramètres des algorithmes

Les paramètres des classificateurs furent déterminés de façon empirique, c'est-à-dire que des simulations effectués avec différentes valeurs de paramètres permirent de sélectionner les valeurs des paramètres donnant les meilleurs taux de reconnaissance. Pour éviter un nombre exagéré d'itérations, certains paramètres furent fixés à priori. Par exemple, le nombre de voisinage optimal du classificateur *k*-NN fut déterminé en utilisant la distance euclidienne. La répercussion de la métrique est inconnue ; une investigation plus poussée devrait permettre de déterminer l'effet de la métrique sur le nombre de voisinage optimal.

Un creux inexpliqué (Figure 5.1) est présent lorsque le nombre de mixtures du classificateur GMM atteint la valeur 5. Ce phénomène n'a pu être analysé en détail et n'est certainement pas anodin.

Les coefficients MFCC ainsi que l'enveloppe temporelle contiennent beaucoup d'information pertinente sur l'empreinte d'un instrument de musique. On sait que les 13 premiers coefficients MFCC sont suffisants. Néanmoins, seuls sept paramètres de l'enveloppe furent extraits et utilisés ; un système de classification tirerait avantage à construire plus de paramètres à partir de l'enveloppe temporelle.

L'algorithme d'extraction de l'enveloppe sélectionnée utilisait un filtre de type « moyenne mobile » pour lisser le spectrogramme. La détermination du noyau du filtre devrait être analysée en profondeur. Des simulations ont également été effectuées avec la combinaison de différents algorithmes d'extraction de l'enveloppe tout en utilisant les mêmes paramètres pour chacun d'eux. Pris ensembles, ils donnent de bons résultats mais ajoutés au reste des descripteurs, aucune amélioration ne fut observée. On doit donc chercher à diversifier les types de paramètres issus de l'enveloppe.

Seuls quatre moments spectraux furent extraits et utilisés. Cependant plusieurs descripteurs issus de la transformée de Fourier sont connus dans la littérature actuelle. Eronen [3] démontre, en autre, que les descripteurs issus de la modulation en amplitude peuvent contenir de l'information sur l'empreinte d'un instrument de musique.

5.5.2 Classification hiérarchique

La classification hiérarchique obtient un meilleur score que la classification directe. Ceci est attribuable essentiellement au fait que les paramètres les plus représentatifs du vecteur d'observation sont sélectionnées à l'aide d'un algorithme, et ce à chaque niveau de la hiérarchie. L'algorithme de sélection qui remporta le meilleur score fut l'algorithme de sélection séquentielle en arrière (SBS). Rappelons qu'un algorithme de sélection séquentielle est un algorithme itératif qui permet de réduire la dimension du vecteur d'observation en sélectionnant un paramètre et en évaluant le taux de reconnaissance lorsque le paramètre est ajouté (SFS) ou bien retiré (SBS) du vecteur d'observation. Un Tel algorithme peut converger vers un résultat sous-optimal (comme en démontre l'exemple de la Figure 4.2); une sélection manuelle devrait être appliquée en début d'itération pour retirer les paramètres qui causent ce problème.

La taxonomie utilisée dans les simulations fut celle proposée par Martin [2] et réutilisée par Eronen [3] ; Essid [4] compara les performances de la classification hiérarchique construite à partir de cette taxonomie dite naturelle à la classification hiérarchique à partir d'une taxonomie construite par nuage de points. Essid en conclu à une meilleure reconnaissance des instruments avec la taxonomie issue d'un nuage de points. Dans le cas de la classification hiérarchique effectuée dans ce mémoire, la taxonomie naturelle apporte de meilleurs résultats qu'avec la taxonomie automatique. Cependant, aucune simulation exhaustive ne permit de déterminer la meilleure métrique à utiliser et quel classificateur s'apparie le mieux avec la taxonomie automatique générée.

Seuls deux classificateurs furent implantés dans les simulations. Malheureusement, le classificateur GMM démontra des performances décevantes. L'utilisation d'un plus grand nombre de classificateur aurait permit de déterminer lesquels s'appliquent le mieux à chaque nœud de la hiérarchie et de les utiliser conséquemment.

5.5.3 Analyse psycho-visuelle du chromatimbre

L'analyse psycho-visuelle du chromatimbre suggère la possibilité d'utiliser le chromatimbre comme empreinte distinctive d'un instrument de musique. Il faut se doter cependant d'une méthode d'extraction des contours plus complexe qui permettrait d'extraire les subtilités du chromagramme tout en inhibant ses variabilités. Un chromatimbre pourrait être construit à partir de plusieurs chromagrammes, chacun ayant un nombre de bins varié. En conclusion, les techniques d'extraction du chromatimbre restent encore à être définies.

Des descripteurs de contours doivent également être établis. Les moments invariants d'image sont efficaces avec l'utilisation d'images dont les seules variabilités sont affines c'est-à-dire pour une rotation, une homothétie et une translation. Le chromatimbre subit une variabilité plus importante et l'information contenue dans le chromatimbre y est beaucoup plus subtile.

CHAPITRE 6

CONCLUSION

Les outils nécessaires à la reconnaissance d'instruments de musique ont été détaillés et assemblés en un système de reconnaissance automatique. Les deux organes fonctionnels du système, l'extraction des descripteurs sonores et la classification des vecteurs d'observation, ont été analysés en profondeur et leurs performances respectives comparées et commentées. Une classification hiérarchique, en plus d'une classification directe, a été conduite pour faire émerger les similitudes entre familles d'instruments. Une nouvelle représentation, le chromatimbre, fut introduite et qualifiée tant visuellement que quantitativement.

Dans les premières analyses, les paramètres des algorithmes furent déterminés de façon empirique. Ceux-ci permirent aux simulations de s'exécuter de façon optimale. Le chromatimbre fut visité de façon particulièrement détaillée, combinant ainsi plusieurs valeurs de paramètres. Dans la seconde partie des analyses, chacun des descripteurs fut analysé avec différentes normalisations pour chacun des classificateurs. À chaque classificateur fut appliqué deux algorithmes de sélection, soient SFS et SBS, et un algorithme de réduction, soit PCA, tout en analysant l'effet de la normalisation pour chacun des cas. L'agrégation des trames fut également analysée pour chaque classificateur. Une classification hiérarchique fut appliquée en troisième partie d'analyse avec sélection SFS et SBS à chaque niveau de la hiérarchie, toujours en analysant

l'effet de la normalisation sur ces algorithmes. Enfin, dans la dernière partie des analyses, une validation psycho-visuelle fut réalisée avec les formes du chromatimbre des instruments démontrant ainsi le pouvoir discriminant de la représentation.

Il en ressort que le système le plus performant est constitué des paramètres des descripteurs obtenues par sélection SBS dans une classification *k*-NN hiérarchique et normalisation mu-sigma. La méthode de classification hiérarchique est prometteuse puisqu'à chaque nœud, de nouveaux paramètres sont sélectionnés et de nouvelles statistiques de normalisation calculées. Un système performant devrait utiliser le plus grand nombre de descripteurs possibles, fussent-ils redondants mais calculés avec différents algorithmes, agrégés avec plusieurs statistiques différentes, différenciés entre trames successives ou extraits avec des valeurs de paramètres différents. Une sélection automatique pourrait conséquemment être appliquée à chaque niveau de la hiérarchie; cela devrait permettre d'attribuer à chaque groupe d'instruments les paramètres les plus représentatifs quant à la signature des instruments par rapport au groupe.

De plus, il serait préférable pour chaque descripteur que la nature de sa densité de probabilité soit étudiée en profondeur de façon à déterminer la meilleure normalisation à appliquer à ses paramètres. Cette étude permettrait également de préférer un classificateur par rapport à un autre suite à la détermination de sa nature statistique et d'utiliser à chaque niveau hiérarchique le classificateur le plus approprié. Dans une conception adaptative à l'extrême, à chaque descripteur sont associés une normalisation, un classificateur, une règle de décision à l'ensemble des classificateurs, et ce à chaque niveau hiérarchique de la classification.

RÉFÉRENCES

[1] E. M. Von Hornbostel and C. Sachs, "Classification of Musical Instruments: Translated from the Original German by Anthony Baines and Klaus P. Wachsmann," *The Galpin Society Journal,* vol. 14, pp. 3-29, 1961.

[2] K. D. Martin, "Sound-Source Recognition: A Theory and Computational Model," PhD Thesis, Department of Electrical Engineering and Computer Science, Massachusetts Institutes of Technology, Cambridge, MA, 1999.

[3] A. Eronen, "Automatic Musical Instrument Recognition," Master Thesis, Department of Information Technology, Tampere University of Technology, Tampere, Finland, 2001.

[4] S. Essid, "Classification automatique des signaux audio-fréquences : reconnaissance des instruments de musique.," Thèse de doctorat, Université Pierre et Marie Curie, France, 2005.

[5] S. Theodoridis and K. Koutroumbas, *Pattern Recognition*, Academic Press ed., 2009.

[6] F. Opolko and J. Wapnick, "McGill University Master Samples," ed. Montréal: McGill University, 1986.

[7] F. Opolko and J. Wapnick, "McGill University Master Samples," ed. Montréal: McGill University, 2006.

[8] T. Eerola and R. Ferrer, "Instrument Library (MUMS) revised," *Music Perception,* vol. 25, pp. 253-255, 2008.

[9] F. Opolko and J. Wapnick. (2010-03-25). *McGill University Master Samples* [Web]Available: http://www.music.mcgill.ca/resources/mums/html/brief_history.htm

[10] M. Goto. RWC (Real World Computing) Music Database [Online]. Available: http://staff.aist.go.jp/m.goto/RWC-MDB/

[11] L. Fritts. Musical Instrument Sample (MIS) [Online]. Available: http://theremin.music.uiowa.edu/

[12] Vienna Symphonic Library [Online]. Available: http://www.vsl.co.at/

[13] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka, "RWC Music Database: Music Genre Database and Musical Instrument Sound Database," in *Proceedings of the 4th International Conference on Music Information Retrieval (ISMIR 2003)*, 2003, pp. 229-230.

[14] K. D. Martin and Y. E. Kim, "Musical instrument identification: A pattern-recognition approach," presented at the 136th meeting of the Acoustical Society of America, 1998.

[15] A. Eronen and A. Klapuri, "Musical instrument recognition using cepstral coefficients and temporal features," *IEEE International Conference on the Acoustics, Speech, and Signal Processing, 2000,* pp. 753-756, 2000.

[16] A. Eronen, "Comparison of features for musical instrument recognition," *IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics,* pp. 19-22, 2001.

[17] G. Agostini, M. Longari, and E. Pollastri, "Musical instrument timbres classification with spectral features," *EURASIP J. Appl. Signal Process.,* vol. 2003, pp. 5-14, 2003.

[18] A. A. Livshin, G. Peeters, and X. Rodet, "Studies and Improvements in Automatic Classification of Musical Sound Samples," in *International Computer Music conference (ICMC)*, 2003, pp. 25-28.

[19] T. Kitahara, M. Goto, and H. G. Okuno, "Musical instrument identification based on F0-dependent multivariate normal distribution," *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'03),* vol. 5, pp. 421-424, 2003.

[20] S. Essid, G. Richard, and B. David, "Musical Instrument Recognition by Pairwise Classification Strategies," *IEEE Transactions on Audio, Speech and Language Processing,* vol. 14, pp. 1401-1412, 2006.

[21] G. De Poli, P. Cosi, and P. Prandoni, "Timbre Characterization with Mel-Cepstrum and Neural Nets," *International Computer Music conference (ICMC),* pp. 42-45, 1994.

[22] G. De Poli and P. Prandoni, "Sonological models for timbre characterization," *Journal of New Music Research,* vol. 26, pp. 170-197, 1997.

[23] B. Feiten and S. Günzel, "Automatic Indexing of a Sound Database Using Self-Organizing Neural Nets," *Computer Music Journal,* vol. 18, pp. 53-65, 1994.

[24] S. McAdams, "Recognition of sound sources and events," in *Thinking in Sound: the Cognitive Psychology of Human Audition*, edited by S. McAdams and E. Bigand Oxford: Oxford University Press, 1993, pp. 146-198.

[25] J. M. Grey, "An Exploration of Musical Timbre," Ph.D. Thesis, Report STAN-M2, Department of Music, Stanford University, Standford, CA, 1975.

[26] J. M. Grey, "Multidimensional perceptual scaling of musical timbres," *The Journal of the Acoustical Society of America,* vol. 61, pp. 1270-1277, 1977.

[27] A. Eronen, "Musical instrument recognition using ICA-based transform of features and discriminatively trained HMMS," *the 7th International Symposium on Signal Processing and its Applications,* pp. 133-136, 2003.

[28] B. Kostek, "Musical instrument classification and duet analysis employing music information retrieval techniques," *Proceedings of the IEEE,* vol. 92, pp. 712-729, 2004.

[29] T. H. Park and P. Cook, "Nearest Centroid Error Clustering for Radial/Elliptical Basis Function Neural Networks in Timbre Classification," *International Computer Music Conference (ICMC),* pp. 833-866, 2005.

[30] "American National Standard psychoacoustical terminology," *ANSI S3.20-1973,* 1973.

[31] B. Kostek, *Perception-Based Data Processing in Acoustics: Applications to Music Information Retrieval and Psychophysiology of Hearing. Studies in Computational Intelligence,* vol. 3: Springer 2009.

[32] S. Handel, "Timbre Perception and Auditory Object Identification," in *Hearing*, edited by B. C. J. Moore, Second Edition ed.: Academic Press, 1995, pp. 425-461.

[33] J. D. Deng, C. Simmermacher, and S. Cranefield, "A Study on Feature Analysis for Musical Instrument Classification," *IEEE Transactions on Systems, Man and Cybernetics,* vol. 38, p. 11, 2008.

[34] B. P. Bogert, M. J. R. Healy, and J. W. Tukey, "The Quefrency Alanysis of Time Series for Echoes: Cepstrum, Pseudo Autocovariance, Cross-Cepstrum and Saphe Cracking," in *Symposium on Time Series Analysis*, New York, 1963, pp. 209-243.

[35] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on acoustics, Speech and Signal Processing,* vol. 28, pp. 357-366, 1980.

[36] L. R. Rabiner and B. H. Juang, *Fundamentals of speech recognition*: PTR Prentice-Hall, inc, 1993.

[37] D. S. G. Pollock, *Handbook of Time-Series Analysis, Signal Processing and Dynamics*: Academic Press, 1999.

[38] B. Milner, "Display and Analysis of Speech," in *Handbook Of Signal Processing In Acoustics*. vol. 1, edited by D. Havelock, S. Kuwano, and M. Vorländer: Springer, 2008, pp. 449-482.

[39] T. Fujishima, "Realtime chord recognition of musical sound: A system using Common Lisp Music," in *International Computer Music Conference (ICMC)*, Beijing, 1999, pp. 464-467.

[40] K. Lee, "Automatic Chord Recognition Using Enhanced Pitch Class Profile," in *International Computer Music Conference (ICMC)*, 2006, pp. 306-313.

[41] J. Flusser, B. Zitova, and T. Suk, *Moments and Moment Invariants in Pattern Recognition*: Wiley Publishing, 2009.

[42] M.-K. Hu, "Visual pattern recognition by moment invariants," *IRE Transactions on Information Theory,* vol. 8, pp. 179-187, 1962.

[43] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," *IEEE Transactions on Speech and Audio Processing,* vol. 3, pp. 72-83, 1995.

[44] S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach (2nd Edition)*: Prentice Hall, 2002.

[45] M. Radovanović, A. Nanopoulos, and M. Ivanović, "Hubs in Space: Popular Nearest Neighbors in High-Dimensional Data," *Journal of Machine Learning Research,* vol. 11, pp. 2487-2531, 2010.

[46] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of Machine Learning Research,* vol. 3, pp. 1157-1182, 2003.

[47] I. Fodor. (2002). *A Survey of Dimension Reduction Techniques* Technical report UCRL-ID-148494, June 2002, Lawrence Livermore National Laboratory, Center for Applied Scientific Computing. Available: https://e-reports-ext.llnl.gov/pdf/240921.pdf

[48] H. Liu and H. Motoda, *Feature Selection for Knowledge Discovery and Data Mining*: Kluwer Academic Publishers, 1998.

[49] P. Refaeilzadeh, L. Tang, and H. Liu, "Cross Validation," in *Encyclopedia of Database Systems*, edited by L. Liu and M. T. Özsu: Springer, 2009, pp. 532-538.

[50] C. Schörkhuber and A. Klapuri, "Constant-Q transform toolbox for music processing," presented at the 7th Sound and Music Computing Conference, Barcelona, Spain, 2010.

[51] O. Lartillot and P. Toiviainen, "A Matlab Toolbox For Musical Feature Extraction From Audio," in *Proc. of the 10th Int. Conference on Digital Audio Effects (DAFx-07)*, Bordeaux, France, 2007.

[52] I. Nabney and C. Bishop. *Netlab Neural Network Software* Available: http://www1.aston.ac.uk/eas/research/groups/ncrg/resources/netlab/

ANNEXE A

LIBRAIRIES UTILISÉES POUR LA PLATEFORME DE TESTS

Les librairies Matlab suivantes ont été utilisées à l'élaboration de la plateforme de simulations :

- Pour la transformée *constant-Q* nécessaire au calcul du chromagramme, la librairie « CQT-Toolbox » [50] fut utilisée.
- Pour l'extraction des descripteurs classiques tels que MFCC et moments spectraux, la librairie « MIRToolbox » fut utilisée [51].
- Pour la classification avec le classificateur *k*-NN, la librairie « Statistics Toolbox » de Matlab fut utilisée.
- Pour la classification avec le classificateur GMM, la librairie « Netlab » [52] fut utilisée.
- Pour la réduction de la dimension avec l'analyse en composante principale (PCA) et la sélection séquentielle en avant et en arrière (SFS et SBS), la librairie « Statistics Toolbox » de Matlab fut utilisée.

ANNEXE B

PSEUDO-CODE DE LA SEGMENTATION DES SOURCES EN NOTES ISOLÉES

```
Pour chacun des fichiers
    S := signal contenu dans le fichier courant
    Segmenter le signal S en n fenêtres de 100 ms
    Tab := tableau de n entiers indiquant seuil atteint(1) ou non(0)
    Initialiser les éléments de Tab à 0
    i := 0
    Pour chaque fenêtre f de S
          i := i+1
          Centrer le signal de f
          Normaliser le signal de f
          Calculer l'énergie locale de f
          Si l'énergie locale >= seuil alors
                Tab(i) := 1
          Fin si
    Fin pour
    deltas := Tab(2:n) - T(1:n-1)
    idxDebut := indices des éléments de deltas = 1
    idxFin := indices des éléments de deltas = -1
    m := taille de idxDebut
    Pour i allant de 1 à m
          Début := idxDebut(i) * taille des fenêtres
          Fin := idxFin(i) * taille des fenêtres
          Sauvegarder la partie du signal S entre Début et Fin
    Fin pour
Fin pour
```