une radio numérique

une radio numérique

Ce document décrit le travail de recherche exécuté durant mon doctorat en convention CIFRE dans l'entreprise RTL, en cotutelle académique avec le département TSI [1] du laboratoire de l'école TELECOM ParisTech. Ce doctorat est né de la nécessité pour RTL de moderniser ses moyens techniques pour demeurer l'un des principaux acteurs du paysage radiophonique français dans le cadre du projet national de numérisation de la radio. Aujourd'hui l'un des derniers médias encore analogiques, la radio prépare actuellement sa transition vers le numérique, dans le sillage de la Télévision Numérique Terrestre.

Pourtant le contexte est très différent. De par sa simplicité technologique et parce qu'elle peut être une occupation auxiliaire, la radio est le compagnon de notre quotidien et trouve sa place dans une multiplicité d'endroits tels que la cuisine, la salle de bain, le salon, dans un baladeur, ou surtout dans la voiture. Ainsi le pari de la radio numérique implique le renouvellement de 160 millions de postes de radio en France, et, contrairement à l'image hertzienne, la qualité de son est suffisamment satisfaisante pour que de nombreux utilisateurs demeurent sceptiques quant à l'intérêt de renouveler leurs postes pour une offre dont l'avantage n'est pas évident.

C'est ainsi que, sous l'impulsion de plusieurs acteurs, parmi lesquels RTL joue un rôle essentiel, la révolution numérique s'accompagne d'une valeur ajoutée. Le protocole de diffusion T-DMB (*Terrestrial Digital Multimedia Broadcasting*, soit Diffusion Multimédia Numérique Terrestre) permet d'adjoindre au flux audio un flux de services multimédias accessibles à partir d'un écran interactif. Afin de ne pas se dénaturer, la radio se doit de demeurer un média n'accaparant pas l'attention de son auditeur ; aussi le service ajouté n'est pas un flux vidéo qui viendrait en outre concurrencer les acteurs très compétitifs du paysage audiovisuel, mais une offre d'informations auxiliaires qui viennent agrémenter l'expérience radiophonique sans jamais s'y substituer.

Ainsi on pourra par exemple y trouver, dans le cas d'une émission musicale, le titre et l'artiste de la chanson diffusée, voire un lien vers un site de vente en ligne ; ou dans le cas d'une interview, une courte présentation textuelle ainsi qu'une photographie de l'invité. De nombreux services non synchronisés peuvent également venir compléter le flux audio, comme des prévisions météorologiques ou la grille des programmes de la station. La figure 1.1 montre un exemple d'affichage

1. Traitement du Signal et de l'Image

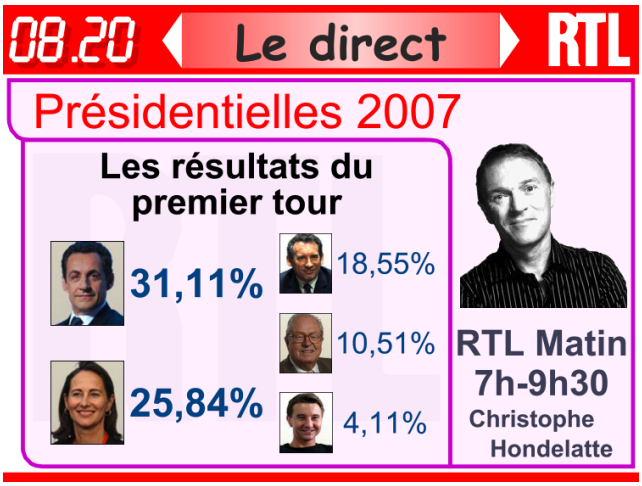


FIGURE 1.1 – Exemple d'affichage interactif accompagnant la radio numérique.

complémentaire pour une soirée électorale, qui permet ainsi à l'auditeur de consulter les résultats à tout moment.

Le projet idéal pour une radio comme RTL consisterait à pouvoir produire ce contenu automatiquement en temps réel ou en ligne [2], à partir du flux audio, ou au moins à réunir le plus possible d'informations pertinentes pour la personne en charge de ce travail. Pourtant, actuellement, la plupart des grandes radio n'ont pas de contrôle en aval sur ce qu'elles émettent. Les logiciels de diffusion exploités sont des applications propriétaires volumineuses et se contentent de réunir les informations sonores voulues, sans fournir d'informations sur ce qu'elles diffusent, qui soient exploitables par un ordinateur. De plus, dans de nombreux cas, comme par exemple une interview impliquant une personnalité et plusieurs journalistes dans un même studio, l'information qui nous intéresserait, à savoir l'identité des locuteurs et la localisation des tours de parole, est totalement inconnue du système de diffusion.

C'est pourquoi RTL, entamant sa mutation numérique, a choisi de se doter des meilleurs atouts en faisant appel aux technologies d'indexation audio, qui substituent à l'indexation manuelle classique l'extraction automatique d'informations (on parle généralement de *méta-données*) à partir du signal audio. Celles-ci ouvrent une autre perspective prometteuse dans la mise en place d'un système d'indexation automatique des archives de la station. En effet, la station conserve depuis 1997 la totalité du flux d'antenne, mais l'annotation manuelle d'un tel volume de données dépasse largement les possibilités d'une entreprise dont le cœur de métier reste avant tout la production d'informations et non l'archivage.

1.2 Applications de l'indexation audio pour la radio

On peut ainsi lister nombre d'applications qui profiteraient directement à un média radio comme RTL :

1. **Reconnaissance des titres musicaux** : l'identification des titres musicaux est un atout essentiel pour une radio puisqu'elle permet de maintenir l'auditeur informé de ce qu'il écoute en lui fournissant les informations de titre, artiste et album, en plus de fournir la pige [3] nécessaire pour les organismes de contrôle de droits d'auteur (SACEM ...). On fait généralement appel pour cela à des techniques d'identification audio qui se concentrent sur la construction d'une empreinte compacte pour chaque titre musical et sa recherche parmi une très vaste collection d'empreintes indexées.

2. Nous faisons la distinction entre le temps réel, qui désigne une réponse quasi instantanée par rapport au contenu d'un flux audio, et le traitement en ligne (*online*) qui correspond à un temps de réponse différé mais régulier et borné par une durée raisonnable (quelques secondes...).

3. Vocable de radio désignant la description détaillée de ce qui est émit par la station.

2. **Recherche de la voix chantée dans un titre musical** : il n'est pas rare que le présentateur d'une émission musicale ou de variété parle sur le début d'une chanson, grapillant ainsi quelques précieuses secondes de parole sur une introduction trop longue, ou assurant simplement la transition entre deux titres. Les présentateurs s'interrompent par convention lorsque l'artiste commence à chanter. À cette fin, il est actuellement nécessaire d'annoter manuellement les chansons afin de permettre au présentateur de savoir précisemment jusqu'à quand il peut parler ou à partir de quand, sur la fin d'une chanson. La détection automatique de voix chantée dans une chanson permettrait ainsi d'automatiser ce processus.
3. **Reconnaissance et suivi de locuteurs** : la reconnaissance de locuteurs permet de fournir à l'auditeur des informations (biographie...) sur le journaliste ou la personne interviewée. Le suivi de locuteur permet de plus d'indiquer en temps réel qui a la parole. L'application d'une telle technique sur les archives permettrait en outre une recherche par locuteur, ce qui se révèlerait un outil très utile pour le travail des journalistes ou pour la constitution des meilleurs moments de certaines émissions (par exemple les fameuses « Grosses têtes »).
4. **Transcription de la parole** : les bulletins d'informations sont généralement fournis aux journalistes à l'antenne sous forme écrite. Le texte est par la suite corrigé manuellement afin de prendre en compte les modifications éventuelles apportées par le présentateur en direct. La transcription automatique permettrait de simplifier ce processus et de le généraliser à l'ensemble des programmes d'antenne. La forme textuelle représente un avantage énorme sur l'archive audio puisqu'elle permet l'application des outils de recherche textuelle beaucoup plus puissants et moins gourmands que l'indexation audio.
5. **Détection de rires, d'applaudissements ou de foule** : les rires et les applaudissements du public ou des invités peuvent être interprétés comme des indices de moments forts de certaines émissions. De même, lors d'une retransmission sportive, la clameur de la foule est généralement révélatrice d'un événement clé du match. La détection de ce type d'événement peut ainsi aider à la constitution du résumé ou des meilleurs moments d'une émission.
6. **Recherche de sons-clés (jingles...)** : la recherche de sons-clés caractéristiques et récurrents, comme les jingles, les habillages sonores ou les publicités, permet de structurer les archives et ainsi de faciliter son exploration.

On remarque que les quatre premières applications énumérées se basent sur une hypothèse forte sur le contenu acoustique analysé. Ainsi les deux premières concernent des plages de musique tandis que les deux suivantes ne s'appliquent que sur des extraits de voix parlée. Le premier outil indispensable à RTL pour l'implémentation de ces traitements plus complexes consiste donc en l'annotation automatique des plages de parole et de musique dans un flux audio. De plus, une fois la musique détectée, la détection de voix chantée constitue une application dont le principe est très similaire. En effet, chacune de ces tâches implique la reconnaissance d'une catégorie acoustique identifiable sans ambiguïté par un être humain.

Les tâches de recherche de sons-clés et de reconnaissance de titre se basent par contre sur un formalisme différent et dépassent le cadre de cette thèse. De même la reconnaissance de locuteur et la transcription automatique sont des sujets de recherche à part entière qui impliquent, l'un la connaissance d'une vaste collection de locuteurs dont la multiplicité a un impact radical sur l'approche suivie, l'autre des notions sur le langage et la sémantique qui dépassent largement le cadre purement audio de cette étude.

Le problème de la classification audio, et particulièrement la classification parole/musique et la détection de chant, constituent donc les sujets couverts par cette thèse, et présentés dans ce document.

1.3 « Qu'est-ce que la musique ? »

Alors que j'expliquais, durant une école d'été, mes travaux de jeune doctorant sur la classification parole/musique à un chercheur expérimenté, celui-ci me posa avec amusement la question suivante, qui me laissa sans réponse :

« Mais qu'est-ce que la musique ? »

En effet définir la musique de manière formelle est problématique. Même si l'on dépasse les querelles sur la musicalité de tel ou tel genre (un éternel débat entre générations), on conviendra que celle-ci est généralement le produit d'un consensus culturel basé sur de nombreuses notions cognitives complexes difficilement formalisables. On trouve les définitions suivantes, respectivement dans le Dictionnaire de l'Académie Française et le Robert :

> Art de composer une mélodie selon une harmonie et un rythme ; théorie, science des sons considérés sous le rapport de la mélodie, de l'harmonie, du rythme.
>
> Art de combiner des sons d'après des règles (variables selon les lieux et les époques), d'organiser une durée avec des éléments sonores ; productions de cet art (sons ou œuvres).

On remarque que dans les deux cas, la musique est caractérisée par son mode de production, à savoir l'acte de composition, qui consiste en un agencement de sons dans le temps. On trouvera pour la parole des définitions qui renvoient au mode de production, ou qui sont même cycliques (« Élément(s) de langage parlé » dans le Robert), liant inévitablement le phénomène sonore à sa source.

Travaillant sur la reconnaissance de ces sources dans un signal audio, je revenais parfois sur cette question, me disant que ne pas y apporter une piste de réponse constituait une lacune. Pourtant j'ai trouvé dans mon incapacité à apporter une réponse formelle la justification de la démarche scientifique employée. Si toute personne est en effet capable d'identifier un son de production musicale ou vocale, c'est bien parce que cette action, comme la plupart des processus cognitifs, échappe à la nécessité d'une définition formelle et repose en réalité sur l'apprentissage empirique de très nombreux exemples associés à une ou plusieurs catégories, qui nous permet de reconnaître celles-ci en présence d'exemples inconnus. Le cerveau est fondamentalement une machine associative, avant d'être une machine logique.

L'apprentissage statistique, qui constitue l'outil prédominant dans le domaine de l'indexation audio, repose précisément sur ce principe, et revient à poser la question plus empirique : « Est-ce de la musique ? »
Cette dernière constitue un problème fondamentalement différent, reposant sur la classification. On peut retrouver dans les deux questions posées la dualité classique entre approches « top-down » et « bottom-up » [4], la première partant d'une définition englobant tous les exemples d'une catégorie et permettant de les reconnaître, la seconde construisant la définition de la catégorie à partir d'une collection d'exemples représentatifs.

Ce que nous appelons « classification audio » consiste en l'application de ce principe de catégorisation d'exemples parmi un ensemble prédéfini de classes, sur un signal audio.

1.4 Classification par Machines à Vecteurs de Support

Le domaine de l'apprentissage statistique est aujourd'hui riche et l'expérimentateur dispose de nombreuses méthodes de classification, généralement formalisées par les statisticiens. Parmi celles-ci, les Machines à Vecteurs de Support (SVM, *Support Vector Machines*) sont une approche récente (datant de la décennie passée) qui modernise le cadre classique de la séparation linéaire en introduisant une non-linéarité dans la surface de décision. La régularité de cette surface de décision est contrôlée par un principe de Minimisation du Risque Structurel qui garantit les bonnes propriétés de généralisation du classifieur. Les excellentes propriétés des SVM nous ont conduit à restreindre notre étude de la classification audio à cette méthode. Une étude préliminaire de l'état de l'art dans le domaine de la classification audio, au chapitre 2, suivie d'une présentation détaillée de la théorie des SVM, dans la partie I, nous permettrons d'étayer notre propos et de justifier ce choix.

L'introduction des SVM dans le domaine de la classification audio est relativement récent (le premier article que nous avons trouvé ne remonte qu'à 2001), et on compte aujourd'hui encore relativement peu d'articles tirant parti de cette méthode pour la tâche en question, par rapport aux autres méthodes plus connues de la communauté. De plus, les SVM restent souvent exploités comme une « boîte noire » de classification que l'expérimentateur n'exploite pas toujours de manière

4. littéralement « du sommet vers le bas » et « du bas vers le haut ».

optimale, en partie en raison des nombreuses *toolbox* publiques lui apportant une interface simple pour employer cette technique sans avoir a en maîtriser les détails théoriques.

Nous verrons que le point central dans la mise en place d'une machine à vecteurs de support est le choix d'une fonction noyau, qui réalise implicitement une transformation sur les données, qui les place dans un espace de dimension supérieure, où la séparation linéaire classique est appliquée. Afin de maximiser la séparabilité des données dans l'espace transformé, la transformation doit donc être directement déterminée par la structure des données dans l'espace d'origine. Un soin particulier doit ainsi être porté à la fois sur le choix de cette transformation et sur la caractérisation des données audio, qui détermine leur répartition dans l'espace d'origine.

Les contraintes propres aux machines à vecteurs de support déterminent donc un certain nombre de problématiques qui constitueront les axes de recherches de cette étude.

1.5 Problématiques

- **Comment employer efficacement les Machines à Vecteurs de Support ?**
  Bien qu'elles réduisent considérablement le nombre de paramètres de réglages par rapport à certaines méthodes classiques comme les réseaux de neurones, les SVM restent fortement dépendantes de l'ajustement de certaines variables, comme le facteur C, fixant le compromis entre régularité et minimisation de l'erreur, ou les paramètres propres au noyau. Le réglage de ces derniers est généralement mené par une procédure de validation croisée dont la complexité devient trop lourde lorsque le nombre de paramètres augmente. Nous étudierons donc les critères permettant d'évaluer de manière fiable et économique les performances d'une SVM par rapport à ses paramètres.
- **Comment appliquer les SVM sur un problème multi-classes ?**
  Cette technique, dérivée de la séparation linéaire, est fondamentalement discriminative. Or nous verrons que le problème de la classification parole/musique, s'il est correctement posé, implique en réalité plus de deux classes. Il nous faut donc déployer une stratégie efficace permettant d'adapter leur usage à ce genre de configuration.
- **Comment caractériser efficacement le signal audio ?**
  Comme la plupart des approches en apprentissage statistique, les SVM se basent sur une modélisation vectorielle des données traitées. Il nous faut donc en premier lieu déterminer les descripteurs qui permettront de décrire au mieux le signal audio, dans l'optique d'une séparabilité maximale entre les classes traitées. En outre, ces données étant groupées et mises en concurrence dans le processus de classification, leur rôle individuel et leurs interactions mutuelles sont difficilement prédictibles pour l'expérimentateur. Il nous faudra donc appliquer des techniques permettant de déterminer automatiquement le sous-ensemble des descripteurs disponibles qui optimise les performances d'un classifieur donné. La prise en compte du noyau, élément central des SVM, sera dans cette opération l'un des enjeux majeurs de cette étude.
- **Comment introduire la donnée temporelle ?**
  Comme nous le verrons par la suite, le calcul des descripteurs résulte d'un découpage du signal audio en trames successives de courte durée. L'approche par « sac de trames » fragmente ainsi le problème en supprimant tout lien ou corrélation entre trames voisines. Nous examinerons donc les techniques de post-traitement qui réintroduisent cette relation temporelle pour augmenter les performances de classification.

Le système développé doit en outre répondre aux contraintes pratiques de l'entreprise RTL. Ainsi le traitement doit être le plus rapide possible, ce qui réduit le champ des possibilités par rapport à une approche purement académique de recherche, et doit pouvoir fonctionner en temps réel, on du moins « en ligne », c'est-à-dire sur un flux audio, avec un retard éventuel mais qui reste contrôlé.

1.6 Résumé des contributions

Afin de répondre aux problématiques exposées dans la section précédente, nous avons apporté durant cette thèse les contributions suivantes.

Nous avons en premier lieu exploité les critères d'alignement du noyau et de séparabilité de classes, que nous présentons dans la section 4.4, pour l'évaluation du noyau dans le contexte de la classification audio. Nous montrerons dans la section 4.6 la pertinence de ces critères en terme de performances et de temps de calcul, par rapport aux autres méthodes plus connues de la communauté. De plus, après avoir montré l'importance du facteur d'erreur C, nous proposerons dans la section 4.5 une procédure d'ajustement de la matrice de Gram pour la prise en compte du facteur C, dans le calcul des mesures d'alignement du noyau et de séparabilité de classes. Nous montrerons dans la section expérimentale 4.6 que cet ajustement améliore sensiblement les résultats de la sélection, pour un coût additionnel minime. En outre, l'inclusion du noyau dans les algorithmes de sélection de descripteurs se révèle équivalente à la sélection de noyau, où la contribution de chaque descripteur constitue un paramètre de ce dernier. Ceci nous conduira donc à proposer, dans la section 7.5, cinq nouvelles méthodes de sélection de descripteurs basées sur les critères sus-mentionnés d'alignement et de séparabilité de classes. Une étude comparative, détaillée dans la section 7.7, viendra confirmer l'efficacité de ces méthodes dans diverses configurations, synthétiques ou réelles.

Après un rapide parcours des paradigmes multi-classes pour les SVM, nous adapterons dans la section 5.1.5.4 le principe des arbres de classification hiérarchique afin d'estimer les probabilités a posteriori par classes. Ces dernières nous permettrons de déployer les techniques de post-traitement. L'examen de diverses configurations hiérarchiques, incluant des taxonomies hybrides basées sur le paradigme *one-vs-one*, fera l'objet d'une étude expérimentale détaillée dans la section 10.3.

Nous proposerons également, dans la section 8.3.2.2, un nouveau paradigme de post-traitement basé sur l'exploitation des probabilités a posteriori estimées comme observation d'un modèle de Markov caché (HMM) dont on estime le chemin optimal. Nous introduirons en outre le modèle moins connu des HSMM (semi-markovien), pour lequel nous proposerons une méthode simple pour la modélisation probabiliste de la durée passée dans un état donné, qui permet de relacher la contrainte de distribution géométrique induite sur cette dernière par le modèle HMM. L'étude sur les post-traitements nous permettra d'introduire dans le chapitre 9 une approche hybride combinant l'approche SVM par trames à un panel de méthodes de segmentation aveugle dont le principe est de détecter automatiquement les frontières entre segments au contenu acoustique homogène. Nous détaillerons ainsi cinq méthodes de segmentation aveugle, dont les plus récentes tirent parti des apports de la théorie des noyaux. Une étude comparative sur les différentes métriques, en section 10.4, montrera l'avantage de l'approche hybride proposée sur les méthodes plus traditionnelles.

Nous aborderons enfin dans la section 10.5 notre participation durant cette thèse à la campagne d'évaluation nationale ESTER 2 qui apporte une comparaison objective aux contributions de l'état de l'art en France sur la classification parole/musique. Le problème de la détection du chant étant moins couvert par la littérature, il est difficile de trouver des corpus publics suffisamment conséquents pour l'évaluation des résultats. Nous avons donc constitué un corpus réunissant des titres libres de droit pour cette tâche, que nous décrirons dans la section 10.1.4, et sur laquelle une expérience comparative est mené pour évaluer notre approche.

On trouvera à la fin de ce document, en page 183, la liste de nos publications.

1.7 Structure du document

Ce document est structuré en trois parties théoriques, suivies d'un quatrième partie expérimentale. La figure 1.2 synthétise la structure en question.

Dans la partie I, nous commencerons par traiter les questions relatives à l'application des Machines à Vecteurs de Support pour la classification. Après avoir présenté en détail la **théorie** et ses implications en terme de **contrôle du Risque Structurel** dans le chapitre 3, nous expliquerons en quoi les SVM constituent une synthèse de nombreuses autres méthodes d'apprentissage, et nous montrerons enfin les avantages qu'implique le principe de maximisation de la marge. Par la suite nous aborderons, dans le chapitre 4, la question de la sélection ou **paramétrisation du noyau**, élément central des SVM, en présentant les différents critères existants dans la littérature, pour mettre l'accent sur le critère d'Alignement, encore très peu utilisé en indexation audio. Nous terminerons cette première partie en comparant dans le chapitre 5 les différentes approches dé-

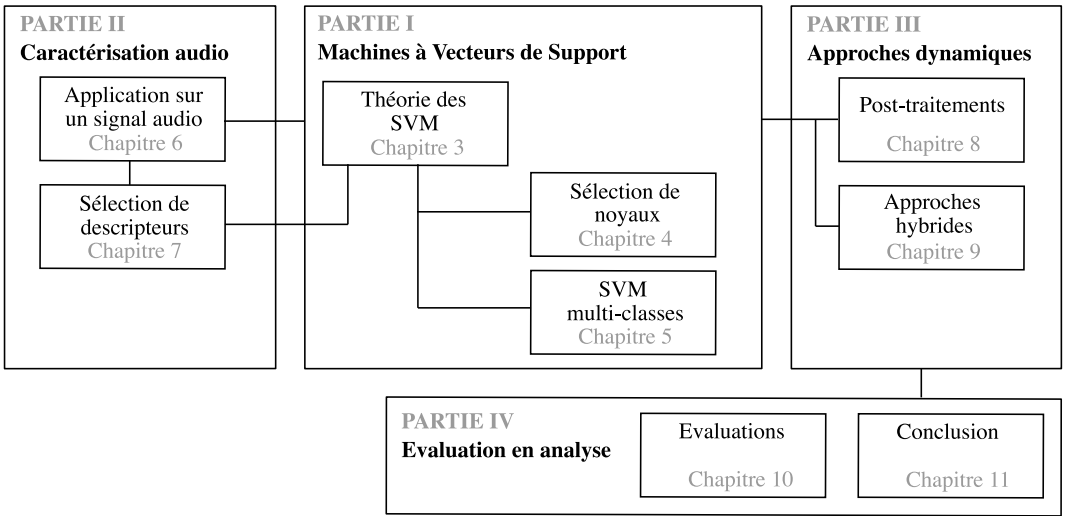


FIGURE 1.2 – Résumé de la structure du document.

diées à l'application des SVM sur un **problème multi-classes**, c'est-à-dire impliquant plus de deux classes. Ceci nous amènera à proposer une stratégie de classification hiérarchique permettant d'estimer les probabilités a posteriori, qui nous seront nécessaires par la suite.

La caractérisation numérique du signal audio sera traitée dans la partie II. Nous commencerons par détailler dans le chapitre 6 le processus de **découpage audio** en trame, puis nous présenterons l'**ensemble de descripteurs** retenus pour leurs propriétés discriminatives sur le problème posé. Cette collection sera complétée par un descripteur proposé dans ce document pour le problème particulier de la classe mixte de parole sur fond musical. Le chapitre 7 traitera des techniques de **sélection automatique de descripteurs**. Après une courte étude sur la notion de pertinence et une proposition de taxonomie des algorithmes, nous présenterons plusieurs approches de la littérature en mettant l'accent sur celles liées aux Machines à Vecteurs de Support. Nous terminerons ce chapitre par la proposition de plusieurs algorithmes exploitant entre autres le critère d'Alignement introduit précédemment.

La dernière partie théorique III examinera les moyens envisagés pour corriger autant que possible les résultats en replaçant les trames dans leur contexte temporel. Ainsi, le chapitre 8 présentera plusieurs **techniques de post-traitement** sur les probabilités a posteriori, allant du simple filtrage à l'application de modèles de Markov, dont nous examinerons une amélioration possible. Cette étude sera complétée dans le chapitre 9 par la proposition d'une **approche hybride** combinant le processus de classification au résultat d'une **segmentation aveugle**, afin de fournir un découpage en segments acoustiquement homogènes. Nous présenterons à cet effet plusieurs algorithmes de segmentation aveugle dont certains sont fortement liés à la théorie des noyaux.

Nous conclurons cette étude dans la partie IV, chapitre 10, par une **évaluation** des différents aspects du système proposé sur plusieurs corpora publics, dont l'un fut créé durant cette thèse, ainsi que dans le cadre de notre participation à la **campagne d'évaluation** nationale ESTER 2. Nous présenterons brièvement ensuite, dans le chapitre 11, l'**implémentation** en C++ du système de classification, que nous avons livré à l'entreprise RTL à la fin de notre thèse. Enfin, le chapitre 12 de **conclusion** apportera quelques commentaires sur ce travail ainsi qu'un aperçu des perspectives ouvertes par ce dernier.

Chapitre 2

État de l'art

Sommaire

2.1 Applications de la classification audio

Le principe de la classification a de nombreuses applications en indexation audio. On trouve ainsi le découpage automatique de données audio pour le parcours structuré d'archives vidéo dans [261], [41] et [204]. La distinction entre parole et musique permet également, dans le domaine du codage audio, d'adopter des algorithmes de codage plus adaptés au contenu et ainsi d'accroître le taux de compression [164] ou d'opérer une allocation intelligente de la bande passante en temps réel [41].

La classification s'applique également sur d'autres classes, en se basant généralement sur la même architecture. On peut ainsi exploiter celle-ci dans le domaine de la Recherche d'Information Musicale (MIR, *Music Information Retrieval*) pour la reconnaissance de genres musicaux [229][68][151][180] ou d'instruments de musique [132][73], ou encore pour l'identification de l'artiste ou du chanteur dans un titre musical [25][124][228] (on distingue les deux dans le cas d'artistes invités pour des duos). Mandel et al. [146] étendent également le domaine d'application de la classification audio à la recherche de titres musicaux par similarité, généralement basée sur d'autres techniques comme le *fingerprint audio* (empreintes audio).

Le signal de parole est également matière à certaines classifications plus approfondies, par exemple pour découper ce dernier en tours de parole successifs [125][155] ou pour la détermination du sexe [102] ou de l'âge [32] du locuteur, qui permettrait d'apporter un complément d'information pour la tâche de reconnaissance de locuteurs. On peut également considérer la reconnaissance de parole comme un exemple de classification audio, bien que celle-ci en dépasse le cadre puisqu'elle fait intervenir des notions linguistiques et sémantiques.

Étendant considérablement le champ acoustique considéré, le domaine de le reconnaissance de scènes auditives [182][194][69] (CASR, *Computational Auditory Scene Recognition*) a pour principe l'identification de l'environnement capté par un enregistrement audio, par exemple la rue, la

Référence	Sil	Par	Tel	Mus	Par+Mus	Ch	Br	Par+Br	Aut
[207],...		X		X					
[85]		X		X					X
[41], [168]		X		X			X		
[52], [202]		X		X		X	X		
[100]		X	X	X					X
[159]		X		X					X
[86]		X		X	X				X
[169]	X	X		X	X			X	
[194]		X		X			X	X	

TABLE 2.1 – Taxonomies de classes exploitées dans la littérature pour la classification parole/musique. Par : parole – Tel : parole au téléphone – Mus : musique – Ch : chant – Br : bruit – Aut : autre – A+B : les deux classes superposées.

nature, un café, l'intérieur d'une voiture, une bibliothèque ou encore une église. Le problème posé est beaucoup plus complexe et les classes moins clairement définies, mais sa résolution, au moins partielle, aurait pléthore d'applications concrètes.

La détection du chant est le plus souvent destinée à mettre en évidence les zones à analyser pour la reconnaissance de chanteurs ou d'artistes, mentionnée plus haut. Elle sert également à d'autres applications comme la reconnaissance de la langue chantée [141], la transcription d'une mélodie [193] ou sa requête dans une base de données [131], ou encore la transcription textuelle ou la synchronisation de paroles (par rapport au texte) [135][141]. Le lecteur intéressé trouvera dans l'étude de Rocamora [199] une liste assez complète des applications possibles de la détection de chant.

2.2 Taxonomie audio

Les exemples précédents montrent un large éventail de possibilités dans le choix des classes employées. Une attention particulière doit cependant être portée à la définition de classes pour que le problème soit bien posé. Burred et Lerch [41] distinguent deux défauts courants dans ce qu'ils nomment les taxonomies audio : la *non-complétude*, qui désigne l'absence flagrante d'une classe implicite importante, par exemple l'absence d'une classe de musique classique dans un problème de reconnaissance de genres musicaux, et l'*inconsistance*, qui désigne une définition trop ambigüe des classes ou leur mauvaise partition, par exemple en présence d'une classe de musique classique et d'une autre d'opéra. Ce second défaut, plus courant dans la littérature, souligne en général l'absence de consensus sur les classes considérées dans certains domaines. Ainsi la reconnaissance de genre se heurte généralement à l'impossibilité de trouver un consensus sur une taxonomie cohérente entre les différents sous-genres musicaux [229], de même pour la reconnaisance de scènes auditives, mentionnée précédemment.

Un problème méthodologique consiste par ailleurs à distinguer des classes définies non pas par un phénomène acoustique identifiable, mais par une notion sémantique qui n'a pas de sens d'un point de vue auditif. Par exemple la distinction de la publicité [170] ou des jingles (par rapport à la musique) [179] dépasse le cadre de la classification audio puisque ces deux classes ne sont pas définies par leur contenu acoustique mais par le sens que leur accorde l'auditeur. De la même manière, la mise en concurrence de classes définies sur des niveaux incompatibles ou se chevauchant (par exemple la publicité et la violence physique [170] peuvent décrire un même signal audio) constitue généralement un obstacle pour le système de classification.

De manière générale on considérera qu'une taxonomie audio est bien définie si elle forme une partition de classes disjointes sur l'ensemble des phénomènes audio couverts par la classification.

Sur le problème le plus basique de distinction entre parole et musique, on constate déjà de nombreux points de divergences entre les taxonomies employées dans la littérature, que nous comparons dans le tableau 2.1, indiquant les classes exploitées dans quelques article.

Les points de suspension sur la première ligne indiquent que l'approche à deux classes (parole

et musique pures) est de loin la plus généralement suivie dans la littérature (on s'en convaincra par le nombre de références : [205],[42],[68],[91],[247],[121],...). Il est cependant naturel que d'autres classes interviennent dans le processus, ces deux seules ne suffisant pas à décrire un signal audio de manière exhaustive. Une solution simple consiste parfois à introduire une classe complémentaire « Autres » [85][100] qui permet d'y ranger tout ce qui ne correspond pas aux autres classes. Mais une telle classe est généralement très mal définie puisqu'elle fait cohabiter des phénomènes acoustiques très différents qu'un classifieur peinera à caractériser dans leur ensemble. On évitera donc en pratique cette solution trop simple.

Certaines classes sont parfois bien définies mais n'apportent pas grand chose au processus de classification, en raison de leur détection très aisée. Ainsi la voix téléphonique, prise en compte dans [100], est caractérisée par un filtrage passe-bande très net ; de même le silence [169] est très simple à localiser et est généralement détecté dans une phase préliminaire à la classification. De manière générale, on parle de *détection* lorsque la classification n'implique que la reconnaissance d'une classe.

On remarque également la présence des classes mixtes parole+musique et parole+bruit dans un grand nombre de publications [86][169][194]. Les approches classiques (parole et musique pures) sont en effet parcellaires puisqu'elles font l'impasse sur l'éventualité de la présence simultanée de parole et de musique dans un même extrait sonore. Cette situation est pourtant très courante, à la radio par exemple, où les titres des bulletins d'informations sont généralement accompagnés d'un fond musical destiné à agrémenter le discours d'une certaine tension, ou par exemple dans le cas d'émissions de variété où le fond musical apporte au contraire une ambiance à l'antenne.

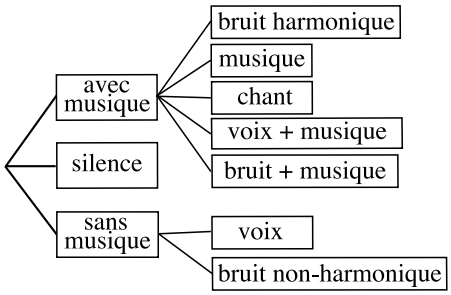
Certains auteurs contournent le problème des classes mixtes en traitant chaque classe par un problème de détection indépendant [150][261]. Ainsi la reconnaissance de parole sur fond musical sera menée implicitement par les détections conjointes de parole et de musique. Si l'approche a le mérite d'être simple et de limiter le nombre de classes, elle est cependant pénalisée par la constitution de classes fortement hétérogènes d'un point de vue acoustique, et donc difficilement caractérisables.

Une autre approche couramment exploitée pour prendre en compte la multiplicité des classes mises en jeu consiste à suivre un arbre hiérarchique de classifications successives permettant d'affiner itérativement la détermination des classes présentes. La figure 2.1 montre plusieurs exemples de taxonomies hiérarchiques employées dans la littérature. La définition de l'arbre est en général empirique et suit une logique intuitive, par exemple détecter dans un premier temps la présence de parole pour affiner par la suite la caractérisation des régions où la parole est absente [5][140][139] (exemples d et e), ou au contraire commencer par détecter la présence de musique [261] (exemple a). On trouve également des graphes de décision plus complexes ne constituant pas des arbres [112]. Certains auteurs, en présence d'un ensemble plus complexe de classes, préfèrent baser la construction de l'arbre sur des critères de séparabilité des classes, comme Essid [73] qui, pour un problème de reconnaissance d'instruments de musique, regroupe itérativement les classes par *clustering* (regroupement) hiérarchique, appliquant ainsi une stratégie *bottom-up*.

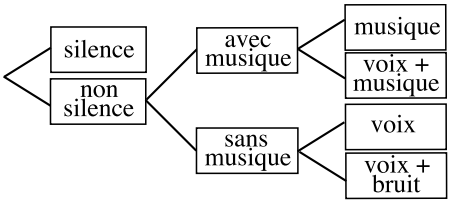
Les problèmes de classification étant posés, nous poursuivons cet état de l'art par un examen des contributions dans ce domaine. Une énorme majorité des publications concentrent leurs efforts sur l'un des deux axes suivants : l'exploitation d'un algorithme de classification original ou efficace, ou bien la proposition de descripteurs destinés à caractériser au mieux les classes pour la tâche en question. Les sections suivantes détaillent les principales propositions pour chacun de ces deux aspects.

2.3 Techniques de classification

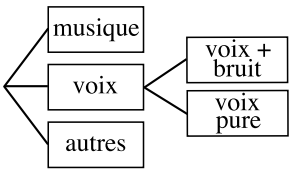
Nous avons évoqué dans la section précédente l'existence de méthodes dites *discriminatives*. Les techniques exploitées en apprentissage statistique se distinguent en effet parmi deux modalités



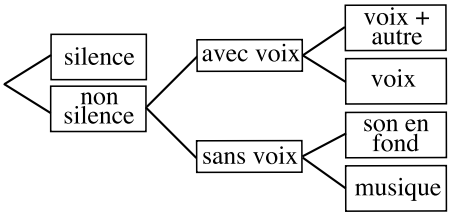
(a) Zhang & Kuo [261]



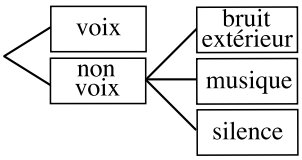
(b) Nitanda et al. [169]



(c) Gauvain et al. [85]



(d) Lu, Li & Zhang [140]



(e) Lu, Jiang & Zhang [139]

FIGURE 2.1 – Quelques exemples de taxonomies hiérarchiques de la littérature sur le problème de la classification parole/musique.

classiques :

- Les **méthodes génératives** ont pour principe de modéliser la distribution des exemples de chaque classe, et ainsi de « comprendre » implicitement la nature des différentes classes dans l'espace de description.
- Les **méthodes discriminatives** portent uniquement l'attention sur la détermination de la frontière séparant les exemples de deux classes.

Les deux approches ont leurs avantages respectifs. Les méthodes génératives apportent une meilleure compréhension de la distribution des classes et permettent de prendre en compte un nombre élevé de classes. En revanche les méthodes discriminatives simplifient généralement le problème en le limitant à la détermination d'une frontière (dont la caractérisation est nécessairement plus compacte que celle d'une distribution), mais se restreignent pour cela à deux classes ; l'application sur plus de deux classes se fera alors par une combinaison de discriminateurs, comme nous l'avons évoqué pour les arbres hiérarchiques de classification. La figure 2.2 illustre le principe des deux méthodes sur un exemple simple à 3 classes.

2.3.1 Méthodes génératives

Les méthodes génératives sont les plus couramment employées dans la littérature, en raison de l'héritage historique des techniques de traitement de la parole. La théorie de la décision de Bayes apporte aux modèles évalués le complément nécessaire pour la classification. Ainsi, si l'on suppose les exemples de chaque classe générés par un modèle aléatoire de densité de probabilité $p(\boldsymbol{x}|\omega_c)$, où ω_c représente la classe d'indice c, la formule de Bayes nous permet de déterminer la probabilité

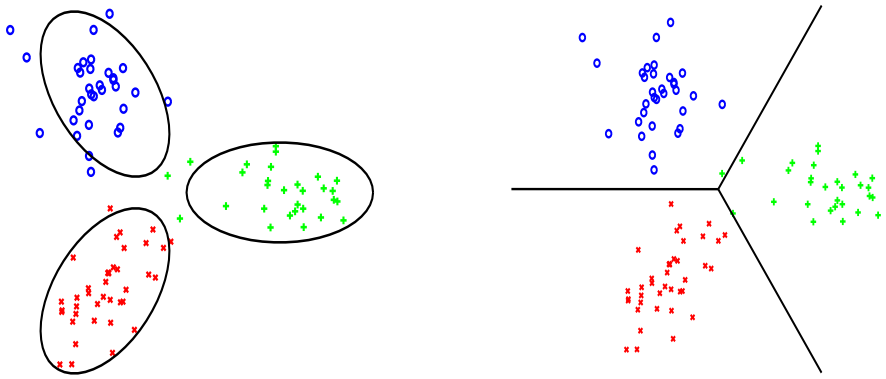


FIGURE 2.2 – Illustration des principes génératifs et discriminatifs (respectivement à gauche et à droite) sur un exemple simple à 3 classes.

a posteriori d'une classe sous l'observation d'un échantillon donné :

$$p(\omega_c|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\omega_c)P(\omega_c)}{P(\boldsymbol{x})}.$$

Le cadre habituel des méthodes génératives consiste à appliquer la stratégie du *Maximum A Posteriori* (MAP), qui associe à l'exemple $\boldsymbol{x}$ la classe $\hat{c}$ maximisant la probabilité a posteriori (la probabilité $P(\boldsymbol{x})$ n'intervient pas dans le choix puisqu'elle est constante au regard de la variable c) :

$$\begin{aligned} \hat{c} &= \arg\max_{1\leq c\leq C} p(\omega_c|\boldsymbol{x}) \\ &= \arg\max_{1\leq c\leq C} p(\boldsymbol{x}|\omega_c)P(\omega_c). \end{aligned}$$

Les probabilités a priori $P(\omega_c)$ sont en général supposées uniformes ou bien estimées à partir de la distribution des exemples du corpus d'apprentissage. Le principe des méthodes génératives consiste ainsi à estimer les densités de probabilités $p(\boldsymbol{x}|\omega_c)$ par des modèles statistiques.

Le **modèle gaussien multi-dimensionnel** caractérise la distribution par sa moyenne $\boldsymbol{\mu}_c$ et sa matrice de covariance $\boldsymbol{\Sigma}_c$, dont l'estimation à partir des exemples du corpus est immédiate. La distribution gaussienne est définie par :

$$\mathcal{N}(\boldsymbol{x}|\omega_c) = \frac{1}{(2\pi)^{d/2}\left|\boldsymbol{\Sigma}_c\right|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_c)^T\boldsymbol{\Sigma}_c(\boldsymbol{x}-\boldsymbol{\mu}_c)\right). \tag{2.1}$$

Celui-ci est par exemple exploité par Scheirer et Slaney [207], ou encore par Saunders [205], sur le problème de classification parole/musique.

Le modèle gaussien est néanmoins généralement trop restrictif et ne permet pas de modéliser la plupart des distributions réelles. On peut montrer, cependant, que toute distribution régulière est asymptotiquement modélisable par une somme de gaussiennes pondérées (par asymptotiquement, on entend : lorsque le nombre de gaussiennes tend vers l'infini), que l'on appelle communément **Modèle de Mélange de Gaussiennes** (GMM, *Gaussian Mixture Model*). On estime donc dans le cadre du modèle GMM la distribution par la somme suivante de M composantes :

$$\hat{p}(\boldsymbol{x}|\omega_c) = \sum_{i=1}^{M} m_i\,\mathcal{N}\left(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\right),$$

où chaque composante d'indice i est définie de manière similaire à l'équation 2.1 et caractérisée par la moyenne $\boldsymbol{\mu}_i$, la matrice de covariance $\boldsymbol{\Sigma}_i$ et le coefficient de pondération m_i. Ces paramètres sont estimés au moyen de l'algorithme Espérance-Maximisation [62][263] (EM, *Expectation Maximization*), guidé par la maximisation de la vraisemblance du modèle par rapport aux exemples.

Le nombre de composantes M reste sujet à une détermination manuelle de l'expérimentateur, et peut se révéler crucial pour la pertinence du modèle puisqu'il constitue un compromis entre la précision et la complexité. De plus, un nombre trop élevé de composantes peut impliquer un sur-apprentissage du classifieur et ainsi pénaliser ses capacités de généralisation sur des exemples inconnus. Le modèle GMM est l'un des modèles les plus largement employés dans la littérature [52][180][100][207][42].

Bien que cette technique soit à priori tout à fait indépendante des GMM, les **Modèles de Markov Cachés** (HMM, *Hidden Markov Models*) sont généralement couplés à ces derniers. Les HMM [190], que nous exploiterons et présenterons en détail dans la section 8.3 de la partie III, modélisent l'évolution temporelle d'un système par une séquence d'états tirés parmi un ensemble fini, où à chaque itération une observation est produite, dont la distribution est classiquement décrite par un modèle GMM. La combinaison HMM/GMM est très populaire dans la communauté pour sa simplicité d'implémentation et son interprétation aisée. On trouve ainsi de nombreux exemples de l'exploitation de ce dernier pour la classification audio [125][9][55][259][109].

2.3.2 Méthodes discriminatives

Le principe général des méthodes discriminatives est la détermination d'une frontière de séparation optimale entre deux classes. La décision sur un exemple se fait en évaluant de quel côté de la frontière ce dernier se situe. Si l'éventail des frontières possibles est infini, nous verrons que, comme le modèle GMM, celles-ci sont avant tout contraintes par une condition de régularité qui influence directement la capacité de généralisation du discriminateur.

La méthode discriminative la plus sommaire consiste à appliquer une **heuristique** consistant en une combinaison logique de seuils sur les descripteurs, empiriquement déterminés à partir des données d'apprentissage. Bien que très basique, et souvent implicitement couverte par des méthodes automatiques plus complexes, cette approche demeure relativement populaire dans de nombreux domaines, y compris la classification audio [139][261]. Elle est souvent employée pour montrer la pertinence d'un nouveau descripteur fortement discriminant pour une tâche donnée [176][111], en particulier dans le domaine de la détection de chant [196][143].

Il est possible de rationaliser l'application d'heuristiques par seuillages successifs en suivant un **Arbre de Classification** (ou **CART**, *Classification And Regression Trees*) comme dans [242].

C'est historiquement le modèle le plus simple d'un hyperplan de séparation linéaire qui a ouvert la voie dans ce domaine. Ainsi l'**Analyse Discriminante Linéaire** (LDA, *Linear Discriminant Analysis*), que nous présenterons dans la section 3.2, est l'une des premières méthodes d'apprentissage automatique, qui consiste en la détermination d'un vecteur $\boldsymbol{w}$ normal définissant l'hyperplan de séparation optimale pour la fonction de décision suivante :

$$f(\boldsymbol{x}) = \boldsymbol{w}^T \boldsymbol{x} + b. \tag{2.2}$$

Malgré son principe discriminatif, on classe généralement la LDA dans les méthodes génératives, car celle-ci implique une modélisation gaussienne des distributions de classes. Néanmoins, parce qu'elle est le dénominateur commun de la plupart des méthodes discriminatives ultérieures, nous préférons l'introduire dans cette section. La LDA reste encore aujourd'hui exploitée dans le domaine de la classification audio [91][71][13], pour sa simplicité, généralement dans des travaux où l'accent est porté avant tout sur la caractérisation du signal audio, et non sur la phase de classification.

Afin de relâcher l'hypothèse de séparabilité linéaire, l'**Analyse Discriminante Quadratique** (QDA, *Quadratic Discriminant Analysis*), employée dans [68] et [151], permet d'étendre le champ des surfaces de séparation à l'ensemble des sections coniques, par la recherche d'un fonction de décision de la forme $f(\boldsymbol{x}) = \boldsymbol{x}^T \boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}^T \boldsymbol{x} + c$, au prix, bien sûr, d'un apprentissage plus complexe.

L'algorithme du **Plus Proche Voisin** (NN, *Nearest Neighbor*) est un exemple d'approche discriminante d'une simplicité extrême puisqu'il consiste à assigner à un exemple de test la classe associée à l'exemple le plus proche parmi le corpus d'apprentissage. La facilité d'implémentation et l'efficacité de cet algorithme lui a offert une certaine popularité dans les dernières décennies et on

trouve plusieurs exemples de son usage sur le problème de classification audio [207][68]. Son extension aux **k Plus Proches Voisins** (kNN, *k Nearest Neighbors*) permet de renforcer l'algorithme en présence d'exemples d'apprentissage marginaux en basant la décision sur un vote majoritaire parmi les k exemples les plus proches dans l'ensemble d'apprentissage. Cette version plus robuste du plus proche voisin est beaucoup plus diffusée dans la communauté scientifique et encore assez exploitée dans notre domaine [139][202][182][13]. Toutefois, les kNN tombent aujourd'hui en désuétude, principalement parce que la recherche des exemples les plus proches se heurte à la fameuse *malédiction de la dimensionalité*, et peut ainsi impliquer, dans des espaces à grande dimension, une recherche exhaustive parmi les exemples de la base, ce qui se traduit par un coût en mémoire et en temps de calcul prohibitifs par rapport aux méthodes plus récentes.

Le **perceptron**, qui est à l'origine une émulation artificielle du comportement d'un neurone proposée par Rosenblatt [200], est en fait strictement équivalent au modèle de la LDA (équation 2.2). La contrainte de linéarité a considérablement limité le développement du perceptron durant de nombreuses années. Dans les années 70 cependant, les techniques neuronales ont connu un regain d'intérêt grâce à l'introduction du **Perceptron Multi-Couches** (MLP, *Multi-Layer Perceptron*) qui introduit la non-linéarité par le biais d'une structure de couches de perceptrons mutuellement alimentées, ce qui explique l'autre nom couramment employé pour cette technique : les **Réseaux de Neurones Artificiels** (ANN, *Artificial Neural Networks*). Cette technique a rencontré un grand engouement et l'on trouve beaucoup d'exemples de son application sur le problème de la classification audio [106][202][121][204]. Cependant l'apprentissage des réseaux de neurones est une procédure difficile qui nécessite généralement une supervision manuelle pour garantir la convergence. De plus ceux-ci constituent une sorte de boîte noire qui n'offre pas ou peu d'interprétations sur les données exploitées.

Les **Machines à Vecteurs de Support** (SVM, *Support Vector Machines*), que nous avons évoquées en introduction, sont actuellement plus populaires que les réseaux de neurones (bien que le débat soit encore vif entre les tenants des deux approches) et ont en outre montré leur équivalence implicite à certaines structures de réseaux. Pourtant on compte aujourd'hui encore peu d'articles [140][48][96][133] tirant parti de cette méthode pour la classification audio, par rapport aux autres méthodes plus connues de la communauté.

On trouve également quelques exemples dans la littérature d'application pour la tâche de classification audio [43][194] du méta-algorithme **AdaBoost** [79][80], qui renforce un discriminateur faible (tel l'algorithme C4.5) en en combinant de multiples instances, et permet ainsi de complexifier la surface de décision, tout en évitant le problème du sur-apprentissage. On montre par ailleurs [206] que celui-ci induit implicitement un principe de maximisation de marge et présente ainsi de fortes similarités théoriques avec les Machines à Vecteurs de Support, notamment concernant les bornes sur l'erreur de généralisation.

2.3.3 Approches hybrides

Sans trop en détailler le contenu, nous noterons que la plupart des propositions visant à améliorer le processus de classification pour le problème parole/musique se basent sur des approches hybrides combinant certains des algorithmes présentés précédemment. Ainsi, Ellis et Williams [247] et Ajmera et al. [9] proposent un algorithme appliquant un modèle HMM sur les sorties d'un réseau de neurones appris pour identifier les phonèmes de parole. Goodwin et Laroche [91] introduisent le facteur temporel dans une approche par LDA en y adjoignant une procédure de programmation dynamique.

Certains auteurs proposent également des solutions pour fusionner les résultats de divers classificateurs. Ainsi, outre les systèmes multi-experts classiques [102][5], on trouve des paradigmes de fusion basés sur la théorie de l'Evidence (qui se substitue à la théorie des probabilités) [150], sur les réseaux bayésiens [87] ou encore sur des combinaisons de modèles gaussiens [102].

On trouve également plusieurs propositions d'approches **hybrides SVM/GMM** censées tirer parti des avantages des deux approches (discriminative et générative), comme les supervecteurs proposés dans [32] pour la reconnaissance du sexe et du genre du locuteur, ou la combinaison de

Milgram et al. [159].

2.3.4 Discussion

On constate, au regard de cet état de l'art, que la majorité des méthodes utilisées pour la classification audio sont de nature discriminative. Ceci s'explique en partie par le fait qu'elles répondent au principe fondamental énoncé par Vapnik [236], qui se résume à ne jamais traiter un problème par la résolution d'un problème plus général, et donc plus complexe. Ainsi, tandis que les méthodes génératives emploient une grande partie de l'effort d'apprentissage dans la modélisation d'une distribution sur l'ensemble de son support, les méthodes discriminatives se limitent à la seule caractérisation de la région délimitant les classes. Le but de la classification étant en définitive d'associer une classe à chaque exemple, on comprend que la modélisation générative des classes se traduit par une recherche d'informations non-pertinentes qui implique soit un coût supplémentaire inutile, soit une pénalisation des performances à coût égal. Ce constat oriente donc notre choix vers l'usage d'une méthode discriminative.

Nous avons en outre évoqué la *malédiction de la dimension* (ou *curse of dimensionality*). Ce phénomène, décrit pour la première fois par Bellman [20], constitue l'un des problèmes majeurs en apprentissage statistique. En effet, le « volume » d'un espace augmente exponentiellement avec sa dimension, si bien qu'un espace à grande dimension peuplé par un nombre fini d'exemples peut être considéré comme quasiment vide [104], et donc difficilement caractérisable. De la répartition éparse des exemples dans l'espace résulte donc généralement un *sur-apprentissage* qui réduit toute capacité de généralisation de l'algorithme sur des exemples inconnus. Les modèles à mélanges de gaussiennes et la technique des k plus proches voisins sont tous deux sujets à ce phénomène, le premier parce que l'augmentation de la dimension oblige à accoître le nombre de gaussiennes pour caractériser correctement les distributions, le second à cause de la structure éparse des exemples dans l'espace et la complexité implicite de la métrique. Nous verrons dans la présentation des Machines à Vecteurs de Support (partie I) que celles-ci se distinguent entre autres par leur moindre sensibilité à la dimension de l'espace des descripteurs.

2.3.5 Un mot sur la détection de chant

Nous ne détaillons pas ici les algorithmes de classification employés dans la littérature pour la tâche de détection de chant parce que ceux-ci sont globalement les mêmes que ceux employés pour la classification parole/musique, à savoir les modèles à mélanges de gaussiennes (GMM) [52][228][135][105], les modèles de Markov cachés (HMM) [24][172][18], les réseaux de neurones [25][237] et les Machines à Vecteurs de Support (SVM) [131][144][237] [199], ainsi que plusieurs heuristiques par seuillages empiriques [124][143][258][129][196].

Nous verrons que les principaux efforts sur cette tâche se concentrent sur la construction de descripteurs pertinents pour identifier le chant sur un fond musical, plutôt que sur la méthode de classification.

2.4 Caractérisation audio

Nous avons vu, dans la section précédente, une collection parcellaire mais assez représentative des méthodes d'apprentissage statistique déployées sur les problèmes de classification audio. Ces dernières sont généralement le fruit des travaux de statisticiens et leur développement théorique est donc indépendant de toute application pratique. La caractérisation du signal audio, par la définition de descripteurs numériques susceptibles d'apporter l'information pertinente pour la tâche voulue, est au contraire fortement liée aux classes en présence. Nous présenterons ainsi les propositions de la littérature pour les deux problèmes posés : la classification parole/musique et la détection de chant.

Le calcul des descripteurs se base presque unanimement sur le principe du « sac de trames » (*bag of frames*), qui consiste à découper le signal audio en trames temporelles successives suffisamment courtes (de l'ordre de quelques centièmes de secondes) pour respecter la contrainte de quasi stationnarité des propriétés acoustiques. Le terme « sac de trames » décrit également le fait que

les trames sont considérées comme indépendantes, identiquement distribuées (IID) [145], et donc classifiées en dehors de toute considération d'ordre temporel, ce qui implique, comme nous le verrons dans la partie III, l'usage complémentaire de techniques de post-traitement pour réintroduire la donnée temporelle dans le processus de classification.

2.4.1 Classification parole/musique

Le problème de la discrimination entre parole et musique se focalise en général sur la caractérisation de la parole. En effet, la musique est un phénomène de nature très hétéroclite, impliquant une diversité de timbres et de dynamiques quasi infinie, et se trouve donc plus difficile à résumer par des propriétés simples.

C'est ainsi que le problème de classification parole/musique fut considéré à l'origine comme un problème annexe au traitement de la parole. Il est donc naturel que de nombreuses publications se soient dans un premier temps contentées de transposer l'usage de descripteurs reconnus dans ce domaine. Parmi ces descripteurs classiques on trouve les coefficients cepstraux sur échelle Mel (MFCC, *Mel Frequency Cepstral Coefficients*), qui constituent sans conteste le groupe de descripteurs le plus populaire dans la littérature [41][59][132][195][77][85][100][86][13], ainsi que les coefficients de prédiction linéaire (LPC, *Linear Prediction Coefficients*) [182][85] ou les coefficients de prédiction linéaire perceptifs (PLP, *Perceptual Linear Predictive analysis*) [9].

Alors que les articles précédents [106] se limitent au problème de la détection de voix parlée (généralement en présence de bruit), Saunders est le premier à publier [205] sur le problème spécifique de la classification parole/musique, suivi l'année suivante par l'article référence de Scheirer et Slaney [207]. Ces deux articles fournissent une analyse des propriétés permettant de discriminer parole et musique, dont nous retenons les points suivants :

- La voix parlée est une alternance de sons voisés (typiquement les voyelles et certaines consonnes), dont le spectre est quasi-harmonique, et de sons non-voisés (la plupart des consonnes) proches d'un bruit modulé. Cette alternance est beaucoup plus marquée que dans un signal de musique, où les parties harmoniques (notes tenues) sont généralement beaucoup plus longues que les parties non-harmoniques (percussives ou attaques transitoires).
- Cette alternance pour la parole se manifeste à une cadence relativement constante de 4 Hz que l'on nomme débit syllabique, et qui se traduit par un pic d'énergie autour de cette fréquence.
- Elle s'observe également en termes d'énergie globale puisque les consonnes non-voisées consistent généralement en attaques très fortes dont l'énergie contraste sensiblement avec les parties voisées. De plus, le signal de parole contient habituellement, si le débit n'est pas trop rapide, de nombreux interstices silencieux qui accentuent également cette alternance énergétique.
- L'alternance décrite précédemment se traduit également par des variations plus fréquentes du spectre d'un signal de parole que d'un signal de musique.
- La musique contient en général de nombreux phénomènes percussifs ou d'attaques qui se traduisent par un spectre centré sur une moyenne supérieure à celle du spectre de parole, qui lui se distingue dans les hautes fréquences par une nette décroissance spectrale d'environ 12 dB par octave.
- La voix est à priori plus localisée en fréquences, et limitée à 8 kHz, de même que la hauteur des sons, qui s'étend sur un ambitus moins large que la musique.

On peut y ajouter deux propriétés, secondaires parce qu'elles ne concernent pas directement le son lui-même :

- La musique populaire suit souvent un schéma rythmique très régulier qui se traduit par une périodicité marquée entre 40 et 200 battements par minutes.
- Les algorithmes de codage de la voix sont optimisés par rapport aux propriétés de cette dernière. Le résultat du codage d'un signal de musique par un algorithme de ce type doit donc, à débit constant, être plus bruité que sur un signal de parole. Le résiduel peut donc servir d'indice discriminant entre les deux sources.

La préoccupation principale de Saunders étant de délivrer un algorithme fonctionnant en temps réel, contrainte assez restrictive en 1996, il propose [205] une série de descripteurs exclusivement

basés sur le taux de tassage par zéro (ZCR, *Zero Crossing Rate*), dont le calcul est très rapide. Ces descripteurs consistent en une collection de processus d'intégration long-terme (moyenne, déviation standard, 3^{e} moment central, ...) appliqués sur les valeurs court-terme du ZCR. La pertinence de ce descripteur pour cette tâche est confirmée par le nombre de publications en faisant usage [201][41][42][261][202][176][169][140].

À la différence de Saunders, Scheirer et Slaney [207] ne se préoccupent pas des contraintes de temps de calcul et exploitent les caractéristiques énoncées plus haut en déployant une batterie de descripteurs beaucoup plus diversifiés, composée de la modulation d'énergie à 4 Hz, du taux de trames à basse énergie, de la fréquence du 95^{e} percentile d'énergie (appelée fréquence de coupure), du centroïde spectral (également appelé « clarté » sonore, *brightness*), du flux spectral, du ZCR, de la magnitude du résiduel après resynthèse spectrale, et enfin d'une mesure de battement rythmique. Les travaux des auteurs auront une certaine influence sur la communauté et l'on retrouvera un grand nombre de ces descripteurs dans la plupart des publications postérieures [42][202][87][182][151][139][169][13][229][140], le centroïde spectral et le flux spectral étant de loin les plus largement repris. On trouve en outre quelques descripteurs proches de ces derniers, comme le taux de hautes valeurs de ZCR (HZCRR, *High ZCR Ratio*) [139][68], le taux de trames silencieuses [71][138] ou le niveau d'activité [13], ou d'autres exemples de descripteurs spectraux mono-dimensionnels assez simples comme la largeur de bande [59][132][248][182][169], ou définis dans le standard MPEG 7 [3], comme la platitude spectrale et l'étalement spectral [41].

Bien qu'elle soit fortement liée aux conditions d'enregistrement, la mesure d'énergie instantanée (ou RMS, *Root Mean Square*), qu'elle soit mesurée sur le signal ou sur le spectre (en vertu du théorème de Plancherel), est également très populaire dans la littérature [41][132][261][87], au point de parfois constituer, de par son coût de calcul très réduit, la base principale de certaines propositions [176]. Certains auteurs préfèrent exploiter une version perceptive de ce dernier appelée *loudness*, qui prend en compte l'échelle de perception humaine quasi logarithmique [248][41][151]. L'energie instantanée apporte cependant peu d'information sur le contenu spectral, si bien qu'il est en général plus intéressant de calculer les énergies de sous-bandes fréquentielles [59][132][140][48], ou même les rapports d'énergie entre sous-bandes [182][151], qui impliquent une invariance par rapport à l'énergie globale ; la largeur de bande est parfois calibrée sur une échelle musicale comme l'octave [242][243]. Nwe et Li [170] font en outre précéder le calcul des énergies de sous-bande d'une phase d'accentuation harmonique (par le biais d'un banc de filtres triangulaires calés sur les partiels de la fréquence fondamentale estimée), destinée à atténuer les signaux non harmoniques.

Dans un article présentant une comparaison de différents descripteurs pour la tâche de classification parole/musique, Carey et al. [42] commentent les travaux de Saunders et Scheirer et Slaney en s'étonnant de ne pas y voir figurer une mesure impliquant la hauteur des sons, dont les variations dans la parole sont plus homogènes que dans la musique. Ils proposent en ce sens une mesure de fréquence fondamentale (ou *pitch*), que l'on retrouve également dans de nombreuses autres contributions [59][132][248][261][151] et dont l'apport peut également se traduire par une mesure du « rapport harmonique » [140][48][4], c'est-à-dire le taux de trames où une fréquence fondamentale peut être mesurée, qui permet ainsi de quantifier l'alternance entre trames voisées et non-voisées. Nielsen et al. [168] présentent une étude plus approfondie sur le pitch et proposent une série de descripteurs pour la classification audio, basés sur cette mesure.

Nous avons également mentionné la présence d'une structure rythmique, et en particulier d'un battement régulier, comme critère caractérisant le signal de musique. Ce point est en partie traité par la mesure de battements de Scheirer et Slaney, et sera repris et amélioré par Burred et Lerch [41] par le biais d'un histogramme d'intensités rythmiques sur lequel sont extraites diverses mesures statistiques (moyenne, déviation standard...) et une mesure de régularité par auto-corrélation. Tzanetakis et Cook ont également proposé [229] un algorithme très détaillé pour le calcul d'histogrammes d'intensités rythmiques basés sur une mesure d'auto-corrélation appliquée sur une estimation de l'enveloppe du signal. On retrouve encore l'usage de l'auto-corrélation pour la détection de rythme dans [111], cette fois appliquée sur les facteurs d'échelles du codage audio MPEG 1.

Les descripteurs spectraux présentés jusqu'ici sont tous construits sur une échelle linéaire des fréquences ou sur une échelle logarithmique. Plusieurs auteurs tentent de reproduire plus fidèlement le comportement auditif humain en appliquant des échelles perceptives pour le calcul de certaines grandeurs. Ainsi, dans [164], l'échelle Bark se substitue à l'échelle linéaire pour le calcul

du centroïde spectral. L'aspect psychoacoustique peut également être pris en compte sous d'autres formes, par exemple à travers la détection des phénomènes de rugosité (caractérisés par la modulation de l'enveloppe entre 20 et 150 Hz) ou d'enveloppes temporelles sur un banc de filtres d'échelle perceptive [151]. Des études beaucoup plus poussées visent à reproduire de manière précise le comportement du système auditif humain à travers différentes modalités, comme la représentation Taux-Echelle-Fréquence-Temps (*Rate-Scale-Frequency-Time*) introduite dans [194] ou le modèle cochléaire de [157]. Ces modèles sont toutefois d'une complexité sensiblement supérieure aux descripteurs de la littérature, et souvent inadaptés à un traitement en temps réel ou en ligne.

2.4.2 Détection de chant

Bien que la voix soit le point central dans les deux cas, le problème de la détection de chant diffère sensiblement de la classification parole/musique car les propriétés de la voix chantée ne sont pas les mêmes que celles de la voix parlée.

En premier lieu c'est le débit qui, sous la contrainte du temps musical, est profondément modifié par rapport à la voix naturelle (parlée). En effet, les notes musicales étant tenues par les chanteurs sur les voyelles, la proportion de trames voisées, qui n'est que de 60% en moyenne sur la parole, monte à 90% pour un signal de chant [53]. La plupart des descripteurs décrits précédemment basés sur l'alternance voisé/non-voisé sont donc moins pertinents dans ce cadre particulier. De plus, le débit syllabique à 4 Hz qui caractérise la parole devient ici caduque et ne permet plus d'identifier la voix chantée [52].

En définitive, on retient pour caractériser la voix chantée les critères suivants :

- Un des traits les plus souvent cités est sans doute le fameux « formant du chanteur », une résonance dans la bande de fréquences 2000-3000 Hz qui aide le chanteur à se faire entendre par dessus un accompagnement instrumental [223]. Mais l'accentuation de ce formant nécessite une technique très particulière que l'on ne retrouve guère qu'en musique lyrique, et ne permet donc pas d'identifier la voix chantée dans la musique populaire, qui constitue pourtant la cible principale de notre application.
- Le chant étant intrinsèquement de la musique, on retrouve certaines des propriétés énoncées précédemment pour différencier cette dernière de la parole. En particulier la dynamique des hauteurs musicales est beaucoup plus développée que dans la parole où la hauteur suit principalement une fonction prosodique qui se caractérise par des variations moindres et plus subtiles. À titre d'exemple on considère que la parole évolue habituellement entre 80 et 400 Hz tandis qu'une chanteuse soprano peut raisonnablement atteindre les 1400 Hz [199].
- Cette dynamique accrue se retrouve également sur les intensités, celles-ci faisant partie intégrante du langage musical (mais peut toutefois se trouver fortement réduite par les procédés de compression dynamique, couramment employés par les radios populaires).
- Comme nous l'avons mentionné plus haut, la voix chantée étire les sons voisés et peut en général être modélisée comme une séquence de hauteurs relativement constantes par morceaux (en suivant la terminologie mathématique), à la différence de la parole où la hauteur fluctue constamment pour les besoins de l'expression prosodique.
- La prédominance des sons voisés, et leur importance musicale, a pour effet de rendre le signal de chant beaucoup plus harmonique (dans le sens d'un spectre à structure de peigne régulier très marqué) que la parole.
- Enfin, l'un des traits qui caractérisent sans doute le mieux le chant est la présence quasi systématique d'un vibrato, que l'on peut plus ou moins différentier des vibratos instrumentaux [196].

La difficulté principale réside dans le fait que le signal de chant est mélangé au fond instrumental, qui peut être d'intensité comparable à la partie de voix, et dont le contenu est généralement fortement corrélé à celle-ci, en termes de schéma rythmique ou de notes jouées ; il est donc d'autant plus complexe de distinguer les deux contributions dans le signal, que la musique couvre une bande fréquentielle très large et ne peut donc être isolé du mélange qu'au prix d'une forte dégradation du signal de chant.

Pour certaines publications, la phase de détection de chant n'est qu'un pré-traitement pour

l'application d'algorithmes de reconnaissance du chanteur, aussi les solutions proposées y sont généralement assez simples et l'on retrouve quelques cas d'algorithmes de détection de la parole adaptés pour l'occasion [134][105]. La filiation évidente avec la détection de la parole, malgré les différences énoncées plus haut, se traduit également par l'exploitation de descripteurs classiques dans ce domaine, tels les PLP [1] [25][131][237], les LFPC [172], les MFCC [237][135][141], ainsi que d'autres descripteurs que nous avons présentés pour la classification parole/musique (énergie, ZCR, flux spectral...) [258].

Nous avons mentionné comme caractéristique principale de la voix chantée la présence de vibrato. Celui-ci se manifeste par une modulation conjointe du son en fréquence et en intensité (cette seconde modalité est parfois appelée *tremolo* pour la distinguer du vibrato fréquentiel), à la différence des instruments de musique qui dans leur grande majorité ne produisent qu'un seul de ces phénomènes à la fois. Ainsi dans le cas des instruments à vent c'est le tremolo qui prédomine tandis que les cordes favorisent le vibrato fréquentiel [196]. Lachambre et al. [129] proposent ainsi un critère de mesure de vibrato basé sur la recherche d'un pic fréquentiel entre 4 et 8 Hz. Afin de répondre à une contrainte de stabilité spectrale, le signal est segmenté en trames temporelles dont les frontières sont déterminées à partir de la structure des pics fréquentiels dans le spectrogramme. Le critère proposé consiste à calculer le taux de trames où le vibrato est détecté, parmi les trames d'un même segment temporel. Regnier et Peeters [196] accroissent la robustesse du critère en combinant les mesures de modulation de fréquence et d'intensité pour la détection de partiels dits « vibrants ». L'observation de plusieurs partiels vibrants simultanés détermine alors la détection de chant. Ces deux approches sont très efficaces car les critères proposés sont fortement discriminants pour la tâche considérée. Cependant elles impliquent toutes deux une phase très coûteuse de détection de partiels dans le spectrogramme. Nwe et Li [171] proposent à l'inverse une approche plus économique, mais moins efficace, basée sur le calcul de coefficients cepstraux après l'application de « filtres numériques de vibrato », dont la définition manque de clarté.

Les mêmes auteurs complètent cet apport en construisant d'autres filtres caractérisant certaines propriétés du chant. Ainsi un banc de filtres centrés sur les moyennes des formants permet également d'accentuer les résonances vocales. Un autre processus, appelé « attenuation harmonique » et initialement introduit dans [172], vient atténuer le signal de musique par rapport au chant par un filtrage harmonique triangulaire, le vibrato fréquentiel ayant un effet d'étalement des pics harmoniques spectraux qui rend donc le signal de chant moins harmonique que la musique. Kim et Whitman [124] emploient paradoxalement un traitement similaire (filtrage par peigne harmonique sur la fondamentale, après filtre passe bande entre 200 et 2500 Hz) en le justifiant par l'argument contraire, à savoir que le chant est plus harmonique et qu'un seuil sur la mesure d'harmonicité peut ainsi constituer un critère de décision satisfaisant.

On retrouve également les descripteurs introduits à l'origine par Williams et Ellis [247] pour la tâche de classification parole/musique, adaptés dans [24] pour la détection de chant. Les PPF (*Post Probability Features*) sont le résultat à 54 composantes d'un réseau de neurones de reconnaissance de phonèmes de la parole. Les auteurs comparent par la suite l'efficacité de ces descripteurs dans une approche classique du maximum de vraisemblance avec des critères d'information (entropie, dynamisme, ...) calculés sur ces derniers.

Enfin, Maddage et al. proposent une approche originale [143] qui consiste à itérer deux fois la transformée de Fourier sur une fenêtre de signal. En effet, si l'on considère que la FFT [2] d'un signal périodique est un train de pulsation périodique (les partiels), alors la FFT de cette FFT est un sinus cardinal, dont les premières composantes contiennent plus d'énergie dans le cas du chant, parce que son spectre harmonique est plus dense. Un seuil sur l'énergie cumulée des premières composantes permet ainsi de décider entre chant et musique. Cependant, pour pouvoir appliquer la FFT sur un spectre stationnaire, l'algorithme nécessite une première phase de détection du rythme assez coûteuse afin d'être appliqué sur chaque fenêtre encadrée pardes battements successifs.

1. se référer à la section 6.5 sur les descripteurs employés pour la classification parole/musique pour la signification des acronymes.
2. *Fast Fourier Transform*, désigne ici le résultat du calcul numérique de la transformée de Fourier.

2.4.3 Discussion

Cet état de l'art donne une idée de la diversité des descripteurs mis en jeu sur les problèmes posés. Bien que la plupart s'accompagnent d'une argumentation raisonnée concernant leur efficacité réelle ou supposée, il est a priori impossible pour l'expérimentateur qui voudrait tirer parti de ces contributions de déterminer lesquels choisir en premier lieu, d'un point de vue théorique, d'autant que beaucoup d'entre eux sont largement redondants.

Beaucoup d'auteurs proposent ainsi des protocoles expérimentaux comparant les résultats obtenus pour un même classifieur avec chacun des descripteurs d'un ensemble donné. C'est l'approche que suivent par exemple Scheirer et Slaney [207] sur l'ensemble des descripteurs qu'ils proposent pour la tâche de classification parole/musique, et que l'on retrouve dans une longue liste de publications : [42][195][68][202][176][87]...

Il peut bien sûr être avantageux de grouper les différents descripteurs dans la phase de classification, et l'on trouve parfois dans ces comparaisons les résultats obtenus pour différentes combinaisons, qui montrent en général l'avantage à exploiter tous les descripteurs en même temps. Ce résultat met pourtant en évidence les limites d'un tel protocole, d'abord parce qu'en définitive il ne fait que montrer que l'accumulation d'information profite d'une façon ou d'une autre au classifieur, ensuite parce qu'il exclut la prise en compte de la complexité induite par la superposition ded descripteurs. Or, pour les applications visées par notre système (l'annotation de gros volumes d'archives et la classification du flux audio en direct), le coût en temps de calcul est un aspect essentiel. De plus, nous verrons que l'accumulation d'un grand nombre de descripteurs finit par être préjudiciable au processus de classification parce qu'une partie de cette information peut se révéler non pertinente pour la tâche choisie, et le reste largement redondant.

Une alternative raisonnable consiste à choisir une combinaison de taille raisonnable parmi une collection de descripteurs disponibles. L'opération, si elle est guidée par une mesure de performances après apprentissage du classificateur, se révèle vite irréaliste car l'évaluation de tous les sous-ensembles possibles implique une explosion combinatoire.

La sélection automatique de descripteurs est en fait un sujet d'importance croissante dans le domaine de l'apprentissage statistique, dont le développement est en grande partie dû aux besoins de la bioinformatique, qui traite couramment des données contenant plusieurs milliers de composantes fortement redondantes ou bruitées. Un examen poussé de la littérature sur les deux problèmes posés nous indique pourtant que pratiquement aucun article ne met à contribution ces techniques de sélection de descripteurs. On peut néanmoins citer Peeters [179] qui exploite l'algorithme IRMFSP (que nous présenterons dans la section 7.3.2), algorithme relativement simple mais efficace qu'il avait proposé précédemment [181] pour la reconnaissance d'instruments de musique, et Rocamora [199] qui utilise un algorithme basé sur les mesures de corrélation entre descripteurs, pour la détection de chant.

Nous proposerons donc dans ce document une approche qui se démarque des publications antérieures en mettant l'accent non sur le développement de nouveaux descripteurs mais sur la mise en place d'un cadre de sélection efficace appliqué sur un grand ensemble de descripteurs collectés dans la littérature. Nous verrons en outre que la notion de pertinence n'est pas absolue et que l'efficacité d'un ensemble de descripteurs est fortement liée au classifieur mis en jeu, ce qui nous amènera à proposer de nouveaux algorithmes de sélection adaptés aux Machines à Vecteurs de Support.