Détection de la voix chantée dans un morceau de musique

Master ATIAM
(Acoustique, Traitement du signal, Informatique Appliqué à la Musique)
Année : 2007-2008

Lise REGNIER **sous la direction de Geoffroy PEETERS**

Institut de Recherche et de Coordination Acoustique / Musique
1 Place Igor Stravinsky, 75004 Paris

Université Paris VI Pierre & Marie Curie
Jussieu, 75012 Paris

Résumé

De tous les éléments constitutifs d'un morceau de musique, la voix est sans conteste celui qui focalise le plus l'attention de l'auditeur. Ceci provient du fait que la voix est porteuse d'un message (le texte) et d'une identité (celle du chanteur). Pour ces raisons, de nombreuses applications dans le domaine de l'indexation repose sur sa description. Afin de décrire ou d'extraire la voix chantée dans un mélange polyphonique de signaux sonores il est nécessaire de procéder, dans un premier temps, à sa détection.

La localisation de la voix dans un flux audio étant la base de tous les travaux portant sur la voix chantée de nombreuses recherches ont déjà été effectuées comme le montre l'état de l'art dressé dans le second chapitre de ce rapport. Ces méthodes reposent toutes sur des approches dites "aveugle" consistant à extraire du signal des descripteurs audio "génériques" utilisés ensuite pour apprendre les deux classes de segments "chanté" et "non chanté" à l'aide d'algorithmes statistiques. La classification obtenu est généralement lissé a posteriori selon différentes techniques répertoriées dans le sixième chapitre. Dans les meilleurs cas, ces méthodes permettent d'obtenir une classification correcte à 85%.

Dans un premier temps, nous avons également utilisé ce type d'approche afin de disposer d'une méthode de référence. La description de cette méthode fait l'objet du troisième chapitre. Les descripteurs utilisés sont les MFCC (Mel Frequency Cepstral Coefficients) et la SFM (Spectral Flatness Measure) ainsi que leurs dérivées temporelle première et seconde. L'apprentissage est effectué en assignant chacune des trames d'analyse à une des classes "chanté" ou "non chantée". Chaque classe est modélisée par un mélange de gaussiennes (GMM). Nous disposons, pour tester cette méthode d'une base de test constituée de 90 chansons annotées manuellement dont la majeure partie est utilisée pour l'entraînement du modèle statistique. L'évaluation de cette méthode, effectuée sur 16 chansons, permet d'obtenir une classification correcte à 75% avant lissage et à 77% après lissage.

Ces approches, dites aveugles, sont souvent utilisées pour la classification d'instrument, la discrimination parole/musique, la reconnaissance de locuteurs... Nous reprochons à ces approches de ne pas prendre en compte les caractéristiques propres à la voix chantée dont une description est donnée dans le premier chapitre de ce document. En s'appuyant sur l'analyse de spectrogrammes de voix chantée nous avons constaté que la voix possédait trois caractéristiques majeures : le voisement (ou harmonicité), le vibrato et les formants. Nous avons étudié, dans le quatrième chapitre, différents descripteurs permettant de mesurer le coefficient d'harmonicité de la voix. Le cinquième chapitre est consacré à l'étude du vibrato. La voix possède la caractéristique d'être naturellement vibrée. Son vibrato qui, par définition, est une modulation de fréquence entraîne une modulation d'amplitude caractéristique. C'est cette propriété que nous allons particulièrement exploité dans l'approche totalement nouvelle que nous proposons.

Nous avons mis en place une méthode automatique de détection de la voix chantée basée sur le suivi de partiel. Dans cette méthode, seuls les partiels avec modulation de fréquence et d'amplitude sont conservés. Dans un premier temps, cette méthode est appliquée sur un seul canal des fichiers audio. Cette méthode qui ne nécessite aucun apprentissage et utilise un descripteur unique permet d'obtenir une classification correcte à 74%. Les résultats détaillés de cette approche figurent dans le septième chapitre de ce document.

En partant du constat que la voix est souvent présente sur les deux canaux du fichier audio

nous avons amélioré la classification en exploitant la stéréo. La méthode automatique appliquée sur chacun des canaux permet d'obtenir deux segmentations. Nous avons étudier les résultats trouvés en utilisant l'intersection et la réunion des deux segmentations.

Les résultats obtenus par notre méthode sont extrêmement encourageants. Dans de futurs travaux, on prévoit de combiner ce descripteur à ceux utilisés dans la méthode de référence pour améliorer le résultat final. Par ailleurs le critère d'harmonicité étudié au quatrième chapitre n'a pas été incorporé à la méthode.

Remerciements

C'est une chance inouïe de pouvoir se consacrer à ses passions et cette chance m'a été offerte par Geoffroy Peeters et Xavier Rodet qui m'ont accepté en stage au sein de l'équipe analyse synthèse de l'ircam.

Je remercie particulièrement Geoffroy Peeters, mon encadrant, qui n'a pas ménagé son temps et son énergie pour m'aider tout au long de ce projet de recherche. Je le remercie pour sa disponibilité, sa compréhension et surtout pour m'avoir transmis une partie de ses connaissances et de son expérience. Ses nombreux conseils seront un pilier dans ma vie professionnelle future.

Je remercie également Damien Tardieu, mon collège de bureau, pour sa précieuse aide, ses nombreux conseils scientifiques, sa patience et sa gentillesse.

Je remercie chaleureusement Emmanuel Deruty et Gaétan Parseihian pour leur soutient moral et affectif tout au long de ce stage.

Je remercie également Maël Derio et Nicolas Obin pour le temps qu'ils ont consacré à la relecture de ce rapport.

Je tient à remercier à Mathieur Ramona qui m'a fournit la base de test qui a servie pour cette étude.

Pour finir, je tient à remercier toutes les personnes qui ont rendu cette année d'ATIAM agréable : l'équipe pédagogique du master ainsi que les étudiants.

Table des matières

Chapitre 1

Introduction

1.1 Importance de la voix chantée dans la musique et ses caractéristiques acoustiques

La diffusion et la distribution de musique numérique devenant de plus en plus importantes les bases de données, professionnelles comme personnelles, augmentent de façon considérable. Il est devenu nécessaire de disposer d'outils permettant de classer et d'accéder à ces bases en effectuant des recherches par le contenu musical. C'est ce que tente de faire l'inter discipline appelée Music Information Retrieval (MIR).

Dans la musique populaire, un groupe de musique est souvent caractérisé par son chanteur. On considère que la voix est l'élément le plus important d'une chanson dans la mesure ou elle porte la mélodie principale et qu'elle contient le message à faire passer par les paroles. Ce constat nous pousse à placer le chant comme élément principal dans l'organisation, le parcours et l'accès au morceaux de musique.

Parallèlement au problème d'organisation et d'accès des bases de données il est devenu nécessaire d'archiver un certain nombre d'éléments comme les émissions de radio, de télévision... Cet archivage peut être basé sur la segmentation de longues heures d'enregistrements en parole ou musique par exemple. Lors des recherches sur la discrimination parole/musique, on a pu constater que la voix chantée était la plupart du temps classée comme de la parole. Cette erreur qui limite la performance de la classification parole/musique à poussé des recherches spécifiques sur la reconnaissance de la voix chantée [CG01].

1.1.1 Problématique

Le problème de la détection de segments chantés dans un flux audio peut être posé de la façon suivante : pour une chanson donnée, trouver les instants qui permettent de partitionner une chanson en segments purement instrumentaux, ou chantés. Le problème majeur réside dans le fait que la voix est dans la pratique accompagnée par d'autres instruments.

1.1.2 Motivations

La segmentation d'un morceau en parties chantée et non-chantée peut trouver de nombreuses applications. Parmi celles-ci on trouve la synchronisation des paroles, la recherche par le chant, le "thumbnail generation". Une fois les segments chantés isolés, la caractérisation de ces derniers permet de procéder à l'identification du chanteur. Un tel outil permet à l'utilisateur de retrouver les informations sur le chanteur de n'importe quelle chanson ou encore de retrouver toutes les chansons interprétées par un chanteur en particulier. Cette technologie permet également de retrouver les chansons dont le chanteur à une voix proche de celle d'un chanteur donné [Zha03a]. Il va de soi que l'application des droits d'auteurs est vivement intéressée par cette recherche. Des recherches concernant l'extraction de la mélodie principale, la reconnaissance de la langue,

la transcription et le suivi des paroles complètent ces travaux et trouvent leurs applications dans la création d'outils pour le karaoké.

Afin de trouver une méthode susceptible de détecter la voix chantée dans un morceau de musique, on s'interroge sur ses caractéristiques. On se questionne sur ce qui permet de différencier la voix parlée de la voix chantée puis sur ce qui nous permet de différencier la voix des autres instruments de musique.

1.1.3 Différence entre la voix parlée et la voix chantée

Les recherches sur la discrimination parole/musique ont montré que la voix chantée s'apparente plus à de la parole qu'à de la musique. En effet, l'analyse à court terme d'un signal de parole et d'un signal de voix chantée ne comporte pas de différence significative. On analyse dans les deux cas la présence de formant. Par contre, l'analyse à long terme permet de mettre en évidence certains éléments discriminants comme la hauteur ou l'aspect sinusoïdal.

Modulation d'énergie

Scheirer met en évidence, dans [SS97], que les signaux de parole ont un pic de modulation d'énergie autour de 4Hz qui correspond à la vitesse syllabique. Ainsi lorsque l'énergie normalisée est filtrée à 4Hz par un filtre passe-bande, les signaux de parole ont une valeur de sortie supérieure à celle obtenue pour les signaux de voix chantée. Chou fait remarquer, dans [CG01], que ce descripteur n'est pas très robuste puisque les signaux de parole peuvent avoir une faible énergie à 4Hz si le locuteur parle vite. Paradoxalement, les signaux de voix chantée peuvent avoir une très forte énergie à 4Hz ce qui équivaut à une noire de pulsation 240 bpm.

Voisement

Par ailleurs, Cook fait remarquer dans [Coo90] que la voix chantée est voisée à 90% par opposition à la voix parlée qui n'est voisée qu'à 60%. En traitement de la parole, on distingue les voyelles qui sont voisées car elle nécessitent la mise en vibration des cordes vocales des consonnes qui s'apparentent à du bruit. Dans la voix chantée, les voyelles sont tenues plus longtemps que les consonnes puisqu ?elles seules produisent une hauteur. Par ailleurs, le contour de pitch de la voix chantée est constant par morceaux avec des changements radicaux entre deux alors que ces changements sont continues dans le cas de la voix parlée. [RH07]

Tessiture

Les fréquences fondamentales de la voix chantée s'étendent de 80Hz (cas d'une basse) à 1400Hz (cas d'une soprano) alors que la limite supérieure est de 400Hz dans la voix parlée.[LW06a]. La majorité de l'énergie d'une voix chantée se situe entre 200 et 2000Hz. Cette région est la plus sensible de l'oreille humaine. Cette région est également caractérisée par son effet de masquage fréquentiel [KW02].

1.1.4 Caractéristique de la voix chantée

A la lecture d'un spectrogramme, il est assez aisé de voir les segments comportant de la voix. On essaye donc de trouver les éléments qui rendent la voix si facilement discernable des autres instruments aussi bien à la lecture d'un spectrogramme qu'à l'écoute d'une chanson.

L'harmonicité

De nombreux auteurs ont mis en évidence le caractère harmonique de la voix. En effet, la voix possède de nombreuses harmoniques qu'il est facile de visualiser sur un spectrogramme. De

plus la voix est très souvent doublée par l'accompagnement de façon harmonique ce qui renforce le caractère harmonique des parties chantées d'un morceau.

Le formant du chanteur

Il existe une autre caractéristique, spécifique au chant lyrique : le formant du chanteur. Ce formant est une extra résonance située entre 2000 et 3000 Hz qui permet de laisser passer la voix au dessus de l'accompagnement [KW02] [SR90].

Le vibrato

Pour finir, il existe une caractéristique très marquée du chant : le vibrato. Tout son monophonique, auquel on applique une modulation de fréquence d'environ 6Hz peut être naïvement perçu comme du chant. Comme le suggère Fujihara dans [FKG$^+$05] la vibrato est une clé importante dans la discrimination chanté, non chanté. Pour la voix, la présence de vibrato implique la présence de trémolo, puisque les harmoniques suivent en amplitude les formants. Ainsi, la fréquence de modulation est la même que la fréquence de modulation du vibrato comme l'explique Sundberg dans[SR90].

Nous verrons dans un second temps comment ces différentes caractéristiques ont été exploitées, dans la littérature, pour discriminer les segments chantés des segments non chantés.

Chapitre 2

État de l'art

L'ensemble des éléments cités ci-dessous sont extraits des recherches menées dans le domaine de la segmentation d'un morceau de musique en parties chantées et parties non chantées ou dans le domaine de l'identification des chanteurs. Ces dernières nécessitant une segmentation au préalable.

2.1 Problématique

La voix chantée se situe à l'interface entre la musique et la parole, ainsi les études concernant le chant sont héritées de ces deux domaines. La détection de la voix peut être considéré comme un cas particulier de l'identification d'un instrument parmi un mélange d'instruments. C'est pourquoi les méthodes et les descripteurs utilisés pour la classification d'instruments peuvent être appliquées. On distingue trois types d'approches : celles issues de la parole (reconnaissance de phonèmes), les méthodes statistiques basées sur la combinaison de descripteurs et sur des techniques d'apprentissage et enfin les méthodes automatiques (sans apprentissage) utilisant des seuils.

2.2 Méthodes issues du domaine du traitement de la parole

Les premières recherches concernant la détection de voix chantée sont issues du domaine de la parole. Dans [BE01], les auteurs utilisent un outil de reconnaissance de la parole pour distinguer les segments vocaux de l'accompagnement. Un réseau de neurones est d'abord entraîné pour distinguer les différents phonèmes de l'anglais. Ce réseau de neurones génère des vecteurs de "posterior probability feature" (PFF's). La pertinence des données est représentée par un "posteriogramme". Lorsque la correspondance avec un phonème est bonne le posteriogramme est caractérisée par une valeur élevée pour un seul phonème. En l'absence de parole, le posteriogramme comporte des valeurs plus ou moins élevées pour plusieurs phonèmes puisque le choix est incertain. Cette différence est modélisée par un HMM à deux états.

2.3 Méthodes de classification statistique

Les techniques basées sur des méthodes d'apprentissage (ou systèmes de classification statistique) sont celles que l'on rencontre le plus souvent. On trouve des chaînes de Markov cachées (HMM) dans [BE01] et [NW04], des modèles de mixture de Gaussiennes (GMM) dans [TW06], [LW06a] et [LGD07], des réseaux de neurones artificiels (ANN) dans[BE01], [Tza04] et des machines à support de vecteur (SVM) dans [MXW03] ,[MWXW04] et [RR08].

Un système de classification statistique modélise chaque classe de signaux par une variable aléatoire qui, le plus souvent, est une distribution de Gauss. Dans notre cas, on distingue la

classe "purement instrumental" (PI) et la classe "instrument plus voix" (I+V). Dans de rares cas, on peut disposer d'une classe chant seul. Ces méthodes nécessitent deux étapes : une phase d'apprentissage et une phase de classification.

- L'apprentissage a pour but d'estimer les paramètres des distributions de Gauss qui composent le modèle. Cet apprentissage n'est pas réalisé sur les signaux eux-mêmes mais sur un ensemble de paramètres, judicieusement choisis, extraits à partir de ceux ci.
- La phase de classification permet de déterminer la classe d'appartenance d'un signal donné. Les paramètres extraits du signal à classer sont comparés aux paramètres de chacune des classes. Le calcul de la vraisemblance des informations permet de décider la classe d'appartenance la plus probable du signal.

2.3.1 Lissage, amélioration de la classification

Les résultats obtenus lors de la classification peuvent êtres améliorés par des systèmes de filtrage comme le montrent des études récentes. Dans [RR08] , Ramona propose de filtrer les probabilités obtenues à posteriori par un filtre médian. Il propose également de lisser ces probabilités à l'aide d'un modèle de Markov caché à deux états (V+I et I). Lukashevich propose une méthode pour dériver les fonctions de décision et les lisser à l'aide d'un filtre ARMA (Autoregressive Moving Average) dans [LGD07].

Les paramètres extraient du signal acoustique sont généralement appelés descripteurs ou "features". Afin d'améliorer la rapidité et la performance de la classification il est nécessaire de trouver des descripteurs dont la dimension est raisonnable comparée à celle du signal. La performance d'un système de classification dépend essentiellement du choix des descripteurs.

2.4 Descripteurs utilisés dans la discrimination chanté non-chanté

Il a été montré, lors des recherches sur la classification des instruments, qu'il était intéressant de combiner les descripteurs temporels et spectraux [Mar99]. De nombreux descripteurs, souvent issus du domaine de la parole ou de la classification d'instruments, ont été testés pour la discrimination chanté / non-chanté.

2.4.1 Le timbre

Les descripteurs les plus souvent utilisés sont ceux qui visent à décrire le timbre. On admettra que le timbre est l'élément qui permet de distinguer deux instruments jouant la même note sur une même durée à la même sonie. Le timbre est en partie décrit par le spectre du signal. Les descripteurs de timbre sont des vecteurs contenant un nombre réduit de coefficients qui permettent de modéliser la forme de l'enveloppe spectrale. Le descripteur le plus souvent rencontré dans la littérature ([TW06], [MXW03], ...) est le vecteur de MFCC (Mel Frequency Cespstral Coefficients) . Ce sont d'ailleurs ces coefficients qui fournissent les meilleurs résultats comme le montre Rocamora dans sa comparaison des différents descripteurs utilisés pour identifier les segments de voix chantée [RH07]. Les MFCC, calculés à partir de l'échelle des mels, peuvent trouver des variantes. Nwe propose, dans [NL07], une comparaison des coefficients cepstraux calculés sur les bandes de Mel, les bandes de Bark ou encore les bandes d'octaves.

Les coefficients de prédiction linéaire (LPC) sont également utilisés pour décrire le timbre dans [KW02] et [MXW03]. À l'origine, ces coefficients sont utilisés pour localiser les formants dans la voix parlée. On trouve également des variantes de la LPC. Berenzweig et Ellis utilisent des "perceptual linear prediction coefficients" (PLP) dans [BEL02] et [BE01]. Kim utilise les warped LPC dans [KW02] .

Ces descripteurs (MFCC, LPC et PLP), issus du domaine du traitement la parole, ne sont pas forcément appropriés pour mettre en évidence les caractéristiques de la voix chantée en

présence d'accompagnement comme le montre Nwe et Maddage dans [NW04] et [MWXW04] respectivement.

Des descripteurs comme les Log Frequency Power Coefficient (LFPC) [NW04] et les LFPC obtenus après filtrage harmonique (HA-LFPC) [NSW04], le flux spectral [Zha03b], le centroïd spectral ou encore le rolloff sont également utilisés pour décrire le timbre [Tza04].

Les dérivées temporelles première et seconde des descripteurs sont utilisées pour capturer l'évolution des informations qu'ils contiennent au cours du temps comme le propose Berenzweig dans [BEL02].

2.4.2 Variation d'énergie

Comme le fait remarquer Zhang dans [Zha03b] la voix entre généralement alors que l'accompagnement a déjà commencé ce qui a pour effet d'augmenter brutalement l'énergie du signal. Zhang utilise les variations d'énergie pour détecter les instants de départ de la voix [Zha03a], Nwe constate que les segments vocaux contiennent plus d'énergie dans les hautes fréquences que les segments non vocaux [NW04]. Tzanetakis utilise également [Tza04] les énergies relatives de certaines bandes de fréquence pour caractériser les segments vocaux.

2.4.3 Coefficient d'harmonicité

La voix chantée possède la particularité d'être fortement harmonique. Cette caractéristique, relatée à l'aide d'un coefficient d'harmonicité ou de voisement, a été utilisé à de nombreux reprises. Chou définit le coefficient harmonique comme la moyenne harmonique du maximum de l'auto corrélation dans le domaine temporel et dans le domaine fréquentiel [CG01]. Cette définition du coefficient d'harmonicité a été reprise par Zhang dans [Zha03a]. Nwe utilise un ensemble de filtres destiné à diminuer les harmoniques en prenant en compte la tonalité du morceau [NW04]. Apres atténuation les signaux vocaux possèdent plus d'énergie que les signaux non vocaux car les harmoniques sont plus régulières dans le cas de ces derniers signaux non vocaux à cause du vibrato et de l'intonation. Kim utilise un filtre en peigne inverse dont le délai varie [KW02] afin de trouver la fréquence pour laquelle le signal est le plus atténué. La mesure d'harmonicité est définie comme le rapport de l'énergie du signal sur l'énergie du signal le plus atténué. Enfin, un seuil permet de détecter les sons harmoniques. Dans [SWW05] les harmoniques sont filtrées par un peigne après avoir trouvé la fréquence fondamentale. En général après filtrage des sons harmoniques la voix chantée reste présente à cause du vibrato et de l'intonation. L'énergie dans les différentes sous bandes est calculée et les trames possédant la plus forte énergie sont considérées comme chantées. Maddage propose une approche différente, appelée TIFCT pour détecter la structure harmonique de la musique instrumentale et vocale dans [MWXW04]. Cette approche part du postulat que la voix chantée possède de fortes harmoniques dans les hautes fréquences. Les valeurs de la transformé de Fourrier dans les hautes fréquences sont modélisées par un train d'impulsion périodique. Une seconde FFT est alors appliquée au modèle pour produire une enveloppe de sinus cardinal. Cette double transformée de Fourrier à pour effet de compresser l'énergie originale dans les premiers coefficients de la deuxième FFT. Un seuil basé sur l'énergie présente dans ces coefficients est utilisé pour discriminer les trames chantées des trames non chantées.

2.4.4 Vibrato de la voix chantée

À ce jour, seuls les articles de Nwe [NSW04] et [NL07] essaient de discriminer la voix chantée en utilisant le vibrato. Un ensemble de filtres est construit afin de détecter différents types de vibrato. D'autres recherches, dont les applications se trouvent dans la classification des sons, proposent des techniques d'extraction et d'estimation des paramètres du vibrato. Rossignol dans sa thèse [RDS+99] propose différentes méthodes pour détecter le vibrato ou extraire ses paramètres sur des sons monophoniques.

2.4.5 Attaque-décroissance

Nwe exploite également deux particularités de la voix chantée qui lui permettent de passer facilement au-dessus de l'accompagnement instrumental : l'attaque-décroissance et le formant du chanteur.

Des études ont montré que l'oreille humaine est particulièrement sensible à l'attaque et au temps de décroissance des différents instruments de musique et de la voix. Ces deux éléments permettent de reconnaître les différents instruments. En général, pour un signal vocal, la durée d'attaque est très brève et la décroissance est progressive comparée à celles des signaux non-vocaux.

2.4.6 Les formants et le formant du chanteur

Les fréquences des formants sont déterminées par la forme du conduit bucco nasal. En général la localisation de trois formants permet d'identifier les voyelles de la voix parlée. Sundberg à montré dans [SR90] que le spectre du chant lyrique est caractérisé par une importante énergie entre 2.5 et 3kHz. Ce pic du spectre, appelé formant du chanteur, permet à la voix de passer facilement au-dessus de l'accompagnement orchestral. Pour intégrer les formants et particulièrement le formant du chanteur aux paramètres acoustiques, les auteurs de [NSW04] utilisent un ensemble de filtres en sous-bandes centrés sur les fréquences des formants.

Pour exploiter ces trois dernières caractéristiques, la sortie des filtres est analysée. Une approche similaire à celle utilisée pour le calcul des MFCC est employée pour obtenir un nombre réduit de coefficients : Une transformée en cosinus discret (DCT) est appliqué au logarithme de l'énergie de chaque sous bande du signal filtré puis les 13 premiers coefficients sont conservés. Les descripteurs de vibrato et de formant sont appelés $OFCC_{vib}$ et $OFCC_{har}$ respectivement.

2.5 Segmentation du signal

Dans ces approches, les descripteurs sont extraits à partir de courtes portions du signal appelées trames durant lesquelles le signal peut être considéré comme stationnaire. Plusieurs stratégies sont adoptées pour partitionner temporellement le signal. Lors de la classification à court terme, chaque trame contient une information locale et la classification obtenue est généralement trop bruitée. Plusieurs auteurs suggèrent d'utiliser les informations à long terme pour lisser la classification obtenue [TW06] ou pour découper le signal en segments regroupant plusieurs trames appartenant à la même classe. Cette segmentation peut être basée sur le tempo comme le propose Nwe et Maddage dans [NW04] et [MXW03], sur le changement d'accords comme le suggère Maddage dans [MWXW04] ou encore sur les changements de timbre [LW06a]. La segmentation peut également être réalisée en modélisant la structure du morceau en intro, couplet, refrain, pont, etc grâce a l'utilisation de HMM comme le suggère Nwe dans l'article [NW04].

2.6 Résultats annoncés

Il est difficile de comparer les résultats de toutes les références puisque les bases de test, les critères d'annotation et d'évaluation varient de façon considérable d'un article à l'autre. On recense dans le tableau suivant pour chaque études les descripteurs et les systèmes de classification qui ont fournit les meilleurs résultats. La base utilisée et la méthode de segmentation sont également précisées. On note que les études concernant l'identification des chanteurs ne nécessitent pas une détection de la voix aussi complète que celles dont le but est la segmentation automatique d'un morceau en parties chantés, non-chantées.

Étude	But	Descripteurs	Classification	Base	Trames	Résultats
[BE01]	Détection par reconnaissance de phonèmes	PPF $log(L_{mus})$ $\log(L_{voc})$ Entropie des PPF	HMM	245 segments de 15 sec	16ms	82.2%
[CG01]	Discrimination parole musique amélioré par la détection de chant	MFCC + Δ MFCC + ΔΔ MFCC + Energie + Δ Energie + ΔΔ Energie	GMM Parole/Musique Cross-validation	26 minutes de parole, musique et chant		
[KW02]	Identification du chanteur	Harmonicité	Seuil	20 chansons	100ms	PI : 79.3% I+V : 30.7% Moy : 54.9%
[BEL02]	Identification du chanteur	PLP + Δ PLP + ΔΔ PLP	MLP (multi-layer perceptron neural)	80 échantillons de 15sec	32ms	40% des trames vocales
[Zha03a]	Identification du chanteur, instant de départ de la voix	Energie + ZRC + harmonicité + flux spectral	GMM pour chaque chanteur	45 chansons (8 chanteurs)	16ms	
[Tza04]	Méthode semi-automatique de détection	Mean Centoid + Rollof + Flux + Relative subband energy + Δ centroid + Δ rolloff + Δ flux	ANN + bootstraping	10 chansons de jazz (5 chanteuses, 1 chanteur)		75 %
[NW04]	Détection automatique	LFPC	MM-HMM + Boostraping	20 chansons	Structure du morceau en Intro, couplet, refrain... 1000ms	PI : 82% I+V : 86.6% Moy :84.3%

suite page suivante

[NSW04]	Détection automatique	HA-LFPC	MM-HMM + Boostraping	20 chansons		PI : 79.2% I+V : 94.1% Moy :86.7%
[LW06b]	Extraction de la mélodie chantée	MFCC	GMM	10 chansons		86.65%
[FKG+05]	Identification du chanteur	MFCC	64 mixture vocal -GMM, 80 mixture non-vocal -GMM	242 chansons avec un seul chanteur, 22 avec deux chanteurs et 174 instrumentales	32ms	82.3%
[NL07]	Détection automatique basée sur des descripteurs perceptifs	SFCC + $OFCC_{har}$ + $OFCC_{vib}$	HMM	84 chansons	55ms	82.8 %
[RR08]	Détection automatique	centroid + slope + width + asymetry + flux + MFCC + LPC + Zero Crossing rate + Estimation de F0 + algorithme IRMFSP pour trier les descripteurs	SVM + lissage par HMM	84 chansons	190ms	PI : 82% I+V : 84.4% Moy : 83.2%
[LGD07]	Détection automatique avec filtrage ARMA	MFCC Δ MFCC + $\Delta\Delta$ MFCC	GMM	84 chansons (5 chanteuses, 5 chanteurs)	30ms	PI : 90.5% I+V : 75% Moy : 82.75
[RH07]	Comparaison des descripteurs	MFCC + Δ MFCC moyenne et stdev	SVM + post processing Cross-validation	46 chansons	25ms	85.1%

TAB. 2.1 – Tableau récapitulatif des différentes études

Chapitre 3

Test de référence

La méthode de référence choisie pendant ce rapport est l'approche dite "aveugle" consistant à extraire du signal des descripteurs audio "génériques" utilisés ensuite pour apprendre les deux classes de segments "chantés" et "non chantés" à l'aide d'algorithmes statistiques (approche dite *machine learning*).

Les descripteurs ainsi que les algorithmes utilisés ici sont les mêmes que ceux utilisés dans [Pee07] pour des tâches de segmentation parole/musique, de reconnaissance du "music genre" et du "music mood". Pour plus de détails nous nous référons à [Pee07].

Dans cet étude, le signal est d'abord ré-échantillonné à 11Khz et converti sur un seul par moyenne des deux canaux.

3.1 Matériel d'étude

La base de test utilisée pour ce projet à été gracieusement donnée par M Ramona. Cette base est constituée de 90 chansons libres de droit provenant du site internet Jamendo. Ces fichiers sont représentatifs de la musique commerciale actuelle et contiennent tous de la voix chantée. L'ensemble des chansons constitue environ 6 heures de musique annoté manuellement, par un seul annotateur, en catégories chantés (ch), non chantés (n).

Le matériel audio et les annotations utilisés sont décrits et disponibles à l'adresse suivante : http ://www.enst.fr/ ramona/icassp08/ . Les fichiers disponibles sur Jamendo sont au format Vorbis OGG codés en stéréo à la fréquence 44.1 kHz avec un bitrate de 112kB/s.

La base est divisé en trois sous-ensembles : 58 chansons sont consacrées à l'apprentissage, 16 à la validation et les 16 dernières aux tests.

3.2 Descripteurs utilisés

Pour chaque fichier, on extrait à chaque instant (trame d'analyse) du signal des descripteurs audio à l'aide d'une analyse à fenêtre glissante : fenêtre de type Blackman de taille égale à 40ms, pas d'avancement de 20ms (voir [Pee04]).

3.2.1 Mel Frequency Cepstral Coefficient (MFCC)

Les descripteurs MFCCs représentent une approximation de l'enveloppe spectrale à chaque instant. Ces coefficients sont calculés de la façon suivante :

- Calcul du spectre d'amplitude de chaque trame fenêtrée.
- Conversion du spectre d'amplitude en bande de Mel. Des études ont montré que l'oreille humaine se base sur une échelle fréquentielle spécifique appelée échelle des Mels. L'échelle des Mels est linéaire pour les basses fréquences (< 1000 Hz) et logarithmique pour les

hautes fréquences. La formule de conversion est la suivante :

$$mel(f) = \begin{cases} f & \text{if } f \leq 1000, \\ 1000.(1 + log_{10}\frac{f}{1000}) & \text{sinon.} \end{cases} \tag{3.1}$$

Ainsi on modélise le système auditif par un ensemble de filtres triangulaires. Nous utilisons ici 40 filtres.

– Pour finir, le spectre en échelle de Mel est convertit en cepstre par transformée de Fourier à cosinus discret (DCT). Les 13 premiers coefficients dont la composante continue, sont conservés pour décrire la forme de l'enveloppe spectrale.

3.2.2 Spectral Flatness Measure (SFM)

Les SFMs représentent le taux de composantes sinusoïdales (à travers une mesure de la "platitude" des bandes du spectre) présentes dans chaque bande de fréquence. Il est donné par le rapport de la moyenne géométrique et de la moyenne arithmétique du spectre :

$$SFM(bande) = \frac{(\prod_{k \in bande} a(k))^{1/K}}{\frac{1}{K}\sum_{k \in bande} a(k)} \tag{3.2}$$

où $a(k)$ représente l'amplitude du spectre à la fréquence k et K le nombre de fréquence k dans la bande.

Pour un signal tonal, la SFM est proche de zéro. Pour un signal bruité, la SFM est proche de 1. On calcule les SFM pour les 4 bandes de fréquence suivantes : (250 à 500 Hz), (500 à 1000 Hz) , (1000 à 2000 Hz) et (2000 à 4000 Hz).

3.2.3 Évolution temporelle des descripteurs

Nous calculons de plus les dérivées première et seconde de chaque descripteur (ΔMFCC, $\Delta\Delta$MFCC, ΔSFM et $\Delta\Delta$SFM) afin de décrire leur évolution temporelle.

Dans un second temps, nous effectuons une modélisation de l'évolution temporelle de chacun des descripteurs (analyse de type "texture window") à l'aide d'une seconde analyse à fenêtre glissante de taille 500ms et de pas d'avancement 166ms. Sur chacune de ces fenêtres nous estimons la valeur moyenne et l'écart type de chaque descripteur.

Ceci conduit à un ensemble de 102 descripteurs audio.

3.3 Classifieur utilisé

L'apprentissage est effectué en assignant chacune des trames d'analyse (longueur de 500ms) à une des classes "chantés" ou "non chanté" à partir de l'annotation fournie. Remarquons que la base d'entraînement est bien équilibrée (42762 données pour la classe "chantée", 40702 données pour la classe "non chantée").

L'apprentissage est effectué sur le sous-ensemble "d'entraînement" de la base, les résultats sont donnés pour les 16 chansons de la partie "test" (la partie "validation" n'est pas utilisée ici).

L'apprentissage comporte les étapes suivantes :

– Normalisation des descripteurs par "l'inter-quantile-range" (IQR) de leur valeur sur l'ensemble de la base d'entraînement.
– Retrait des descripteurs d'IQR nul.
– Sélection automatique des descripteurs par l'algorithme "Inertia Ratio Maximization with Feature Space Projection" (IRMFSP) [Pee03] : nous gardons les 40 premiers descripteurs
– Analyse linéaire discriminante (LDA) sur les deux classes.
– Modélisation de chacune des classes par mélange de gaussienne à 8 composantes avec diagonale pleine.

3.3.1 Modèle : Mélange de Gaussienne

Pour chaque classe un modèle de mélange de Gaussienne (GMM) est utilisé pour modéliser la distribution des données dans un espace de descripteur à D dimensions. Cet espace est obtenu par combinaison linéaire de plusieurs fonctions de densité de probabilité (PDF) à D dimensions qui permet d'approximer la forme globale de la collection de descripteurs. La probabilité d'observer le vecteur de descripteur x sachant le modèle multi-gaussien défini par ses paramètres $\lambda = \{\mu, \Sigma\}$ est alors :

$$p(x|\lambda) = \sum_{i=1}^{M} p_i g_i(x) \tag{3.3}$$

où

- x est un vecteur de descripteur à D dimensions,
- p_i les poids associés aux différentes gaussiennes et
- $g_i(x)$ une densité de probabilité définie par :

$$g(x) = \frac{1}{(2\pi)^{D/2}|\Sigma^{1/2}|} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)} \tag{3.4}$$

où μ représente la moyenne estimée et Σ la matrice de covariance.

Les paramètres des densités de probabilité des gaussiennes sont estimés par l'algorithme EM (Expectation-Maximization) dont la description peut être trouvée dans [Moo96].

Le GMM de la classe "chantée" (λ_{ch}) est entraîné sur les descripteurs extraits des sections annotées comme chantées et le GMM de la classe "non chantée" (λ_n) sur les segments purement instrumentaux.

3.3.2 Critère de classification

Le critère de classification retenu pour la classification est le critère de maximum a posteriori. Suivant ce critère, tout vecteur d'observation x est affecté à la classe qui maximise le probabilité d'observation sachant le modèle λ associé. Ici, $classe(x) = \{i|p(x|\lambda_i)\} = \max\{p(x|\lambda_{ch}), p(x|\lambda_n)\}$

3.4 Résultats

Les résultats suivants ont été obtenus :

- Recall : 0.758
- Precision : 0.7580
- F-measure : 0.7581

Nous avons ensuite testé l'application d'un filtrage median (sur 21 points) sur le flux temporel d'assignation des classes. Celui-ci permet d'augmenter légèrement le taux de reconnaissance :

- Recall : 0.774
- Precision : 0.775
- F-measure : 0.774

Remarquons que ces résultats varient peu selon le nombre de gaussiennes utilisés.

Chapitre 4

Critère de voisement ou d'harmonicité

La production vocale en général est une succession de sons voisés et non voisés. Les sons voisés, comme les voyelles, possèdent une hauteur. Cette hauteur est donnée par la vitesse de vibration des cordes vocales. Les sons non-voisés, comme certaines consonnes, sont produits par la turbulence de l'air contre les lèvres ou la langue. Ils s'apparentent à du bruit et ne possèdent pas de hauteur. Il existe également des consonnes voisées comme ('m', 'd' ...).

De nombreuses recherches, visent à segmenter un signal de parole en parties voisées et non voisées. Cette distinction, nécessaire pour la synthèse de parole, est réalisée à l'aide d'un coefficient de voisement. Ce coefficient mesure la périodicité du signal autour d'un temps donné dans une bande de fréquence donnée. Il permet de décrire le caractère périodique des sons voisés ou le caractère aléatoire des sons non-voisés. On trouve dans [Rd96] un résumé des méthodes d'estimation du coefficient de voisement. Ces méthodes mesure le rapport harmonique/bruit.

Les méthodes de segmentation en parties voisées et non voisées peuvent être appliquées à la voix chantée. Dans le chant, 90 % des sons utilisés sont voisés puisque ce sont les seuls à posséder une fréquence fondamentale. Les sons non-voisés sont également présent mais sont très brefs.

Les sons voisés ont la particularité d'être harmoniques, c'est à dire que les multiples de la fréquence fondamentale sont présents. La caractéristique de voisement ou de l'harmonicité a été exploitée à plusieurs reprises pour détecter la présence de chant dans la musique. On trouve ici quelques approches proposées dans la littérature.

4.1 Coefficient d'harmonicité de Kim and Whitman

4.1.1 Description

Dans [KW02], les auteurs utilisent un filtre en peigne inverse pour détecter les quantités élevées d'harmonique.

Filtre en peigne

Un filtre en peigne est utilisé en traitement du signal pour ajouter une version retardée du signal à lui même provoquant ainsi des interférences destructives ou constructives. La structure d'un filtre en peigne inverse peut être décrite avec l'équation différentielle suivante :

$$y(n) = x(n) - gx(n - M) \tag{4.1}$$

où M est la longueur du retard (en échantillon) et g le facteur de gain appliqué au signal retardé.

La fonction de transfert de ce filtre est donnée par :

$$H(z) = \frac{Y(z)}{X(z)} = 1 - gz^{-M} \tag{4.2}$$

La réponse en fréquence du filtre se présente sous la forme de pics régulièrement espacés, d'où la nomination de filtre en peigne. L'espacement des pics, la largeur des filtres dépend de la fréquence et donc du retard. Le schéma suivant représente la réponse en fréquence de deux filtres destinés à atténuer toutes les harmoniques d'un son à 400 Hz et 800 Hz respectivement c'est à dire toutes composantes spectrales se trouvant à des multiples de 400 (400, 800, 1200) ou de 800 (800, 1600, 2400).

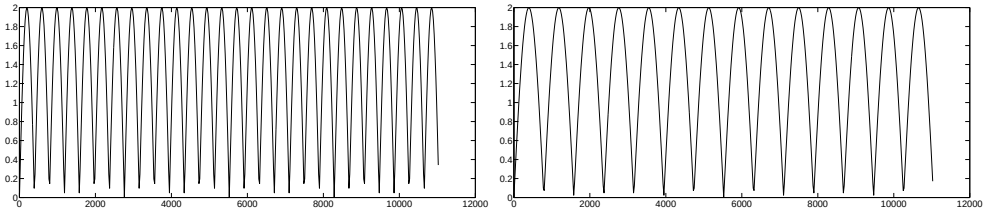


FIG. 4.1 – Filtre en peigne inverse (400 Hz, 800 Hz)

Application

Imaginons que l'on dispose du signal d'un son harmonique dont la fréquence fondamentale est à 100 Hz. Le spectre de ce signal présente des pics à 100, 200, 300 Hz ...Lorsque le signal est filtré par un filtre en peigne dont le retard correspond à 100 Hz toutes les composantes de ce signal seront atténuées. Il ne reste alors qu'une très faible énergie dans le signal filtré. Le rapport de l'énergie du signal initial sur l'énergie du signal filtré est donc élevé. A l'inverse, si l'on filtre un bruit blanc par le même filtre le rapport de l'énergie du signal initial sur l'énergie du signal filtré sera proche de 1 puisque le filtrage n'aura pas apporté de changements significatifs.

Mise en oeuvre

Ce procédé est utilisé dans [KW02] de la façon suivante : Un ensemble de filtre en peigne, dont les délais varient, permet de trouver la fréquence pour laquelle le signal est le plus atténué. Cette fréquence est considérée comme la fréquence fondamentale du signal à traiter. En divisant l'énergie du signal le plus atténué par l'énergie du signal original on obtient une mesure d'harmonicité notée H_c. Dans [KW02], le coefficient proposé est l'inverse de celui-ci. Afin d'obtenir une mesure borné, nous avons préféré définir le coefficient d'harmonicité de la façon suivante.

$$H_c = \frac{min_i(E_{filtre,i})}{E_{initial}} \tag{4.3}$$

Plus H_c est proche de zéro plus le signal est harmonique, plus il est proche de 1 moins il est harmonique. En comparant la valeur trouvée à un seuil on détecte les sons harmoniques.

4.1.2 Évaluation du descripteur

Afin de rendre compte de l'efficacité de ce descripteur on effectue une série de tests sur des sons synthétiques et réels.

Sons synthétiques

Observations :

– On constate que le coefficient d'harmonicité est élevé pour du bruit (cas 5) comparé à celui trouvé pour les sons harmoniques (cas 1 2 3 4). On note qu'il est très sensible à la présence de bruit dans un son harmonique (cas 6).

Cas	1	2	3	4	5	6
Son	Son harmonique à 440 Hz : 3 harmoniques	Son harmonique à 440 Hz : 5 harmoniques	Son harmonique à 440 Hz vibré 10 Hz : 3 harmoniques	Son harmonique à 440 Hz vibré 10 Hz : 5 harmoniques	Bruit blanc	Son harmonique vibré (3 harmoniques) + Bruit SNR = 31 dB
Énergie Initiale	7.01	27.58	7.01	27.58	0.96	7.27
Énergie après filtrage	0.17	0.7	0.28	1.55	0.46	0.45
H_c	0.024	0.026	0.04	0.057	0.479	0.062

TAB. 4.1 – Évaluation coefficient d'harmonicité

– Pour un nombre d'harmonique fixé on obtient des valeurs de H_c légèrement plus élevées pour les sons non-vibrés. Ceci est du au fait que la fréquence varie légèrement en présence de vibrato. Cette remarque nous amène à effectuer l'analyse sur des segments qui comportent au plus une période du vibrato.

Sons réels

Ce détecteur d'harmonicité a ensuite été testé sur un morceau de musique où la voix et sas harmoniques sont facilement visibles sur le spectrogramme. Il s'agit d'un extrait d'une pièce de L. Berio, chantée par C. Berberian, "I wonder as I wander". L'harmonicité à été calculée sur des segments d'une durée de 60 ms. Les valeurs ont ensuite été multipliées par un coefficient avant d'être superposés au spectrogramme afin d'être plus facilement lisibles (voir figure 4.2).

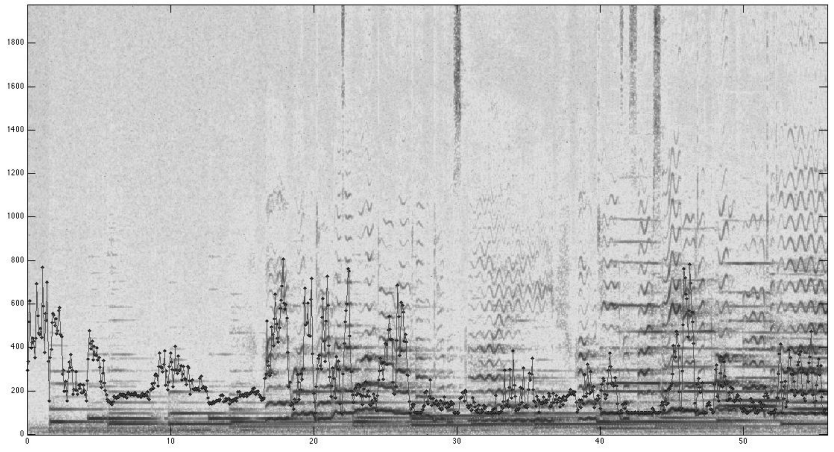


FIG. 4.2 – Illustration du coefficient d'harmonicite

Observations :
– On constate tout d'abord que ce coefficient est cohérent, qu'il suit la structure harmonique du morceau.
– Cependant ce descripteur ne détecte pas toutes les zones chantées de la chanson. Particulièrement aux endroits où sont superposé deux ensembles d'harmoniques.

– Ce coefficient ne permet évidement pas de discriminer les caractéristiques harmonique du chant de celle des autres instruments de musique harmonique.

Améliorations :

– La recherche de la fréquence fondamentale de chaque segment est extrêmement longue, il serait préférable de procéder d'abord à la détection de fréquence fondamentale puis d'effectuer les filtrages.
– Une recherche de "multi-pitch" permettrai de filtrer plusieurs fois les segments polyphoniques et résoudrai le problème de superposition de deux ensembles d'harmoniques non harmonique entre eux.

4.2 Coefficient d'harmonicité de Chou et Gu

Dans [CG01], les auteurs proposent une autre approche pour détecter les caractéristiques de la structure harmonique de la parole voisée. Ils définissent un coefficient d'harmonicité calculé comme la moyenne du maximum de l'auto-corrélation dans le domaine temporelle et dans le domaine fréquentiel. Cette approche avait déjà été utilisée dans [CKK98] pour améliorer l'estimation de fréquence fondamentale de la voix parlée.

4.2.1 Description

On rappelle que l'auto-corrélation est un outil utilisé en traitement du signal pour détecter des régularités, des événement répétés dans un signal. L'auto-corrélation d'un signal peut aussi bien être mesuré sur le signal temporel que sur le spectre d'amplitude du signal.

Afin d'améliorer la performance de la détection d'événements répétés les corrélations spectrale et temporelle peuvent être combinées pour construire une auto-corrélation spectro-temporelle. Notons $R^T(\tau)$ l'auto-corrélation dans le domaine temporel et $R^S(\sigma)$ l'auto-corrélation dans le domaine fréquentiel, la corrélation spectro-temporelle est définie par :

$$R(\tau) = \frac{R^T(\tau) + R^S(\tau\prime)}{2} \tag{4.4}$$

où $\tau\prime = \frac{1}{\sigma}$

Les auteurs définissent le coefficient d'harmonicité par :

$$H_c = max_\tau R(\tau) \tag{4.5}$$

où τ représente le décalage temporel.

Calcul de l'auto-corrélation temporelle par transforme de Fourier inverse

La corrélation temporelle d'un signal (ACF) mesure la corrélation du signal décalé avec lui même. Il existe plusieurs façon de calculer l'auto-corrélation dans le domaine temporelle, nous choisissons une méthode proposée dans [Pee06] basée sur la transformée de Fourier inverse du spectre en puissance. Comme le spectre en puissance est réel et symétrique sa transformée de Fourier inverse est par définition réelle. Pour alléger les calculs on considère que l'estimateur de l'ACF $\tilde{R(l)}$ est une projection du spectre en puissance sur une base de cosinus dont les fréquences sont égales aux décalages (lag) $\tau_l = l/sr_{hz}$ où sr_{hz} est la fréquence d'échantillonnage, τ_l le décalage temporel et l la fréquence. Ainsi l'estimateur non biaisé de l'ACF est donné par :

$$\tilde{R}(l) = \frac{1}{N-l} \sum_{k=0}^{N-k-1} (|X(k)|^2) cos(2\pi k \frac{l}{N}) \tag{4.6}$$

où $|X(k)|^2$ représente le spectre en puissance du signal et N le nombre de point sur lequel il est calculé. On calcul donc la périodicité des pics du spectre.

Les valeurs positives de $\tilde{R}(\tau_l)$ apparaissent pour les valeurs τ_l pour lesquels les valeurs positives du cosinus coïncident avec les pics du spectre.

Auto-corrélation spectrale

L'auto-corrélation du spectre d'amplitude mesure la périodicité de la position des pics du spectre en utilisant la fonction d'auto-corrélation. L'auto-corrélation à la fréquence k du spectre X est estimé par calcul suivant :

Moyenne des deux corrélations

Il faut faire correspondre les décalages temporels aux décalages fréquentiels. Tout d'abord les corrélations sont normalisées par rapport au premier coefficient :

$$\tilde{R}(k) = \frac{\tilde{R}(k)}{\tilde{R}(0)} \tag{4.7}$$

A l'axe temporel des τ on fait correspondre l'axe fréquentiel tel que $k_\tau = \frac{1}{\tau}$. On connaît les τ pour les valeurs allant de $\frac{1}{N_{fft}} \dots \frac{sr_{hz}}{2}$. En inversant cet axe on obtient les valeurs de l'axe fréquentiel pour les Ainsi pour connaître tous les décalages fréquentiels il est nécessaire d'utiliser une FFT calculé pour un nombre très élevé de points .

4.2.2 Évaluation du descripteur

Sons de synthèse

Les schémas suivant illustrent la calcul de l'auto-corrélation spectro-temporelle sur des signaux de synthèse. L'auto-corrélation temporelle est représentée par la courbe en trait longs, l'auto-corrélation spectrale en pointillés et la moyenne des deux par la courbe en trait plein. Enfin, la valeur de l'harmonicité et son emplacement sont matérialisés par un rond. Les signaux utilisés sont les même que ceux de l'étude précédente.

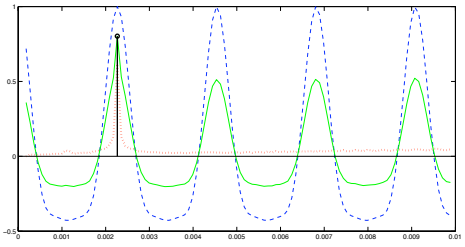


FIG. 4.3 – Corrélation spectro-temporelle d'un signal de synthèse composé de 3 harmoniques

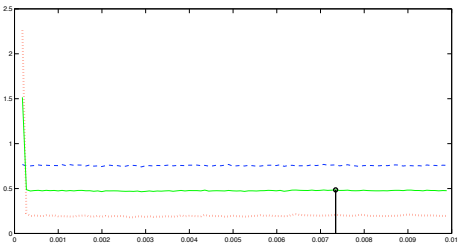


FIG. 4.4 – Corrélation spectro-temporelle d'un signal de synthèse (bruit)

On reporte dans le tableau suivant les valeurs du coefficients d'harmonicité pour les signaux traités :

Plus le coefficient est proche de 1 plus le signal est harmonique.

Observations :

- On constate que le coefficient est deux fois plus faible pour du bruit (cas 5) que pour des sons harmoniques (cas 1 2 3 4). Par ailleurs, comme le montre la quatrième figure il est relativement peu sensible à la présence de bruit (cas 6).
- Cas des sons harmoniques : La présence de vibrato n'influe pas sur le coefficient.

Sons réels

Ce coefficient à également été testé sur des sons réels. Nous observons sur la figure suivante les résultats obtenus pour un son harmonique et pour un son non harmonique :

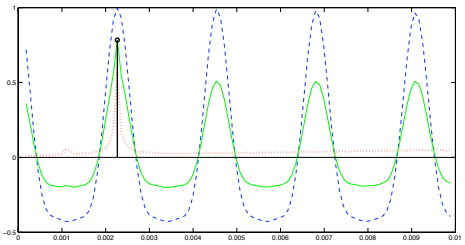


FIG. 4.5 – Corrélation spectro-temporelle d'un signal de synthèse composé de 3 harmoniques avec vibrato

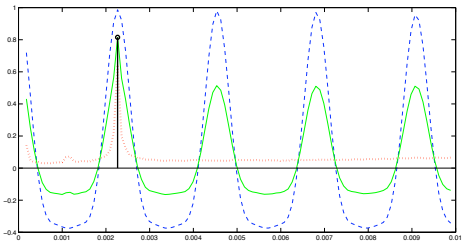


FIG. 4.6 – Corrélation spectro-temporelle d'un signal de synthèse composé de 3 harmoniques avec vibrato + bruit

Cas	1	2	3	4	5	6
Son	Son harmonique à 440 Hz : 3 harmoniques	Son harmonique à 440 Hz : 5 harmoniques	Son harmonique à 440 Hz vibré 10 Hz : 3 harmoniques	Son harmonique à 440 Hz vibré 10 Hz : 5 harmoniques	Bruit blanc	Son harmonique vibré (3 harmoniques) + Bruit
H	0.82	0.89	0.80	0.88	0.43	0.83

TAB. 4.2 – Évaluation coefficient d'harmonicité par corrélation spectro-temporelle

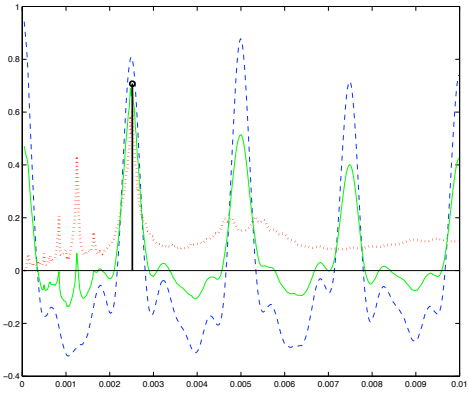


FIG. 4.7 – Corrélation spectro-temporelle d'un signal harmonique

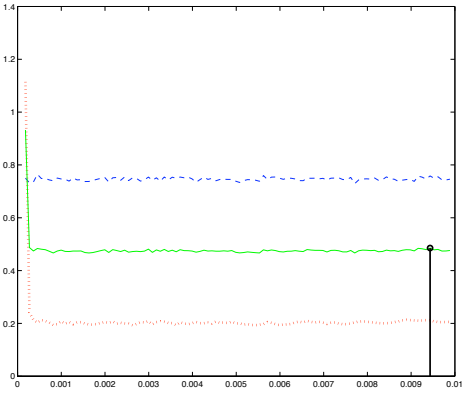


FIG. 4.8 – Corrélation spectro-temporelle d'un signal non harmonique

Cette méthode a également été testée sur le même morceau de musique que précédemment. Les coefficients d'harmonicité ont été centré puis multiplié par 10000 avant d'être reporté sur le spectrogramme. Le calcul du coefficient dharmonicité par l'auto-correlation spectro-temporelle est extrêmement rapide.

4.2.3 Mise en application

Pour chaque instant d'analyse du signal de chaque fichier audio de la base de test on calcul le coefficient d'harmonicité. Les informations sont extraites à l'aide d'une analyse à fenêtre glissante de type Hanning de durée égale à 60ms avec un pas d'avancement de 20ms. On répertorie les résultats trouvés pour la classe chantée (figure 4.10) et la classe non chantée (figure 4.11)

Les histogrammes cumulés montrent qu'il n'existe pas de différence significative entre les

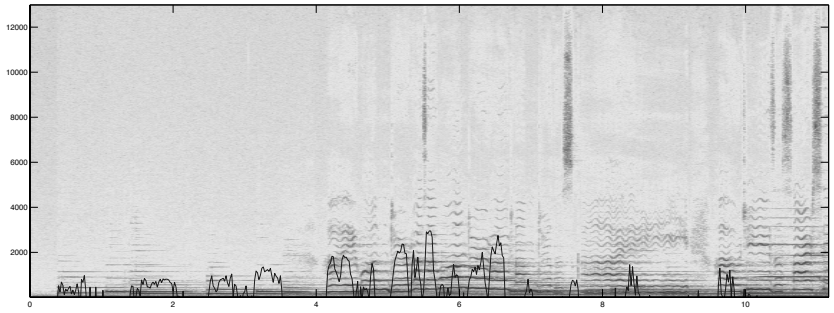


FIG. 4.9 – Illustration du coefficient d'harmonicité par corrélation spectro-temporelle

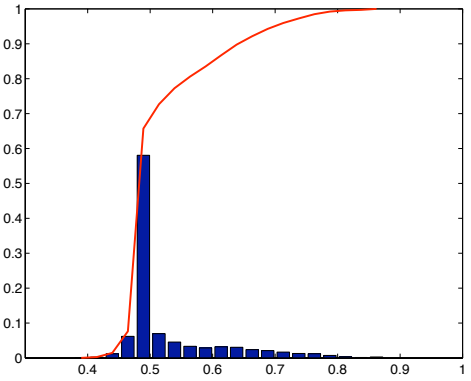


FIG. 4.10 – Coefficients d'harmonicité pour la classe chant

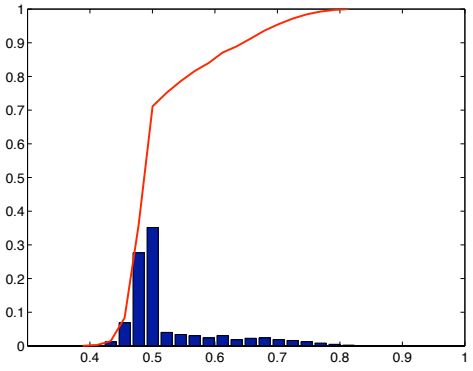


FIG. 4.11 – Coefficients d'harmonicité pour la classe non chant

valeurs trouvées pour chacune des classes. On pouvait prédire ce résultat puisque de nombreux instruments de musique possède aussi la particularité d'être harmonique. Ce descripteur pourra néanmoins être utilisé s'il est combiné à d'autres dans un modèle d'apprentissage.

4.3 Coefficient de voisement de superVP

Un autre critère de voisement à été testé sur le même morceau de musique à l'aide du logiciel superVP de l'IRCAM. Le coefficient de voisement à pour fonction de caractériser l'activité des cordes vocales, il est équivalent au coefficient d'harmonicité puisque les vibration des cordes vocales produisent un signal quasi-périodique. Le coefficient de voisement est une mesure à un instant donné pour une bande de fréquence choisie de l'énergie expliquée par la périodicité du signal.

4.4 Description

Le critère de voisement de superVP est calculé de la façon suivante.

- Classification de chacun des pics en terme de pics sinusoïdaux et non sinusoidaux : Pour une bande de fréquence donnée à un instant précis on mesure l'énergie expliqué par la fréquence fondamentale.
- La classification est ensuite amélioré grace à l'utilisation d'un seuil basé sur le rapport du signal au bruit.

Cette analyse est contrôlée par plusieurs paramètres :

- Un seuil d'énergie normalisée par bande permet de définir un critère à partir duquel la bande sera considérée comme voisée.
- Un autre seuil permet de contrôler la classification des pics dans les deux classes voisée et non voisée. Plus le seuil est élevé moins il y aura de pics considérés comme non voisés. Plus le seuil est faible, moins il y aura de pics non voisés déclarés comme voisés.
- Il est également possible de contrôler la largeur des bandes pour la classification des pics en terme de largeur des bandes des sinusoïdes stationnaires.

On trouvera une description précise dans [**?**].

4.5 Résultats

On peut voir sur la figure suivante le coefficient de voisement calculé par audiosculpt.

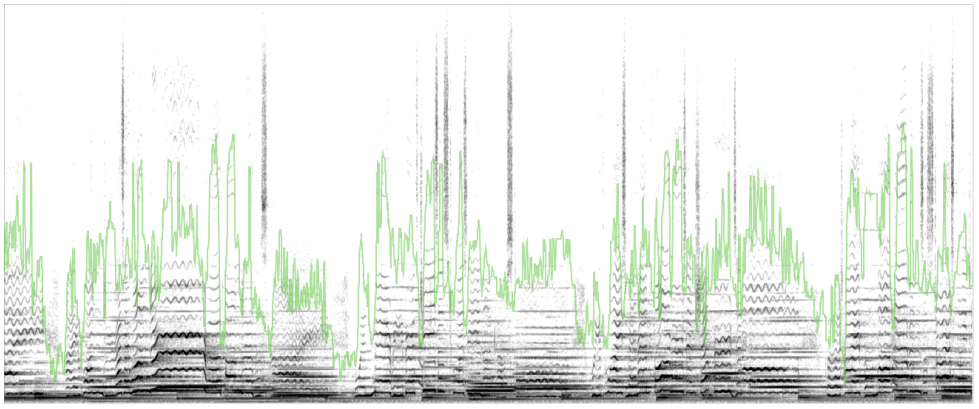


FIG. 4.12 – Coefficient de voisement d'audiosculpt

On remarque que ce coefficient suit particulièrement bien les harmoniques du morceau utilisé. A partir de l'ensemble des valeurs lissées (filtre médian) on propose de segmenter le flux en fonction du voisement.

Chapitre 5

Vibrato de la voix chantée

5.1 Le vibrato

Pour une note donnée, le vibrato correspond à une variation quasi-périodique de f_0 autour de sa fréquence centrale. Cette variation dépend de la nature de l'instrument et de la technique de l'intrumentiste. Le vibrato a une fonction esthétique et expressive liée à une culture et une époque.

Le vibrato a été introduit au 17 ième siècle comme un ornement pour mettre de l'emphase sur certaines notes. Au 19 ième le vibrato émerge sous une forme plus continue pour devenir un attribut du timbre. Cet effet de timbre, contrôlé par le musicien, est aujourd'hui utilisé par la plupart des instrumentistes pour imiter le vibrato naturel de la voix [VGD05].

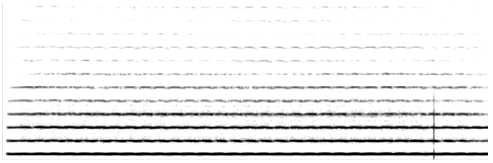
5.1.1 Instruments à cordes

Pour les instruments à cordes, le vibrato est obtenu en faisant légèrement varier la position du doigt autour de sa position centrale. Le changement de longueur de la corde fait ainsi varier la fréquence fondamentale. La modulation de fréquence est appelé par les musiciens vibrato. Par ailleurs le mouvement du doigt ajoute un faible taux d'énergie : la note peut durer sans être entretenue (exemple de la guitare). Cette modulation d'énergie est une modulation d'amplitude.

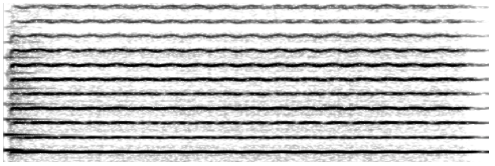
5.1.2 Instruments à vent

Pour les instruments à vents le vibrato est obtenu en modulant le flux d'air : ceci entraîne une modulation d'amplitude (AM) et une modulation de fréquence (FM). La modulation du flux d'air peut être obtenue de différentes façons en fonction des instruments. Pour le saxophone l'instrumentiste peut vibrer en modulant la pression sur l'anche ou en modulant la pression dans la bouche.

Lorsqu'il s'agit d'une modulation d'amplitude (ou modulation d'intensité) on parle de trémolo. Les musiciens appellent souvent ce type de modulation vibrato d'intensité ce qui peut prêter à confusion.



Trémolo : flute



Vibrato : violon

5.1.3 Voix chantée

Pour la voix, la présence de vibrato implique celle de trémolo. En effet, l'intensité n'est pas maintenue constante et la modulation périodique concerne à la fois la hauteur et l'intensité dans des proportions variables selon la technique.

Pour la voix chantée, le vibrato est due aux modulations du flux d'air par la source glottique couplé aux modulations de résonances [SR90] : les variations de fréquence (FM) générées par la source glottique modifient le timbre et l'amplitude (AM). Les modulations de résonance sont aussi responsables de la modulation d'amplitude. Pour conclure, un vibrato produit par de la voix est une modulation de fréquence qui entraîne nécessairement une modulation d'amplitude. Bien que le vibrato soit naturel il doit être travaillé pour être maîtrisé. Il s'agit de contrôler ses paramètres l'amplitude (ou l'étendue) et sa fréquence (ou pulsation).

La figure suivante 5.1 illustre un vibrato de voix.

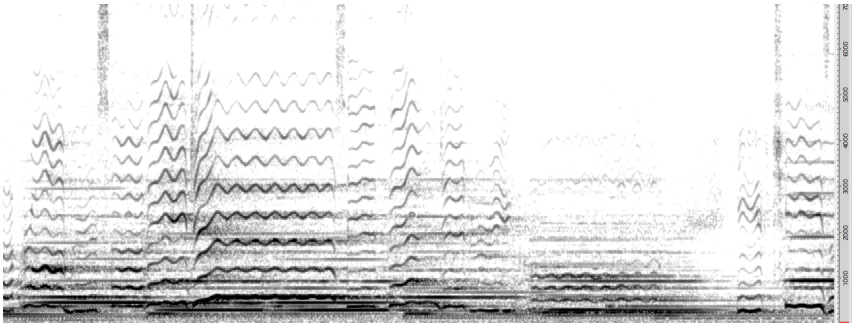


FIG. 5.1 – Vibrato voix

En résumé, le vibrato au sens global du terme est due à au moins l'une de ces modulations :
- Modulation de fréquence (prédominant dans les instruments à cordes et la voix)
- Modulation d'amplitude (prédominant dans les instruments à vent et les cuivres)

5.1.4 Paramètres du vibrato

- **La fréquence :** Pour les instruments de musique la fréquence de vibrato est comprise entre 4 et 12 Hz [DH89] et augmente à la fin des notes. L'accroissement en fin de note de la fréquence du vibrato est particulièrement important chez les chanteurs . Pour la voix chantée la moyenne est d'environ 6 Hz avec une variation d'environ 8% [Pra94]. Lorsque la fréquence de vibration est plus faible les professeurs de chant parlent de modulation : le son émis n'est pas perçu comme du vibrato. Quand la fréquence de vibration est élevée on parle de vibrato serré [GHD+05].
- **L'amplitude :** L'amplitude ou l'étendue du vibrato s'étend de 0.6 à 2 demi tons pour les chanteurs et de 0.2 à 0.35 pour les instruments à cordes [TD00]. En dessous de 0.2 ton le son est mat et dur, au dessus d'un ton le son devient instable. Sundberg à montré [BS03] que l'étendue du vibrato est corrélé au niveau sonore.

5.2 Formalisation du modèle de vibrato

Soit un signal définit, selon le modèle additif, comme une somme de sinusoïdes modulées [MQ86] autour d'un instant τ sous l'hypothèse de stationnarité locale ce signal à l'instant n s'écrit :

$$s(n) = \sum_{h=1}^{H} a_{h,\tau}.\cos(\phi_h(n)) \tag{5.1}$$

ou

– H représente le nombre de partiel
– a_h : l'amplitude de modulation du h-ième partiel
– $\phi_h(n)$: la phase instantanée du h-ième partiel

La phase à l'instant n est donnée par l'intégrale de la fréquence instantanée $f_{h,\tau}$:

$$\phi_h(n) = \phi_h(n-1) + 2\pi \frac{f_{h,\tau}}{sr_{hz}} \tag{5.2}$$

où $\phi_h(0)$ représente la phase initiale et sr_{hz} la fréquence d'échantillonnage.

Le vibrato étant considéré comme une variation périodique il peut être décomposé en série de fourier. Dans le cas non stationnaire, les amplitudes $a_{h,\tau}$ et les fréquences $f_{h,\tau}$ peuvent s'exprimer sous la forme de sommes de sinusoïdes [MR04] : L'amplitude modulé est donnée par :

$$\tilde{a}_h(n) = \sum_{l=1}^{L_a} \tilde{a}_l^a(n) \cos((\phi_l^a(n)) \tag{5.3}$$

La fréquence modulée est donnée par :

$$\tilde{f}_h(n) = \sum_{l=1}^{L_f} \tilde{a}_l^f(n) \cos((\phi_l^f(n)) \tag{5.4}$$

où L_a et L_f représente le nombre de composantes de l'amplitude et de la fréquence respectivement. Lorsque le vibrato est réduit à une modulation d'amplitude (AM) on a : $\tilde{f}_h(n) = f_{h,\tau}$.

Lorsque le vibrato est réduit à une modulation de fréquence (FM) on a : $\tilde{a}_h(n) = a_{h,\tau}$.

5.3 Comportement des fréquences et amplitudes des harmoniques d'un son vibré

On considère des sons instrumentaux, lorsque l'on parle d'harmoniques on parle en fait des partiels qui peuvent être harmoniques ou quasi-harmoniques.

Modulation d'amplitude Quand seule la modulation d'amplitude est présente, les fréquences de chaque partiels sont inchangées alors que les amplitudes ont une pulsation. L'amplitude de la modulation est proportionnelle à l'amplitude du partiel.

Modulation de fréquence Quand seule la modulation de fréquence est présente les fréquences de toutes les partielles ont la même pulsation. L'ambitus (ou amplitude) du vibrato de chaque partielle est proportionelle à la fréquence du partiel.

Modulation de fréquence et d'amplitude Dans le voix chantée les deux modulations sont simultanées. Elles peuvent être synchrones ou en opposition de phase mais ont la même pulsation.

5.4 Détection de vibrato

En supposant que la voix chantée est naturellement vibrée et que ce vibrato est constitué à la fois d'une modulation de fréquence et d'une modulation d'amplitude on cherche à détecter dans un signal les composantes sinusoïdales qui comportent de telles modulations. On peut extraire les paramètres du vibrato à l'aide du signal analytique.

5.4.1 Paramètres du vibrato à partir du signal analytique

Une onde modulée en amplitude ou en fréquence est décrite dans le plan complexe par un vecteur tournant dont le module et la vitesse de rotation sont des grandeurs variables au cours du temps. A partir d'un signal réel $x(t)$ on peut le transformer dans le plan complexe à l'aide de la transformée de Hilbert $z_x(t)$:

$$z_x(t) = x(t) + i(Hx)(t) \tag{5.5}$$

où

- x représente le signal réel
- Hx la transformée de Hilbert de x. La partie imaginaire est une version du signal original dont la phase est décalée de 90 °. Les sinus sont transformés en cosinus et vice-versa. La transformée de Hilbert contient les mêmes amplitudes et fréquences que le signal original mais inclut une information sur la phase.
  L'amplitude et la phase se déduisent de cette représentation complexe appelée signal analytique.

  L'amplitude instantanée est donnée par le module du signal analytique :
  $A_x(t) = |z_x(t)|$

  La fréquence instantanée, dérivée de la phase instantanée est donnée par :
  $f_x(t) = \frac{1}{2\pi}\frac{\delta}{\delta t} arg(z_x(t))$

5.4.2 Détection des partiels correspondant à la voix chantée basé sur la modulation d'amplitude et de fréquence

On considère que les partiels correspondant à la voix chantée sont ceux qui possèdent une modulation de fréquence et d'amplitude. A partir de tous les partiels extraits du signal on effectue une sélection à partir de seuils des partiels qui comportent de tels modulations. Tous les autres partiels sont supprimés. Ainsi seuls les segments comportants des partiels après cette transformation sont classés comme chantés.

Description

1. Extraction des partiels à l'aide de superVP pm2 :
   A chaque instant d'analyse n (ici 16ms), extraction de la fréquence $f_s(n)$, de l'amplitude $a_s(n)$ et de la phase $\phi_s(n)$ de toutes les composantes sinusoïdales de la trame analysée.
   Les valeurs trouvées au cours de l'analyse sont interpolées afin de suivre l'évolution des différentes composantes que l'on nomme partiels.
   Pour un partiel p_k commençant à l'instant n_i et finissant à l'instant n_j les vecteurs f_{p_k}, a_{p_k} et ϕ_{p_k} sont nuls pour tous les instants n en dehors de l'intervalle $[n_i, n_j]$. Entre les instants n_i et n_j ces vecteurs contiennent les différentes valeurs (non nulles) que peut prendre le partiel en fréquence, amplitude et phase respectivement. On note L la durée du partiel en nombre d'échantillon. On a $L = n_j - n_i$.

2. Sélection des partiels vibrées (modulation de fréquence) :
   On s'intéresse ici à la modulation de fréquence, on analyse donc pour chaque partiel p_k les valeurs de f_{p_k}. La transformée de Fourier F_{p_k} de f_{p_k} est donné par :

$$F_{p_k}(f) = \sum_{n=n_i}^{n_j} f_{p_k}(n) e^{\left(-2i\pi f \frac{n}{L}\right)} \tag{5.6}$$

Étendue du vibrato Soit p_k un partiel vibré de fréquence centrale f_0 et d'étendue Δf_{p_k} alors $f_{p_k} \in \left[f_0 - \frac{1}{2}\Delta f_{p_k}, f_0 + \frac{1}{2}\Delta f_{p_k}\right]$. Soit p_l le partiel correspondant à sa première harmonique, de fréquence centrale $2f_0$ alors $f_{p_l} \in [2f_0 - \Delta f_{p_k}, 2f_0 + \Delta f_{p_k}]$.
Pour extraire l'amplitude relative du vibrato ($\Delta f/f_0$) on procède de la façon suivante :

- Afin de ne conserver que les valeurs de variation de fréquence on calcul la transformée de Fourier sur les valeurs de f_{p_k} centrées.

$$F_{p_k}(f) = \sum_{n=n_i}^{n_j} (f_{p_k}(n) - \mu_{f_{p_k}}) e^{-2i\pi f \frac{n}{L}} \tag{5.7}$$

  où $\mu_{f_{p_k}} = \frac{1}{L}\sum_{n=n_i}^{n_j} f_{p_k}(n)$ correspond à la fréquence centrale estimée du partiel.
- L'étendue de la modulation en Hz pour toutes les fréquences de modulation est obtenue par le calcul suivant :

$$\Delta f_{p_k}(f) = \frac{F_{p_k}(f)}{L} \tag{5.8}$$

- L'étendue relative du vibrato est donnée par

$$\Delta f_{relp_k}(f) = \frac{\Delta f_{p_k}(f)}{\mu_{p_k}} \tag{5.9}$$

- Pour finir on détermine l'amplitude relative du vibrato autour de 6 Hz comme la valeur maximale de $\Delta f_{relp_k}(f)$ entre 4 et 8 Hz.

$$\Delta f_{p_k} = \max_{f \in [4,8]} \Delta f_{relp_k}(f). \tag{5.10}$$

Notons P_{vib} l'ensemble des partiels considérés comme vibrés.

$$p_k \in P_{vib} \Longleftrightarrow \Delta f_{p_k} \geq seuil_{vib} \tag{5.11}$$

3. Sélection des partiels harmoniques.
   En partant du constat que la présence de voix chantée est matérialisé par la présence de multiples partiels harmoniques vibrées on effectue une seconde sélection selon le critère suivant : un partiel vibré p_k est conservé si et seulement si il existe un autre partiel vibré au même instant. Notons P_{hvib} l'ensemble des partiels considérés comme vibrés et harmoniques. Soit $p_k \in P_{vib} | p_k(n) \neq 0, \forall n \in [n_i, n_j]$. Alors

$$p_k \in P_{hvib} \Longleftrightarrow \exists p_l \in P_{vib}, \exists n \in [n_i, n_j] \, | p_l(n) \neq 0 \tag{5.12}$$

   On pourra remplacer cette sélection par le critère suivant : Un partiel vibré est conservé si et seulement si il existe un autre partiel vibré au même instant dont la fréquence centrale est un multiple de celle du premier (harmonicité).
4. Sélection des partiels avec modulation d'amplitude
   Comme la modulation de fréquence entraîne une modulation d'amplitude pour la voix à cause des modulations de résonance du conduit vocal on décide de conserver uniquement les partiels qui présentent les deux types de modulation. Cette étape permet d'éliminer les partiels qui pourraient être produits par des instruments à cordes par exemple. On effectue une analyse similaire à celle effectué pour détecter les partiels avec modulation d'amplitude mais sur les valeurs de l'amplitude de chaque partiel à chaque instant.
5. Détection des segments chantés
   On considère que les partiels sélectionnés correspondent à ceux produit par de la voix chantée. Tous les autres ayant été supprimés, les parties chantées sont celles ou il reste des partiels.

Chapitre 6

Post-Processing

Toutes les méthodes conduisent vers une segmentation que l'on pourrait caractérisée de trop précise. Lors de l'évaluation d'une méthode il s'agit de comparer la segmentation automatique à la segmentation de référence. Or la segmentation de référence est souvent assez grossière puisque elle ne prend pas en compte les courtes césures comme les respirations. Afin d'automatiser une segmentation qui correspondent à l'annotation manuelle il est nécessaire de lisser temporellement les résultats obtenus. Différentes stratégies peuvent être misent en oeuvre. Dans la littérature on trouve trois approches qui sont décrites ici. Une méthode très simple basée sur une longueur minimale de segment, a été proposée dans ce travail.

6.1 Filtrage ARMA

Lukashevich [LGD07] part du constat que la présence de voix chantée est une information qui tend à être continue sur plusieurs trames consécutives. Elle considère que la valeur de la décision pour une trame i est en partie déterminée par les k trames précédentes. Cette valeur peut être interprétée comme le résultat d'un processus auto régressif (AR). Par ailleurs, le lissage de la fonction de décision pour éliminer les segments trop courts peut être efficacement réalisé par un processus de moyenne mobile (AM). La combinaison de ces deux traitements peut être interprétée comme un processus ARMA. Ce traitement peut donc être approximé par la fonction de transfert correspondant à l'équation aux différences suivantes :

$$x_i = \sum_{l=1}^{p} b_l y_{i-l} - \sum_{k=0}^{q} a_k x_{i-k} \tag{6.1}$$

La relation entre l'entrée y_i (la segmentation trouvé par la méthode) et la sortie x_i (segmentation lissée) est donnée par la fonction de transfert $H(z) = \frac{B(z)}{A(z)}$ où $A(z)$ et $B(z)$ représentent les transformés en z des processus AR et AM respectivement. Pour finir le résultat filtré est lissé par une convolution avec une fenêtre de Hanning .

6.2 Filtrage médian

Ramona [RR08] montre que les probabilités à posteriori de l'évolution de la présence de voix chantée produisent une classification assez chaotique. De nombreux segments sont mal classés. La première idée pour améliorer le résultat est de lisser la classification en utilisant un filtre médian sur 30 trames consécutives qui correspond dans son étude à des segments de 0.5 seconde. Ce filtrage permet d'améliorer le résultat cependant les changements de classe apparaissent plus souvent dans les régions principalement chantés que dans les segments purement instrumentaux. Ce filtrage permet de passer de 71.6 % à 81.1% de bonne classification.

6.3 HMM

Pour améliorer encore la classification, Ramona propose un lissage temporel basé sur les probabilités a posteriori d'un modèle de Markov caché (HMM) à deux états (purement instrumental et chanté). Les distributions sont modélisées par mélanges de 5 gaussiennes dont les paramètres sont estimés par l'algorithme EM (Expectation Maximization). Le meilleur trajet entre les deux états est obtenu par l'algorithme de programmation dynamique de Viterbi. Le résultat de ce post-traitement présente toujours des trames mal classées surtout aux frontières des segments. Ce résultats reste néanmoins meilleurs que le précédent puisque qu'il permet de passer de 71.6 % de bonne classification à une classification correcte à 83.2%.

6.4 Détermination d'une durée minimale de segment

On propose d'éliminer les segments classés comme purement instrumentaux dont la durée est inférieure à 1 seconde. Ceci permet d'introduire de manière heuristique des contraintes sur la durée des états. Ce traitement heuristique représente une première approximation. Il pourra être notablement amélioré par une modélisation explicite de la durée des états "chanté" et "non chanté" .

Chapitre 7

Résultats

7.1 Protocole d'évaluation

7.1.1 Base de test

La base de test utilisée pour ce projet à été gracieusement donnée par M Ramona. Cette base est constituée de 90 chansons libres de droit provenant du site internet Jamendo. Ces fichiers sont représentatifs de la musique commerciale actuelle et contiennent tous de la voix chantée. L'ensemble des chansons constitue environ 6 heures de musique annoté manuellement en catégories chanté (ch), non chanté (-). Le matériel audio et les annotations utilisés sont décrits et disponibles à l'adresse suivante : http ://www.enst.fr/ ramona/icassp08/ . Les fichiers disponibles sur Jamendo sont au format Vorbis OGG codés en stéréo à la fréquence 44.1 kHz avec un débit de 112kB/s.

7.1.2 Mesure d'évaluation de performances

Lors de l'évaluation des méthodes, nous disposons pour chaque fichier audio d'une segmentation de référence et d'une segmentation automatique. Pour chacune des segmentations, chaque trame du signal est assignée à la classe "chantée" ou à la classe "non chantée". Afin d'évaluer la performance de notre méthode nous utilisons les éléments suivants :

- les trames classée comme chantées lorsqu'elles sont effectivement chantées (TP)
- les trames classée comme chantées lorsqu'elles sont non chantées (FP)
- les trames classée comme non chantées lorsqu'elles sont chantées (FN)
- les trames classée comme non chantées lorsqu'elles sont effectivement non chantées (TN)

Ces informations peuvent être organisées dans une matrice de confusion comme représenté sur la figure 7.1.

Référence / Détecté	chanté	non chanté
chanté	TP (True Positive)	FP (False Positive)
non chantée	FN (False Negative)	TN (True Negative)

Tab. 7.1 – vrai/faux

On mesure :

- **Le rappel :**
  Le rappel (ou recall) est une mesure de sensibilité. C'est la fraction des documents cherchés trouvés sur l'ensemble des documents trouvés. Le rappel peut être vu comme la probabilité qu'un document cherché soit cherchés.

$$Rappel = \frac{TP}{TP + FN} \tag{7.1}$$

– **La précision :**
La précision est une mesure de confiance. C'est la fraction des documents cherchés trouvés sur l'ensemble des documents trouvés.

$$Precision = \frac{TP}{TP + FP} \tag{7.2}$$

Il est trivial d'accéder à une précision de 100% c'est pourquoi cette mesure doit être combinée à une autre, celles du rappel par exemple, pour mesurer la compétence.

– **F-score :**
Le F-score, mesure de compétence, est calculé comme la moyenne harmonique de la précision et du rappel :

$$F_{score} = \frac{2(Rappel * Precision)}{Rappel + Precision} \tag{7.3}$$

– **Courbe précision-rappel**
Les mesures précédentes peuvent êtres calculées pour différentes méthodes. Afin de les comparer il est d'usage de représenter les résultats sous forme de courbe de performance connue comme courbe de précision-rappel. Ces courbes sont similaires aux courbes ROC.

7.2 Résultats de la méthode basée sur la sélection des partiels avec modulation de fréquence et/ou d'amplitude

Les études de cette section ont été réalisées sur un seul des canaux des pistes audio.
Les différents paramètres de cette méthode automatique sont :
– le seuil de l'étendue relative de la modulation de fréquence
– le seuil de l'étendue relative de la modulation d'amplitude
– la durée minimale d'un segment

7.2.1 Étude des seuils

Seuil sur le vibrato

Une étude statistique a été menée afin de déterminer le seuil à partir duquel les partiels peuvent êtres considérés comme vibrés.

Pour chaque partiel, l'étendue relative de la modulation de fréquence autour de 6 Hz est calculée comme expliqué en section 5.4.2. L'ensemble des valeurs trouvées sur les segments purement instrumentaux est reporté dans l'histogramme 7.1. La voix étant toujours accompagnée il est impossible d'effectuer une analyse similaire sur les segments chantés. Il existe toujours des partiels, correspondant aux autres instruments, qui ne sont pas nécessairement vibrés.

Cette étude nous permet de définir une première approximation du seuil à utiliser pour détecter les partiels vibrés.

La voix n'étant pas le seul instrument à vibrer il existe nécessairement des partiels dans les segments purement instrumentaux pour lesquels les valeurs d'amplitude relative du vibrato sont élevées. On note que 80% des partiels ont $\Delta f / f_0$ inférieur à 0.02. Dans la suite, les évaluations seront effectuées pour des valeurs de $seuil_{vib}$ allant de 0.001 à 0.02.

Seuil tremolo

Une étude similaire a été menée sur les valeurs de l'étendue relative de la modulation d'amplitude. Les résultats sont reportés sur l'histogramme 7.2. On trouve également sur cette figure les valeurs de l'histogramme cumulé. On note que l'axe des abscisses représente les valeurs de seuil et l'axe des ordonnés

On constate que 80% des valeurs trouvées sont inférieures à 0.02.

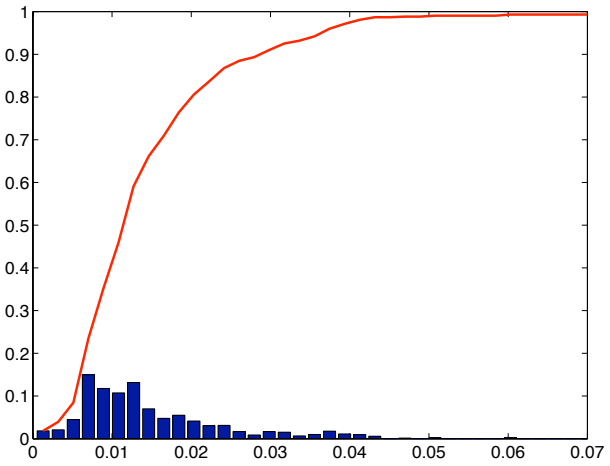


FIG. 7.1 – Seuil amplitude relative vibrato

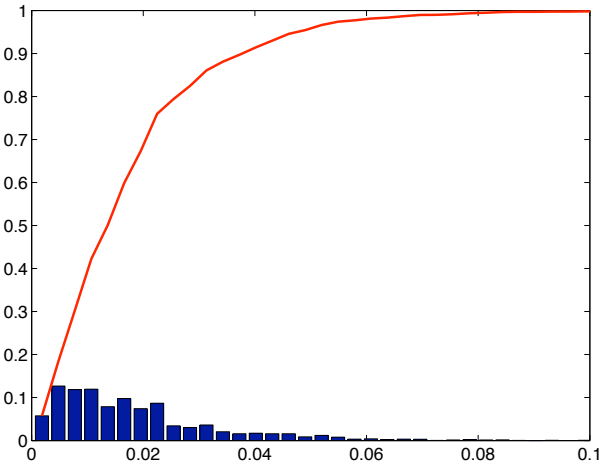


FIG. 7.2 – Seuil amplitude relative tremolo

Étude du nombre de partiels par trame

Dans la partie 5.4.2 nous avons proposé une méthode reposant sur l'hypothèse qu'après suppression des partiels non vibrés les partiels isolés ne pouvait correspondre à de la voix chantée puisque celle ci est hautement harmonique. Dans cette méthode nous avons proposé de supprimer les partiels qui ne rencontrait aucun autre partiel sur toute sa longueur. Nous démontrons ici la validité de cette hypothèse.

Dans un premier temps nous analysons le nombre de partiels total (vibré + non vibré) présents par trames pour les classes "non chanté" 7.3 et "chantées" 7.4 respectivement.

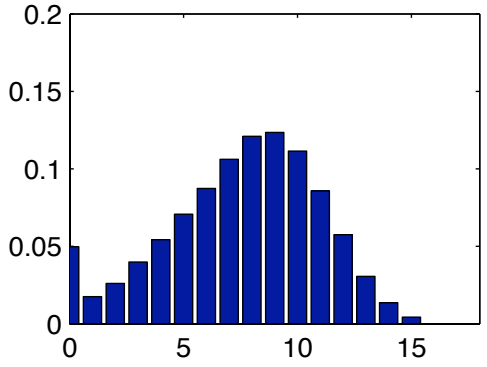


FIG. 7.3 – Nombre de partiels total pour la classe "non chantée"

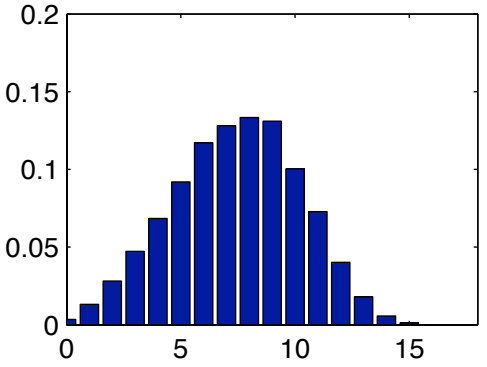


FIG. 7.4 – Nombre de partiels total pour la classe "chantée"

On constate que le nombre total de partiels n'est pas significatif pour la discrimination chantée / non chantée. On note cependant que certaines trames ont été annotés comme chantées alors qu'elles ne contiennent aucun partiel (cas des plosives et des fricatives).

Les figures suivantes (7.5 et 7.6) représentent le nombre de partiels par trame après sélection des partiels vibrés pour un seuil de vibrato égal à 0.01 :

On voit que 71% des trames non chantées comportent au plus un partiel et que 66% des trames chantées comportent au moins deux partiels. On en déduit que le critère qui consiste à considérer comme purement instrumental les trames ayant moins de deux partiels vibrés permet effectivement d'éliminer 71% des trames qui comportent des notes vibrés qui ne correspondent pas à de la voix mais élimine 34% des trames chantées.

Dans un travail futur, ce critère sera remplacé par un des critères de voisement développé dans la section 4.2

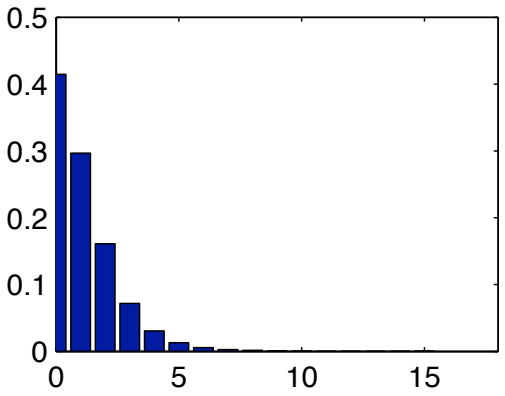


FIG. 7.5 – Nombre de partiels vibrés pour la classe “non chantée”

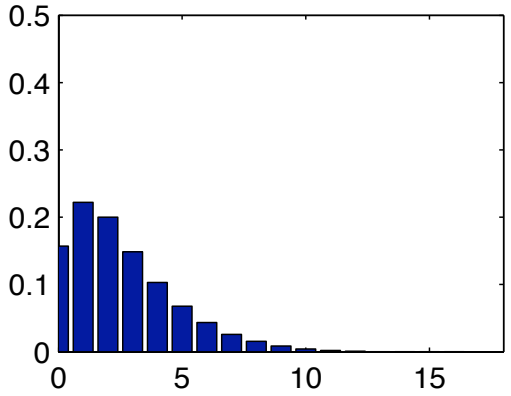


FIG. 7.6 – Nombre de partiels vibrés pour la classe “chantée”

7.2.2 Détection basée sur les modulations de fréquence seule

Dans un premier temps on compare les résultats trouvés pour différentes valeurs de seuil de vibrato pour des segments d’une durée minimal égale à 1 seconde (figure 7.7) et 1.5 seconde (figure 7.8).

Seuil FM	F-score	Rappel	Precision
0.001	0.6736	0.5229	1
0.002	0.6748	0.5243	1
0.003	0.676	0.5255	1
0.004	0.679	0.5291	0.9998
0.005	0.682	0.5326	0.9992
0.006	0.6866	0.5384	0.9972
0.007	0.6924	0.5464	0.9918
0.008	0.7009	0.5589	0.9835
0.009	0.7082	0.5728	0.9687
0.010	0.7159	0.5892	0.952
0.011	0.7212	0.6092	0.9246
0.012	**0.7246**	**0.6317**	**0.8926**
0.013	0.7237	0.6551	0.8543
0.014	0.7157	0.6761	0.8101

FIG. 7.7 – Seuil vibrato, durée minimale segment 1 sec

Seuil	F-score	Rappel	Precision
0.001	0.6734	0.5227	1
0.002	0.6745	0.524	1
0.003	0.6754	0.5248	1
0.004	0.6776	0.5274	0.9998
0.005	0.6798	0.5299	0.9995
0.006	0.6829	0.5336	0.9987
0.007	0.6867	0.5387	0.9956
0.008	0.6929	0.5468	0.9917
0.009	0.7006	0.5581	0.9846
0.010	0.7082	0.5709	0.9747
0.011	0.7172	0.5893	0.9581
0.012	0.7234	0.6084	0.9343
0.013	**0.7252**	**0.6275**	**0.9029**
0.014	0.7231	0.6466	0.8674

FIG. 7.8 – Seuil vibrato, durée minimale segment 1.5 sec

On constate qu’il n’existe pas de différence significative entre ces résultats. Logiquement, plus le seuil du vibrato est élevé plus la précision diminue et plus le rappel augmente. Le seuil de vibrato qui permet la meilleure classification est situé au environ de [0.012, 0.013]. Il serait nécessaire de faire un test statistique pour voir si la différence entre les résultats pour deux seuils proches est significative.

7.2.3 Détection basée sur les modulations de fréquence et d’amplitude

Afin d’évaluer l’apport de la sélection des partiels qui sont modulés en fréquence et en d’amplitude, on fixe le seuil de la modulation d’amplitude puis on étudie les résultats de la classification pour différentes valeurs de seuil de vibrato. Les résultats de 7.2.1 nous conduisent à fixer le seuil de trémolo à 0.002. La durée minimale d’un segment est fixée ici à 1 seconde.

Les résultats de la classification figurent dans le tableau 7.9.

On note que pour un seuil de vibrato fixé (0.009) cette méthode permet d’améliorer le F-score de 3% environ. Les différentes mesures pour les 2 approches sont reportées dans la figure 7.10. On constate que cette amélioration est principalement due au gain apporté au rappel (environ 10%). Par ailleurs, le fait de sélectionner seulement les partiels qui possèdent les deux types de

Seuil FM	F-score	Rappel	Precision
0.008	0.7325	0.627	0.9206
0.009	**0.7355**	**0.6532**	**0.8816**
0.010	0.7347	0.6776	0.8425
0.011	0.7255	0.701	0.7955
0.012	0.7105	0.7234	0.7435
0.013	0.6905	0.743	0.692
0.014	0.665	0.7602	0.6394

FIG. 7.9 – Seuil vibrato, durée minimale segment 1 sec, seuil modulation d'amplitude 0.002

seuil vib 0.009	Avec AM	Sans AM
F-score	0.7355	0.7082
Rappel	0.6532	0.5728
Precision	0.8816	0.9687

FIG. 7.10 – Comparaison des résultats avec/sans AM

modulation fait diminuer la précision. On note que le meilleur résultat de cette dernière méthode est atteint pour un seuil de vibrato de 0.009 qui est inférieur au seuil précédent (0.012).

0n compare ces trois dernières évaluations à l'aide des courbes de rappel-précision.

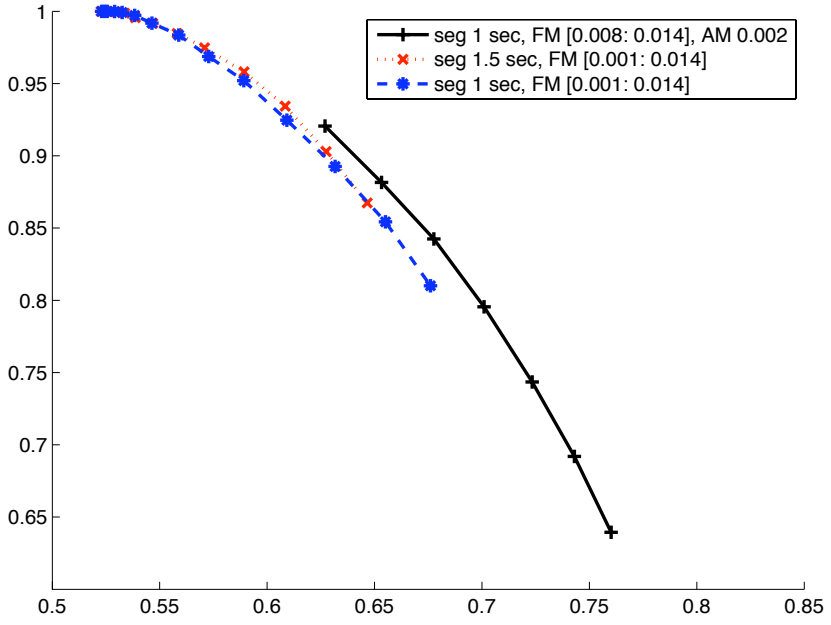


FIG. 7.11 – Courbe rappel-précision

La méthode incluant la modulation d'amplitude est meilleure que les deux précédentes puisqu'il est possible d'obtenir un meilleur pourcentage de rappel pour une précision fixée.

Meilleurs paramètres

Afin d'optimiser le seuil de modulation d'amplitude, on fixe le seuil vibrato puis on compare les résultats pour différentes valeurs de seuil de tremolo. On fixe le seuil de vibrato à 0.008 (7.12) et à 0.009 (7.13) car ces valeurs produisent la meilleure classification dans l'étude précédente 7.9.

On compare les résultats de ces deux méthodes à l'aide des courbes de rappel-précision. Afin de faire un lien avec les résultats précédents on a reporté sur la figure 7.14 la courbe correspondant à l'étude 7.2.3

Seuil AM	F-score	Rappel	Precision
0.001	0.7325	0.627	0.9206
0.002	0.737	0.6489	0.8948
0.003	**0.7399**	**0.6714**	**0.868**
0.004	**0.7399**	**0.6939**	**0.839**
0.005	0.7337	0.7157	0.8036
0.006	0.7208	0.7338	0.7625
0.007	0.7071	0.7508	0.7257
0.008	0.6919	0.7684	0.6894
0.009	0.6759	0.785	0.6536
0.010	0.6555	0.7995	0.6169
0.011	0.6351	0.8118	0.5823
0.012	0.61	0.8221	0.5455
0.013	0.5857	0.8311	0.5118

FIG. 7.12 – Seuil vibrato fixé 0.008, durée minimale segment 1 sec, seuil modulation d'amplitude

Seuil AM	F-score	Rappel	Precision
0.003	0.7345	0.6294	0.9254
0.004	0.7403	0.6513	0.9023
0.005	**0.7419**	**0.6718**	**0.8764**
0.006	0.74	0.6934	0.8457
0.007	0.7314	0.7123	0.8085
0.008	0.721	0.729	0.7734
0.009	0.7101	0.7469	0.7414

FIG. 7.13 – Seuil vibrato fixé 0.009, durée minimale segment 1.5 sec, seuil modulation d'amplitude

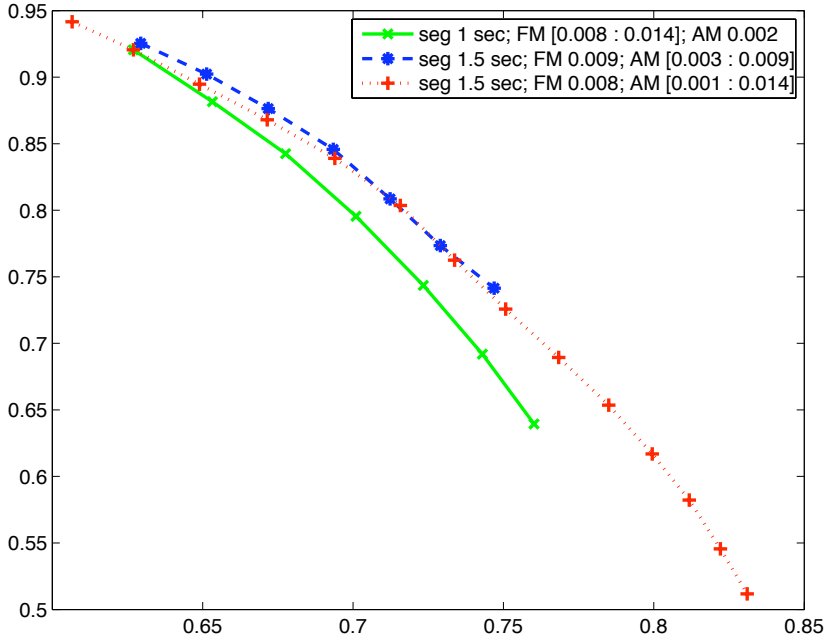


FIG. 7.14 – Courbe rappel-précision

7.2.4 Conclusions

Dans le meilleur cas, nous arrivons à un F-score de 74.2% :
En règle général

- La classification basée sur les modulation de fréquence seules permet d'obtenir un F-score de 72.5% avec une précision de 90% et un rappel de 63%. Ces valeurs ont été obtenues pour un seuil de vibrato égal à 0.012.
- La classification basée sur les modulation de fréquence et d'amplitude permet d'obtenir un F-score de 74% avec une précision de 87% et un rappel de 67%. Ces valeurs ont été obtenues pour un seuil de vibrato de 0.009 et un seuil de trémolo égal à 0.005.
- Le fait de sélectionner les partiels avec modulation d'amplitude et de fréquence augmente le pourcentage de rappel et diminuer la précision.

7.3 Utilisation de la stéréo

Les évaluations précédentes ont été menées sur un seul canal des pistes audio. En partant du constat que la voix chantée est souvent spatialisée au centre, on en déduit qu'elle doit être présente sur les deux canaux c'est pourquoi on étudie l'intersection et la réunion de la segmentation de chacun des canaux.

L'intersection : Il s'agit de considérer comme chantées les trames qui ont été classées comme chantées sur le canal droit et sur le canal gauche.

L'union : Il s'agit de considérer comme chantées les trames qui ont été classées comme chantées sur le canal droit ou sur le canal gauche.

On analyse les résultats trouvés pour deux chansons avant et après lissage de la segmentation. Ces chansons correspondent au meilleur et au pire résultat trouvés pour la méthode 7.13.

	Canal droit	Canal gauche	Union	Intersection
F-score	0.85	0.86	0.90	0.83
Rappel	0.97	0.97	0.89	0.81
Précision	0.76	0.77	0.85	0.84

FIG. 7.15 – Évaluation de la chanson "Alice" avant lissage

	Canal droit	Canal gauche	Union	Intersection
F-score	0.95	0.96	0.95	0.90
Rappel	0.96	0.97	0.94	0.83
Précision	0.94	0.95	0.96	0.99

FIG. 7.16 – Évaluation après lissage

Dans cet exemple le fait de choisir l'intersection ou la réunion des deux segmentations n'améliore pas le résultat.

Lorsque la classification est assez mauvaise le fait de sélectionner les parties chantées du canal droit et du canal gauche permet d'améliorer de façon surprenante le résultat final comme le montre l'exemple suivant.

	Canal droit	Canal gauche	Union	Intersection
F-score	0.37	0.42	0.50	0.71
Rappel	0.88	0.88	0.84	0.65
Précision	0.23	0.27	0.35	0.78

FIG. 7.17 – Évaluation de la chanson "10min" avant lissage

	Canal droit	Canal gauche	Union	Intersection
F-score	0.47	0.54	0.66	0.80
Rappel	0.91	0.87	0.84	0.67
Précision	0.31	0.39	0.54	1

FIG. 7.18 – Évaluation après lissage

Ces exemples montrent que :

- En général, le lissage, pour un seul canal, permet d'améliorer la classification de 10% en améliorant considérablement la précision mais diminuant le rappel. On rappelle que la méthode de lissage utilisée ici est décrite en section 6.4
- Pour de très bons résultats, avant lissage, l'union permet d'augmenter la précision d'environ 8% faisant ainsi accroître le F-score d'environ 5%. Les résultats après lissage sont sensiblement similaires.
- Pour de mauvais résultats, l'utilisation de l'intersection permet d'accroître la précision de façon considérable : 55%. Cette amélioration permet d'augmenter le F-score de 34% avant lissage. On remarque que l'utilisation de l'union des deux segmentations permet également d'augmenter la précision et le F-score. Ces résultats sont encore meilleurs après lissage.
- On en conclut que cette approche permet effectivement d'améliorer le résultat si l'on choisit au cas par cas l'union ou l'intersection des segmentations de chaque canal. En règle général, l'utilisation de l'intersection permet d'augmenter la précision.

Pour finir, on étudie les apports de cette approche sur tous les éléments de la base d'entraînement (45 chansons).

On pourrait également effectué des mesures sur l'union et la réunion de deux segmentations lissées.

	Canal droit	Canal gauche	Intersection	Union
F-score	0.64	0.65	0.61	0.69
Rappel	0.74	0.75	0.52	0.71
Précision	0.62	0.62	0.77	0.73

FIG. 7.19 – Évaluation sur 45 chansons avant lissage

	Canal droit	Canal gauche	Intersection	Union
F-score	0.73	0.74	0.67	0.74
Rappel	0.71	0.71	0.53	0.67
Précision	0.82	0.82	0.99	0.89

FIG. 7.20 – Évaluation après lissage

7.4 Conclusion

Les résultats obtenus sont très encourageants. En comparant ces résultats à ceux obtenus avec la méthode de référence (chapitre 3) on constate qu'on arrive à des performances similaires alors qu'il s'agit d'une méthode sans apprentissage basé sur un unique descripteur.

Chapitre 8

Conclusion et perspectives

La détection des segments de voix chantée dans un flux audio est un problème de classification. Ces problèmes sont généralement résolus par des méthodes utilisants d'une part des descripteurs et d'autre part un modèle d'apprentissage. Le travail effectué lors de ce stage s'inscrit dans la recherche de descripteurs pertinents pour la détection de la voix dans un morceau de musique. Le descripteur mis en place s'appuie sur les modulations de fréquence et d'amplitude ainsi que sur le voisement qui sont des éléments caractéristiques de la voix chantée. Les formants qui sont une autre caractéristique de la voix nont pas été utilisés ici. La classification obtenue, sans apprentissage, à l'aide de cet unique descripteur est très encourageante. Cette méthode à été évaluée par les mesure de précision et de rappel et fournit un F-score de 74%. Une méthode de référence utilisant 102 descripteurs issus du calcul des MFCCs et de la SFM et des GMM comme classifieur a été évaluée sur le même ensemble de donnée. Cette méthode trouve une classification correcte à 77%. Compte tenu du fait que notre méthode n'utilise aucun apprentissage et qu'elle utilise unique descripteur nous estimons que le résultat est satisfaisant. Nous avons par ailleurs exploité la stéréophonie pour la classification. Cette démarche qui pourrait fournir des résultats intéressants demande à être approfondie.

Dans de travaux futurs, nous espérons améliorer la robustesse du descripteur en travaillant sur les points suivants.

Tout d'abord il faudrait incorporer le coefficient de voisement décrit dans la section **??** afin de détecter les zones harmoniques. Ce coefficient ne permet pas de discriminer les parties chantées des parties non chantées car elles peuvent toutes être harmoniques. Cependant il peut être utilisé pour segmenter un flux audio en parties voisées et non voisées. On pourrait améliorer la classification finale en prenant en compte les deux segmentations.

Par ailleurs le post-traitement qui consiste à lisser la classification obtenue à l'aide d'un seuil pourra être remplacé par un traitement utilisant des semi chaînes de Markov cachées. Ce type de traitement permet d'estimer la durée des trames pour chacune des classe dans des chaînes de Markov.

Les seuil de modulation d'amplitude et de fréquence pourront être appris à l'aide d'un SVM. Pour ce faire il faudrait d'abord trouver des segments purement vocaux afin d'apprendre les caractéristiques du vibrato. Le problème actuel étant que notre base de test ne contient pas de morceau a capela.

Une étude sur les formants de la voix pourra également être menée.

Une fois ces amélioration apportées on pourra intégrer ce descripteur à la méthode de référence afin dévaluer l'impact.

Bibliographie

[BE01] AL Berenzweig and DPW Ellis. Locating singing voice segments within music signals. *Applications of Signal Processing to Audio and Acoustics, 2001 IEEE Workshop on the*, pages 119–122, 2001.

[BEL02] A. Berenzweig, D.P.W. Ellis, and S. Lawrence. Using voice segments to improve artist classification of music. *AES 22nd International Conference*, 2002.

[BS03] J. Bretos and J. Sundberg. Measurements of vibrato parameters in long sustained crescendo notes as sung by ten sopranos. *Journal of Voice*, 17(3) :343–352, 2003.

[CG01] W. Chou and L. Gu. Robust singing detection in speech/music discriminator design. *Acoustics, Speech, and Signal Processing, 2001. Proceedings.(ICASSP'01). 2001 IEEE International Conference on*, 2, 2001.

[CKK98] Y.D. Cho, M.Y. Kim, and S.R. Kim. A spectrally mixed excitation (smx) vocoder with robust parameterdetermination. *Acoustics, Speech, and Signal Processing, 1998. ICASSP'98. Proceedings of the 1998 IEEE International Conference on*, 2, 1998.

[Coo90] P.R. Cook. *Identification of control parameters in an articulatory vocal tract model, with applications to the synthesis of singing.* PhD thesis, to the Department of Electrical Engineering.Stanford University, 1990.

[DH89] P. Desain and H. Honing. The quantization of musical time : A connectionist approach. *COMP. MUSIC J.*, 13(3) :56–66, 1989.

[FKG+05] H. Fujihara, T. Kitahara, M. Goto, K. Komatani, T. Ogata, and H.G. Okuno. Singer identification based on accompaniment sound reduction and reliable frame selection. *Proc. ISMIR*, pages 329–336, 2005.

[GHD+05] M. Garnier, N. Henrich, D. Dubois, M. Castellengo, J. Poitevineau, and D. Sotiropoulos. Etude de la qualité vocale dans le chant lyrique. *Scolia*, 20 :151–169, 2005.

[KW02] Youngmoo E. Kim and Brian Whitman. Singer identification in popular music recordings using voice coding features. In *ISMIR*, 2002.

[LGD07] H. Lukashevich, M. Gruhne, and C. Dittmar. Effective singing voice detection in popular music using arma filtering. *Workshop on Digital Audio Effects (DAFx'07)*, 2007.

[LW06a] Y. Li and D.L. Wang. Separation of singing voice from music accompaniment for monaural recordings. *IEEE Transactions on Audio, Speech, and Language Processing*, 2006.

[LW06b] Yipeng Li and DeLiang Wang. Singing voice separation from monaural recordings. In *ISMIR*, pages 176–179, 2006.

[Mar99] K.D. Martin. *Sound-Source Recognition : A Theory and Computational Model.* PhD thesis, Massachusetts Institute of Technology, 1999.

[Moo96] TK Moon. The expectation-maximization algorithm. *Signal Processing Magazine, IEEE*, 13(6) :47–60, 1996.

[MQ86] R. McAulay and T. Quatieri. Speech analysis/synthesis based on a sinusoidal representation. *Acoustics, Speech, and Signal Processing [see also IEEE Transactions on Signal Processing], IEEE Transactions on*, 34(4) :744–754, 1986.

[MR04] S. Marchand and M. Raspaud. Enhanced time-stretching using order-2 sinusoidal modeling. *Proc. DAFx*, pages 76–82, 2004.

[MWXW04] Namunu Chinthaka Maddage, Kongwah Wan, Changsheng Xu, and Ye Wang. Singing voice detection using twice-iterated composite fourier transform. In *ICME*, pages 1347–1350. IEEE, 2004.

[MXW03] N.C. Maddage, C. Xu, and Y. Wang. An svm-based classification approach to musical audio. *4th International Conference on Music Information Retrieval*, 2003.

[NL07] Tin Lay Nwe and Haizhou Li. Singing voice detection using perceptually-motivated features. In Rainer Lienhart, Anand R. Prasad, Alan Hanjalic, Sunghyun Choi, Brian P. Bailey, and Nicu Sebe, editors, *ACM Multimedia*, pages 309–312. ACM, 2007.

[NSW04] Tin Lay Nwe, Arun Shenoy, and Ye Wang. Singing voice detection in popular music. In Henning Schulzrinne, Nevenka Dimitrova, Angela Sasse, Sue B. Moon, and Rainer Lienhart, editors, *ACM Multimedia*, pages 324–327. ACM, 2004.

[NW04] Tin Lay Nwe and Ye Wang. Automatic detection of vocal segments in popular songs. In *ISMIR*, 2004.

[Pee03] G. Peeters. Automatic classification of large musical instrument databases using hierarchical classifiers with inertia ratio maximization. *115th AES convention, New York, USA, October*, 2003.

[Pee04] G. Peeters. A large set of audio features for sound description (similarity and classification) in the cuidado project. *CUIDADO Project Report*, 2004.

[Pee06] G. Peeters. Music pitch representation by periodicity measures based on combined temporal and spectral representations. *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, 5, 2006.

[Pee07] G. Peeters. A generic system for audio indexing : application to speech/ music segmentation and music genre. *Proc. of DAFX*, 2007.

[Pra94] E. Prame. Measurements of the vibrato rate of ten singers. *The Journal of the Acoustical Society of America*, 96 :1979, 1994.

[Rd96] G. Richard and C. d'Alessandro. Analysis/synthesis and modification of the speech aperiodic component. *Speech Communication*, 19(3) :221–244, 1996.

[RDS+99] S. Rossignol, P. Depalle, J. Soumagne, X. Rodet, and J.L. Collette. Vibrato : detection, estimation, extraction, modification. *Proceedings*, 99, 1999.

[RH07] M. Rocamora and P. Herrera. Comparing audio descriptors for singing voice detection in music audio files. *SBCM - Brasilian Sypmosium on Computer Music 07*, 2007.

[RR08] M. Ramona and B. Richard, G. David. Vocal detection in music with support vector machine. *Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'08). IEEE International Conference on*, pages 1885 – 1888, March 2008.

[SR90] J. Sundberg and T.D. Rossing. The science of singing voice. *The Journal of the Acoustical Society of America*, 87 :462, 1990.

[SS97] E. Scheirer and M. Slaney. Construction and evaluation of a robust multifeature speech/musicdiscriminator. *Acoustics, Speech, and Signal Processing, 1997. ICASSP-97., 1997 IEEE International Conference on*, 2, 1997.

[SWW05] A. Shenoy, Y. Wu, and Y. Wang. Singing voice detection for karaoke application. *Proceedings of SPIE*, 5960 :596028, 2005.

[TD00] R. Timmers and P. Desain. Vibrato : Questions and answers from musicians and science. *Proc. ICMPC*, 2000.

[TW06] W.H. Tsai and H.M. Wang. Automatic singer recognition of popular music recordings via estimation and modeling of solo vocal signals. *Audio, Speech and Language Processing, IEEE Transactions on [see also Speech and Audio Processing, IEEE Transactions on]*, 14(1) :330–341, 2006.

[Tza04] George Tzanetakis. Song-specific bootstrapping of singing voice structure. In *ICME*, pages 2027–2030. IEEE, 2004.

[VGD05] V. Verfaille, C. Guastavino, and P. Depalle. Perceptual evaluation of vibrato models. *Proceedings of Conference on Interdisciplinary Musicology*, 2005.

[Zha03a] T. Zhang. Automatic singer identification. *Multimedia and Expo, 2003. ICME'03. Proceedings. 2003 International Conference on*, 1, 2003.

[Zha03b] T. Zhang. Semi-automatic approach for music classification. In *SPIE Conf. on Internet Multimedia Management Systems*, volume 5242, pages 81–91, 2003.