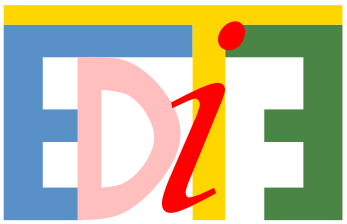
Transcription et séparation automatique de la mélodie principale dans les signaux de musique polyphoniques

Jean-Louis Durrieu



École Doctorale
d'Informatique,
Télécommunications
et Électronique de Paris

Thèse

présentée pour obtenir le grade de docteur

de TÉLÉCOM ParisTech

Spécialité : Signal et Images

Jean-Louis DURRIEU

Transcription et séparation automatique de la mélodie principale dans les signaux de musique polyphoniques.

Automatic transcription and separation of the main melody in polyphonic music signals.

Soutenue le 7 mai 2010 devant le jury composé de

Frédéric BIMBOT	Président du jury
Daniel ELLIS	Rapporteurs
Christian JUTTEN	
Simon GODSILL	Examinateurs
Emmanuel VINCENT	
Bertrand DAVID	Directeurs de thèse
Gaël RICHARD	

À Daniela et Lucien Edwin

Résumé

Nous proposons de traiter l'extraction de la mélodie principale, ainsi que la séparation de l'instrument jouant cette mélodie. La première tâche appartient au domaine de la recherche d'information musicale (MIR en Anglais), parce que nous cherchons à indexer les morceaux de musique à l'aide de leur ligne mélodique. La seconde application est un problème de séparation aveugle de sources sonores (BASS en Anglais), avec pour but d'extraire une piste audio pour chaque source présente dans un mélange sonore.

De nombreux travaux ont récemment jumelé ces deux domaines. En effet, la MIR et la BASS visent un même résultat, la décomposition du mélange en "atomes", avec cependant des interprétations différentes suivant le domaine. Ainsi, en MIR, ces éléments comportent une connotation musicale, avec une sémantique relativement abstraite. En BASS, ces atomes revêtent plutôt un sens physique. En général, on constate alors que les systèmes orientés MIR tendent à éluder du traitement une partie de l'information de sorte à se concentrer directement sur le niveau de description voulu, alors que les systèmes de séparation ont plutôt tendance à ne considérer que peu d'information autre que des aspects physiques tels que les informations spatiales, par exemples. La combinaison des deux types d'approches semble pourtant intuitivement intéressante, étant donnés leurs buts respectifs.

Dans cette thèse, nous proposons d'effectuer le problème de séparation de la mélodie principale et de l'accompagnement ainsi que l'extraction de cette mélodie à l'aide d'un même cadre statistique. Le modèle pour l'instrument principal est un modèle de production source/filtre. Il suppose deux états cachés correspondant respectivement à l'état du filtre et à l'état de la source. Le modèle spectral choisi permet ainsi de prendre explicitement en considération les fréquences fondamentales (ou hauteurs - *pitch*) de l'instrument désiré, afin d'estimer d'abord la séquence de pitches joués, la "mélodie principale", mais aussi de séparer l'instrument qui la joue de l'accompagnement. Deux modèles de signaux sont proposés, un modèle de mélange de gaussiennes amplifiées (GSMM en Anglais) ainsi qu'un modèle que nous avons désigné comme un modèle de mélange instantané (IMM en Anglais). Chacun de ces modèles présente des avantages et des inconvénients. L'accompagnement est lui modélisé par un modèle spectral plus général qui permet d'envisager un éventail assez large de types d'accompagnement possibles.

Le lien entre les modèles statistiques choisis et la factorisation en matrices non-négatives (NMF en Anglais) nous a permis d'employer et d'adapter les algorithmes d'estimation de paramètres déjà existants pour estimer les paramètres de nos modèles. Par ailleurs, afin d'estimer les séquences mélodiques de fréquences fondamentales et de notes, nous proposons des approximations à des degrés variés, des stratégies qui réduisent la difficulté des problèmes posés. Cinq systèmes sont ainsi proposés, trois d'entre eux ont pour but de fournir la mélodie sous forme de séquence de fréquences fondamentales, un autre système vise à fournir la suite de notes musicales jouées et enfin le dernier système permet d'effectuer la séparation de l'instrument principal et de l'accompagnement.

Les résultats obtenus en estimation de la mélodie et en séparation sont du niveau de l'état de l'art, comme l'ont montré nos différentes participations aux évaluations internationales (MIREX'08, MIREX'09 et SiSEC'08). Cela valide la possibilité d'un système intégrant ces deux aspects. Durant cette thèse, nous avons aussi cherché à réduire les différentes approximations réalisées lors de l'estimation. Un résultat important de notre travail est d'avoir intégré de la connaissance inspirée de la communauté MIR afin d'améliorer les résultats de travaux antérieurs sur la séparation de sources sonores.

Enfin, le cadre statistique de nos modèles permet d'imaginer d'éventuels raffinements

du modèle. Des contraintes et des *a priori* peuvent être facilement définis : l'exemple d'un *a priori* sémantiquement motivé, formellement proche des contraintes de parcimonie et de décorrélation, est discuté dans cette thèse. D'autres améliorations des modèles sont possibles, notamment pour celui de l'instrument principal, ou celui de l'accompagnement, voire par l'ajout d'une modélisation des interactions entre ces deux contributions. Quelles sont les notes les plus probables pour la mélodie, étant donné les accords de l'accompagnement ? Répondre à ce type de questions permettra sans doute d'obtenir des transcriptions musicales plus réalistes, en ouvrant aussi la voie à une meilleure compréhension des mécanismes cognitifs qui régissent à notre perception de la musique et de sa structure.

Abstract

We propose to address the problem of melody extraction along with the "monaural lead instrument and accompaniment separation" problem. The first task is related to Music Information Retrieval (MIR), since it aims at indexing audio music signals with their melody line. The separation problem is related to Blind Audio Source Separation (BASS), as it aims at breaking an audio mixture into several source tracks.

Many recent research works have more or less explicitly brought these two fields together. Indeed, MIR and BASS share a common goal: we desire some atomic decomposition of an audio mixture. Of course, the back-end applications may be different. The "atoms" therefore have different meanings: for MIR, the extracted elements should have some musical, high-level semantics, while for BASS, these elements are more related to low-level aspects of the signals. This often leads to MIR systems that tend to discard information in order to directly access to the desired level of description. BASS systems usually consider only physical aspects allowing to distinguish the different sources. Intuitively, combining these approaches can lead to mutual improvements in both areas.

In this thesis, we propose to address leading instrument separation and main melody extraction in a unified framework. We first describe the signal models: the lead instrument is modelled thanks to a source/filter production model. Its signal is generated by two hidden states, the filter state and the source state. The proposed signal spectral model therefore explicitly uses pitch (fundamental frequencies) both to separate the lead instrument from the other instruments and to transcribe the pitch sequence played by that instrument, the so-called "main melody". This model gives rise to two alternative models, a Gaussian Scaled Mixture Model (GSMM) and a model we called the Instantaneous Mixture Model (IMM), each of which having their own advantages and drawbacks. The accompaniment is modelled with a more general spectral model which can describe a large variety of musical backgrounds.

The estimation of the different parameters and of the melody sequence is addressed thanks to methodologies borrowed from Non-negative Matrix Factorization (NMF) literature. Indeed, within the proposed statistical framework, parameter estimation is very similar to an NMF problem. Since the proposed models have several layers of hidden states, several different strategies can be investigated, with various levels of approximation. From these strategies, we have extracted five systems. Three of them aim at detecting the fundamental frequency sequence of the lead instrument, in other words estimating the main melody. Another system is designed to return a musical transcription of the main melody, that is the sequence of notes (pitch in the Western musical scale, onset and offset times) and the last system targets the separation of the lead instrument from the accompaniment.

The results in melody transcription and source separation are comparable to the state

of the art, as shown by our participations to several international evaluation campaigns (MIREX'08, MIREX'09 and SiSEC'08). This means that a completely joint system for transcription and separation is possible. The proposed systems use estimation algorithms for which we have worked on avoiding the approximations that were made. Our results for source separation also provided an interesting insight in the field: the proposed extension of previous works using "MIR" knowledge is a very successful combination.

At last, the chosen statistical framework enables further refinement of the model. Constraints and priors on the parameters can be defined: an example of a semantically motivated prior, recalling sparsity and de-correlation constraints, is given and discussed in this work. Future directions for research go through the improvement of the lead instrument model, as well as the use of a more complex accompaniment model. An interesting path for research could be to model the high level dependency between the lead and the accompaniment: what notes is the lead most likely to play, knowing what kind of chord the accompaniment is playing? Answering this question, among others, may improve the performance in terms of transcription into a musical score and, to a certain extent, our understanding of how we perceive music and its structure.

Contents

Remerciements

Je souhaite en premier lieu remercier Gaël Richard et Bertrand David, mes directeurs de thèse. Leur aide, leur confiance, leur soutien et leur optimisme ont fait progresser ma thèse des balbutiements de ma recherche jusqu'à ce manuscrit, en passant par les diverses publications que nous avons co-signées. Durant ces trois années, un réel équilibre s'est formé entre les intuitions souvent justes de Gaël et la rigueur de Bertrand. Ces dernières m'auront permis d'évoluer librement, avec l'élan nécessaire à l'innovation ainsi que le cadre requis pour me garder de possibles dévoiements. Je souhaite par ailleurs saluer leur travail de gestion des projets qui ont financé ma thèse, travail vital pour une grande partie d'entre nous, doctorants et "post-doctorants", mais bien souvent peu valorisant et peu reconnu.

Un grand merci aux membres de mon jury de thèse : Daniel Ellis et Christian Jutten, mes rapporteurs, Simon Godsill et Emmanuel Vincent, mes examinateurs, et je remercie enfin Frédéric Bimbot d'avoir accepté de présider ce jury.

Merci à mes co-auteurs et proches collaborateurs, Nancy Bertin, Cédric Févotte, Alexey Ozerov et Jan Weil, pour nos échanges intéressants, innovateurs et fructueux.

Merci à toute l'équipe "Audiosig" de Télécom ParisTech, pour l'ambiance chaleureuse instaurée par ses membres, présents et passés, Roland Badeau, Valentin Emiya, Slim Essid, Sarah Filippi, Thomas Fillon, Benoît Fuentes, Yves Grenier, Sébastien Gulluni, Romain Hennequin, Cyril Joder, Mathieu Lagrange, Pierre Leveau, Mounira Maazaoui, Benoît Mathieu, Nicolas Moreau, Laurent Oudre, Mathieu Ramona, Félicien Vallet. Merci aux collègues des groupes STA et MM, entre autres : Aurélia, Christophe, Cyril C., Ismael, Jean, Jean-François, Julien et Teodora. De par les discussions scientifiques, les rencontres furtives ou les brèves de comptoir à la machine à café, ils auront été à divers degrés les artisans de ma formation à la recherche et à la vie professionnelle.

Merci au personnel administratif et technique de Télécom ParisTech, Sophie-Charlotte Barrière, Florence Besnard, Fabienne Lassausaie, Fabrice Planche, Laurence Zelmar, ainsi que toutes les personnes qui m'auront aussi apporté leur aide durant ces années de thèse.

Merci à Olivier Gillet et Antoine Liutkus de m'avoir montré qu'il n'était pas impossible que d'autres implémentent mes algorithmes de séparation de sources, me donnant ainsi l'espoir que mes efforts de recherche n'auront pas été complètement vains et sans suite.

Merci à mes parents, Marie-Ange et Patrick, et ma soeur, Marie-Pierre, d'avoir toujours été un soutien inconditionnel dans les différentes épreuves et aventures qui ont jalonné mon parcours. Merci à mes beaux-parents, Erika et Alfred Häuptli, pour leur accueil et leur présence, notamment durant la dernière année de thèse, au moment de l'attente de la naissance de mon fils.

Merci à ma femme, Daniela, pour les sacrifices consentis afin de me permettre de réaliser cette thèse, et à mon fils, Lucien Edwin, qui nous a rejoint en cours de route, donnant ainsi un cachet tout particulier à cette fin de thèse.

Résumé des travaux de thèse

Ce chapitre est un résumé rédigé en Français du présent document. Il reprend les grands axes du document originellement écrit en Anglais.

Dans une première partie, nous introduisons le problème de l'extraction et de la séparation de la mélodie principale dans les signaux de musique polyphonique. Nous exposons ensuite les modèles de signaux proposés dans cette thèse. Les algorithmes d'estimation et les systèmes mis en oeuvre sont alors décrits. Nous analysons ensuite les résultats obtenus par ces systèmes, en fonction des différentes tâches qu'ils traitent. Enfin nous concluons ce chapitre par un résumé de nos contributions et développons certaines pistes à explorer à l'avenir pour améliorer encore les performances des systèmes présentés.

0.1 Introduction

Explorer une base de données musicale, chercher des morceaux que l'on aime, découvrir de nouveaux titres... Toutes ces activités nécessitent une annotation particulière des signaux de musique. Cette annotation fait l'objet d'une recherche toujours plus intense dans la communauté de la "recherche d'information musicale" (MIR en Anglais). Nous traitons d'abord du traitement automatique de la musique dans sa généralité, puis revenons sur le sujet particulier qui nous intéresse, l'extraction automatique de la mélodie principale. Puis le problème la séparation de l'instrument principal et de l'accompagnement est ensuite présenté. Pour chacune de ces applications, les techniques existantes sont rappelées et commentées. Enfin, les contributions de notre travail sont précisées, notamment par rapport à l'état de l'art.

0.1.1 Le traitement automatique des signaux musicaux

A partir d'un signal audio musical, quel type d'information pouvons-nous extraire ? A l'instar des musiciens, peut-on aisément obtenir une partition de musique à la seule écoute du signal ? Quelles sont les difficultés liées à la transcription musicale, et quelles solutions intermédiaires peut-on envisager avant d'accomplir avec succès une telle tâche ?

Le problème de transcription musicale est en fait très complexe, et reste pour l'instant sans réelle solution. On trouve certes des systèmes qui, mis bout à bout, permettraient d'obtenir des résultats s'approchant d'une partition de musique. De nombreuses études ont par exemple cherché à détecter quels instruments sont présents dans un extrait musical ([Essid et al., 2006a,b, Vincent, 2004]). L'estimation du tempo a aussi bénéficié d'une forte popularité ([Scheirer, 1998, Alonso et al., 2007, Peeters, 2007]), fournissant d'une part un constituant essentiel permettant de définir les rythmes, mais aussi un attribut utile pour la classification en genre, par exemple. L'estimation de fréquences fondamentales dans les signaux de musique polyphonique [Klapuri, 2001], voire l'estimation des notes de

musique [Ryynänen and Klapuri, 2005, Emiya et al., 2009] sont des sujets aussi difficiles à traiter que leurs objectifs précis sont ambigus à définir, comme nous le verrons aussi pour le cas de la transcription de la mélodie principale. Enfin, pour obtenir une partition lisible par un musicien, il faut encore quantifier les notes temporellement, avec l'estimation de la mesure [Peeters, 2009, Weil et al., 2009a], et au final obtenir un bon compromis entre simplicité d'écriture (et donc de lecture) des rythmes et complexité de la musique [Cemgil and Kappen, 2003].

La transcription musicale ne se limite cependant pas à l'écriture d'une partition : les différents résultats intermédiaires évoqués ci-dessus permettent en particulier d'annoter le signal audio. Grâce aux différentes composantes ainsi obtenues, la classification en genres, la recherche de musique par similarité, par requête chantée, la recommandation de morceaux de musique ou la génération automatique de listes de lecture sont autant d'applications qui peuvent être traitées avec une hiérarchie d'attributs allant du bas niveau, le signal, son énergie et ainsi de suite, vers des niveaux plus sémantiques comme la mélodie, les accords, les tonalités, en passant par des "demi-niveaux" comme le tempo, ou les notes de musique.

0.1.2 Extraction automatique de la mélodie principale

Pour identifier un morceau de musique, il suffit souvent d'une mélodie, d'une séquence d'accords ou de rythmes. Nous nous intéressons plus particulièrement à l'estimation de la mélodie principale. Les applications qui peuvent prendre partie d'un tel attribut sont la recherche par requête chantée (*Query-by-Humming*, QbH), la détection de reprises ou plus généralement l'indexation de base de données.

La définition donnée par Paiva [2006] synthétise ce qui est attendu de la mélodie principale :

Definition 1 (Mélodie Principale:) *La mélodie est la ligne pitchée, individuelle et prédominante dans un ensemble musical.*◇

Cette définition oriente les choix que nous avons faits pour les modèles de signaux proposés en Section 0.2, en particulier, les différents éléments qui y apparaissent sont pris en considération de la manière suivante :

- *Ensemble musical :* les morceaux traités contiennent une ligne mélodique accompagnée par de la musique polyphonique. L'accompagnement peut être composé d'un ou plusieurs instruments, éventuellement avec des instruments percussifs comme de la batterie.

- *Pitchée :* la mélodie est jouée par un instrument voisé, et l'estimation se fait soit par la séquence des fréquences fondamentales jouées, soit par les notes de musique jouées.

- *Individuelle :* la mélodie est monophonique, jouée par un seul instrument à la fois. Cela dit, pour des raisons évoquées plus loin dans l'exposition du modèle, cette contrainte n'est imposée que très superficiellement.

- *Ligne :* la ligne mélodique doit être relativement "lisse", sans aspérité. L'instrument jouant la mélodie, l'instrument "principal", ne peut en général pas sauter de manière aléatoire d'une fréquence à l'autre, d'une note à l'autre, et doit plutôt présenter des paliers stables d'une note à l'autre.

- *Prédominante :* la notion de prédominance de la mélodie est probablement la plus difficile à cerner. Intuitivement, cela signifie que la mélodie principale est la séquence de sons que nous considérions comme caractéristique de la chanson, pour l'identifier, notamment. Concrètement, nous considérons dans nos travaux que cette prédominance est essentiellement une prédominance énergétique, de sorte que l'énergie de l'instrument principal domine la majorité du temps celles des autres instruments. Il faudra cependant souvent accomplir un compromis entre l'énergie et la régularité de la ligne mélodique, ce qui motive les modèles d'évolution temporelle choisis pour les séquences de fréquences fondamentales et de notes de la mélodie.

Nous avons suggéré plus haut qu'il y avait deux tâches particulières correspondant à la dénomination d'extraction de la mélodie : l'estimation de fréquences fondamentales (F0) prédominantes et la transcription de la mélodie en notes de musique. Nous donnons dans la suite de cette partie une définition pour chaque tâche, suivie de brefs états de l'art correspondants.

Definition 2 (Estimation de la fréquence fondamentale (F0) prédominante:) *Un système effectuant la tâche d'estimation de F0 prédominante doit prendre en entrée un fichier de musique digitalisée, et fournir en sortie un fichier de description de la séquence des F0. Chaque ligne de ce fichier comporte d'abord le temps où la fréquence est estimée, puis la valeur de cette fréquence en Hz. Pour la ligne correspondant à la trame n:*

```
<Temps de la trame n (s)> <Tabulation> <F0 à la trame n (Hz)>
```

Cette séquence de fréquences fondamentales doit correspondre à la mélodie, jouée par l'instrument principal.◇

Definition 3 (Transcription en notes de la mélodie:) *Un système visant à fournir une transcription en notes de musique de la mélodie doit renvoyer les notes MIDI, les onsets et les offsets pour chacune de ces notes. Une ligne du fichier à remplir doit ressembler à cela : A line of the output file may therefore look like the following:*

```
<Onset (s)> <Tabulation> <Offset (s)> <Tabulation> <Note MIDI>
```

◇

De nombreuses approches ont été proposées afin d'aborder la tâche d'estimation de F0, qui a d'ailleurs donné lieu à une évaluation dans le cadre des campagnes internationales annuelles de MIREX (*Music Information Retrieval Evaluation eXchange*, campagne d'évaluation sur la recherche d'information musicale). Les travaux de Goto [2000, 2004] sont sans doute les travaux fondateurs du domaine. L'auteur a en effet eu l'intuition que l'on pouvait extraire une information sémantique utile, la ligne mélodique et la ligne de basse, sans pour autant devoir décrire finement le signal. Une représentation temps-fréquence du signal est tout d'abord obtenue, utilisant notamment les travaux de Abe and Honda [2006] sur les spectrogrammes ré-assignés à l'aide de l'estimation des fréquences instantanées. Chaque colonne de cette transformée, correspondant chacune à une trame du signal, est alors normalisée, le résultat pouvant alors être considéré comme une densité de probabilité. Celle-ci est alors décomposée sur des mélanges de densités gaussiennes. Chacune de ces densités de probabilité est paramétrée par une fréquence fondamentale f_0 et une envelope spectrale. Les moyennes des composantes gaussiennes sont placées à

toutes les fréquences multiples de f_0, obtenant ainsi un peigne harmonique. Un algorithme Espérance-Maximisation (EM) permet d'estimer les poids attribués à chacune des F0, en maximisant un critère qu'il est intéressant de rapprocher de celui de Lee and Seung [2001]. La technique de Goto [2004] est d'ailleurs assez similaire aux récents travaux effectués sur la factorisation en matrices non-négatives (NMF en Anglais [Lee and Seung, 1999]).

De nombreuses techniques de décomposition se basent sur des algorithmes dits *gloutons*, avec des phases d'estimation suivie de soustraction itérées, comme par exemple le Matching Pursuit [Mallat and Zhang, 1993]. Pour le traitement de la musique, on retrouve ce type de procédé dans les algorithmes d'estimation de F0 multiples de de Cheveigné [1993] ou de Klapuri [2001]. L'approche de Goto [2004] permet, en théorie, d'estimer conjointement tous les paramètres de la décomposition, ce qui devrait limiter la propagation des erreurs d'une estimation à l'autre.

Une majorité des techniques proposées pour la détection de la mélodie est basée sur des sommes harmoniques (SHS [Hermes, 1988]) plus ou moins modifiées. Il en est ainsi pour le pré-traitement utilisé par Ryynänen and Klapuri [2005], ou le cœur même de systèmes tels que [Cao and Li, 2008, Hsu et al., 2009, Wendelboe, 2009]. Le système de Ryynänen and Klapuri [2005] est d'autant plus intéressant, qu'il utilise une modélisation statistique, avec des "note-events" qui permettent une interprétation très pratique des paramètres estimés. Ce dernier système est d'ailleurs l'un des rares systèmes proposés qui effectuent une transcription en note de la mélodie principale, et non une estimation de la séquence de fréquences fondamentales.

0.1.3 Séparation de l'instrument principal et de l'accompagnement

Séparer les différentes contributions d'un enregistrement sonore est un sujet populaire, avec des visées diverses et des approches variées. L'application d'une séparation de sources musicales est d'abord multiple : une telle séparation permet d'obtenir des voix séparées, réutilisables à des fins de re-mixages en studio, par exemple, notamment pour de vieux enregistrements. On peut aussi effectuer une indexation "simplifiée" sur des voix séparées : l'identification d'instruments est par exemple beaucoup plus difficile sur un mélange d'instruments que sur un instrument solo.

Les approches adoptées pour de telles séparations varient suivant le type de signal dont on dispose en entrée. Si un seul canal audio est disponible, alors on parle de séparation monaurale, avec plusieurs canaux, on parle de séparation multi-canale. Le second cas est souvent traité avec des méthodes plus ou moins génériques, fonctionnant pour des signaux non-nécessairement audio, comme la PCA [Pearson, 1901] et l'ICA [Jutten and Herault, 1991, Comon et al., 1994]. Utilisées dans ces circonstances, ces techniques correspondent souvent à effectuer des détections des directions d'arrivée des sons par rapport au champ de capteurs (microphones), et ensuite d'isoler les signaux venant des directions estimées. Cependant, cette information spatiale n'est pas toujours accessible et d'autres méthodes de décomposition, inspirées des premières, permettent malgré tout d'obtenir des résultats intéressants comme [Plumbley, 2003, Abdallah and Plumbley, 2004].

Avec des modèles de signaux spectraux tel que [Benaroya et al., 2006, Ozerov et al., 2007], les décompositions peuvent être interprétées d'un point de vue plus proche de la production même de ces signaux, avec des résultats en séparation très convaincant, et ce même sur des signaux mono-canaux. Ces travaux sont pour cette raison l'un des points de départ de la présente thèse. Le formalisme que nous avons développé s'est peu à peu approché du formalisme des travaux de Vincent [2004], qui considère lui aussi les deux

aspects de transcription et de séparation que nous étudions.

Enfin, plusieurs travaux se sont intéressés au problème de séparation de la voix chantée de l'accompagnement, voire, plus généralement, la séparation de l'instrument principal de l'accompagnement. Ainsi, le système de Ozerov et al. [2007] adapte des modèles spectraux pré-appris afin de mieux correspondre au signal, tout en ayant détecté au préalable les parties du mélange sonore où la voix est absente. Certains travaux [Lagrange et al., 2008, Li and Wang, 2007, Ryynänen et al., 2008] s'appuient sur des techniques d'analyse sinusoidale et des méthodes non supervisées afin de détecter les groupes de sinusoides correspondant à la voix désirée. De plus, les deux derniers travaux [Li and Wang, 2007, Ryynänen et al., 2008] reposent explicitement sur une estimation préalable de la mélodie jouée par l'instrument à séparer.

0.1.4 Contributions

Les modèles et algorithmes proposés dans cette thèse apportent plusieurs nouveautés par rapport aux systèmes d'estimation de mélodie et de séparation existants. Tout d'abord, bien que l'idée de décomposition sur des peignes harmoniques ne soit pas nouvelle [Goto, 2004], notre utilisation du **modèle source/filtre**, particulièrement adapté pour la voix chantée, ainsi que l'utilisation d'une famille de fonctions inspirée par le **mode de production physique** de la voix est inédit dans le domaine.

De plus, à la différence de beaucoup de systèmes d'estimation de la mélodie, nous proposons de **modéliser explicitement la partie accompagnement**. Ainsi, un modèle adapté aux instruments d'accompagnement peut-être utilisé, indépendamment du modèle choisi pour la voix principale. Nous utilisons en l'occurrence un modèle statistique équivalent à la **factorisation en matrices positives (NMF)** de la puissance de la TFCT de la partie d'accompagnement.

En plus d'une **interprétabilité intéressante des paramètres estimés**, en terme de F0 et de formants, un avantage de nos approches par rapport à d'autres méthodes est leur **non-supervision**. En effet, les décisions se basent sur des connaissances d'expert sur la mélodie et le signal d'intérêt, et non sur une base de données d'exemples.

Enfin, le cadre statistique permet d'élargir facilement les horizons de notre recherche, en incluant toujours plus d'*a priori* sur les paramètres ou de complexifier les modèles de signaux de base. Il est à noter que les interprétations que l'on peut donner aux paramètres permettent d'élaborer très intuitivement des distributions *a priori* pour ces paramètres.

0.2 Modèles de signaux

Nous avons développé deux modèles de signaux permettant de détecter la mélodie, en terme de fréquences fondamentales (en Hz). Pour cela, un modèle source/filtre a été intégré aux modèles de signaux pour la séparation de source de Benaroya et al. [2006], Ozerov et al. [2007].

Tout d'abord, le modèle statistique choisi est décrit de manière générale. Nous présentons ensuite le modèle à mélange de Gaussiennes amplifiées (MMGA) de Benaroya et al. [2006], Ozerov et al. [2007], adapté au modèle source/filtre. Ensuite, un modèle alternatif est proposé : ce modèle "à mélange instantané" vise essentiellement à approcher le modèle précédent tout en diminuant les calculs nécessaires à l'estimation des paramètres. Enfin, nous aborderons la manière de modéliser les dépendances temporelles dans chacun des modèles statistiques proposés.

0.2.1 Modèle gaussien pour la transformée de Fourier des signaux

La représentation choisie est la transformée de Fourier à court terme (TFCT ou STFT en Anglais). Une telle représentation temps-fréquence permet d'observer l'évolution des énergies associées à plusieurs bandes de fréquences au cours du temps. Le calcul de la TFCT correspond au calcul de la transformée de Fourier (TF) de fenêtres de signal recouvrantes. Pour un signal temporel y_t, avec t l'indice de temps, alors la matrice de TFCT $\mathbf{Y}$, de taille $F \times N$, est construite en concaténant les vecteurs de TF de sorte à ce qu'ils soient les colonnes de $\mathbf{Y}$.

Le modèles spectraux étudiés dans le présent travail consiste à considérer que chaque TF est une variable aléatoire. Pour un son y dit "élémentaire", la TF $\mathbf{y}$ suit une distribution gaussienne complexe multivariée, de moyenne nulle et de matrice de covariance $\mathbf{\Sigma}^y$ qui caractérise cet élément sonore:

$$\mathbf{y} \sim \mathcal{N}_c(\mathbf{0}, \mathbf{\Sigma}^y)$$

De plus, cette matrice de covariance qui caractérise complètement le son élémentaire est supposée diagonale. Le vecteur de diagonale est dénoté $\mathbf{s}^y$. Sous certaines conditions, ce vecteur peut être assimilé à la densité spectrale de puissance (DSP ou PSD en Anglais), mais on pourra aussi l'interpréter comme une approximation de la puissance de l'élément sonore dans chacune des bandes de fréquences de la TF.

Les modèles présentés dans cette thèse ont principalement pour but de paramétrer les variances des Gaussiennes pour les différents sons élémentaires considérés. Cela permet d'abord de s'assurer qu'ils répondent bien aux caractéristiques désirées pour les sources auxquelles ils correspondent : l'instrument principal ou l'accompagnement. Par ailleurs, l'agencement de ces différents éléments peut aussi être contrôlé afin de correspondre au mieux à un mode de production réel. Les combinaisons de ces "atomes" donnent le mélange observé, et les conditions dont le mélange se forme peuvent être implémentées dans le modèle statistique proposé. Par exemple, pour l'instrument principal, on souhaitera plutôt que les atomes s'excluent mutuellement : il ne peut y avoir qu'un seul atome actif par trame de signal. A l'inverse, on pourra envisager des mélanges de plusieurs atomes pour former l'accompagnement.

0.2.2 Modèle à Mélange de Gaussiennes Amplifiées avec Source/Filtre

Notre premier modèle est une extension hybride du modèle à mélange de Gaussiennes amplifiées (MMGA ou GSMM en Anglais) proposé par Benaroya et al. [2006], Ozerov et al. [2007] pour séparer la voix parlée ou chantée d'un fond sonore musical.

Les différentes contributions du modèle proposé sont les suivantes : les deux sources à séparer, la voix chantée et la musique dans notre cas, sont modélisées par des modèles génératifs de signaux différents. Les travaux pré-cités sont basés sur une supervision préalable permettant de caractériser ces sources grâce à des exemples donnés. Nous proposons un modèle où la supervision reste possible, bien qu'optionnelle. Les modèles génératifs doivent donc être suffisamment discriminants. Un modèle source/filtre est utilisé pour la voix chantée, alors qu'un modèle plus permissif permet à l'accompagnement d'être aussi varié que ce à quoi l'on pourrait s'attendre. Le modèle de source/filtre choisi pour l'instrument principal permet de réduire le nombre d'éléments dans la base de spectres, entrainant ainsi une réduction de la complexité de la mise en service.

Dans cette section, nous détaillons d'abord le modèle pour le mélange, celui de l'instrument principal, puis le modèle pour l'accompagnement.

Le mélange sonore traité $\mathbf{X}$ est supposé être le mélange instantané entre deux contributions, la voix principale $\mathbf{V}$ et l'accompagnement $\mathbf{M}$:

$$\mathbf{X} = \mathbf{V} + \mathbf{M}$$

De manière générique, pour la trame n, le vecteur de TF de l'instrument principal $\mathbf{v}_n$ (resp. l'accompagnement $\mathbf{m}_n$) est supposé suivre une loi gaussienne de moyenne nulle et de matrice de covariance diagonale, avec pour diagonale $\mathbf{s}_n^V$ (resp. $\mathbf{s}_n^M$):

$$\mathbf{v}_n \sim \mathcal{N}_c(\mathbf{0}, \mathrm{diag}(\mathbf{s}_n^V))$$
$$\mathbf{m}_n \sim \mathcal{N}_c(\mathbf{0}, \mathrm{diag}(\mathbf{s}_n^M))$$

Ces deux contributions sont par ailleurs supposées indépendantes, de sorte que $\mathbf{x}_n$ suit aussi une loi gaussienne, de moyenne nulle et de matrice de covariance égale à $\mathrm{diag}(\mathbf{s}_n^V + \mathbf{s}_n^M)$.

L'**instrument principal** est en général une voix chantée. Le modèle génératif source/filtre est donc particulièrement adapté aux signaux considérés. Ce modèle ne limite d'ailleurs pas le type d'instrument principal à la seule voix chantée, et s'étend à de nombreux autres instruments, comme le saxophone ou la trompette, et plus généralement les instruments à vent. Le signal $\mathbf{v}_n$, à la trame n, est modélisé par un MMGA, dont les états cachés sont des états $Z_n = (Z_n^\Phi, Z_n^{F_0})$. $Z_n^\Phi = k$, $k = 1 \dots K$, est l'état de la partie filtre de $\mathbf{v}_n$, et est caractérisé par la réponse en fréquence $\mathbf{w}_k^\Phi$. $Z_n^{F_0} = u$, $u = 1 \dots U$, est l'état de la partie source, caractérisé par des spectres de puissance $\mathbf{w}_u^{F_0}$, qui sont des peignes harmoniques paramétrés par une fréquence fondamentale $f_0 = \mathcal{F}(u)$. La puissance spectrale résultante est le produit terme à terme de ces deux vecteurs, $\mathbf{w}_k^\Phi \bullet \mathbf{w}_u^{F_0}$. De plus, dans le cadre du MMGA, l'amplitude b_{kun} permet d'ajuster l'énergie des formes spectrales à l'énergie du signal et $\mathbf{v}_n$ vérifie donc, conditionnellement à l'état $Z_n = (k, u)$:

$$\mathbf{v}_n | \{Z_n = (k,u)\} \sim \mathcal{N}_c(0, b_{kun}\mathrm{diag}(\mathbf{w}_k^\Phi \bullet \mathbf{w}_u^{F_0})) \quad (1)$$

Le modèle décrit dans l'équation (1) est dénoté GSMM dans nos travaux. La variance de l'instrument principal, pour l'état $Z_n = (k, u)$, est dénotée par $\mathbf{s}_n^{V,\mathrm{GSMM}|ku} = b_{kun}\mathbf{w}_k^\Phi \bullet \mathbf{w}_u^{F_0}$.

Les spectres de la partie filtre, $\mathbf{w}_k^\Phi$, peuvent être contraints à être lisse par construction. Ces spectres sont supposés être le résultat de combinaisons linéaires non-négatives d'une famille de fonctions $\mathbf{W}^\Gamma$, de sorte que la matrice $\mathbf{W}^\Phi$ soit égale au produit $\mathbf{W}^\Gamma\mathbf{H}^\Gamma$. $\mathbf{H}^\Gamma$ est une matrice de facteurs d'amplitude non-négatifs. Le lien évident entre ce formalisme et la factorisation en matrices non-négatives (NMF en Anglais) nous permettra plus loin d'obtenir des algorithmes d'estimation inspirés des techniques existantes pour la NMF. L'enveloppe spetrale pour le filtre k est alors donnée par:

$$\mathbf{w}_k^\Phi = \mathbf{W}^\Gamma \mathbf{h}_k^\Gamma = \sum_p \mathbf{w}_p^\Gamma h_{pk}^\Gamma$$

Ainsi, pour SGSMM, on a $\mathbf{s}_n^{V,\mathrm{SGSMM}|ku} = b_{kun}(\mathbf{W}^\Gamma\mathbf{h}_k^\Gamma) \bullet \mathbf{w}_u^{F_0}$.

L'**accompagnement** est paramétré comme les signaux audio dans [Benaroya et al., 2003], qui est un cadre formellement équivalent à la NMF de $\mathbf{S}^M$, comme montré par Févotte et al. [2009a]. La variance correspondant à l'accompagnement est une combinaison linéaire non-négative de R formes spectrales $\mathbf{w}_r^M$, $1 \geq r \geq R$: $\mathbf{s}_n^M = \mathbf{W}^M\mathbf{h}_n^M$. Les coefficients d'amplitudes $\mathbf{h}_n^M$ permettent d'adapter l'énergie des formes spectrales (dont l'énergie est normalisée) au signal. Si un coefficient h_{rn}^M est nul, cela signifie que l'élément r de la matrice $\mathbf{W}^M$ est absent du signal à la trame n.

Finalement, pour le (S)GSMM, la mixture sonore a pour variance la quantité suivante :

$$\mathbf{s}_n^{\mathrm{GSMM}|ku} = b_{kun}\mathbf{w}_k^{\Phi} \bullet \mathbf{w}_u^{F_0} + \mathbf{W}^M\mathbf{h}_n^M$$

Cela revient à dire que la vraisemblance de la TFCT du mélange sonore $\mathbf{X}$ est la somme des probabilités conditionnelles, pondérée par les probabilités *a priori* :

$$\mathbf{x}_n \sim \sum_{k,u} \pi_{ku}\mathcal{N}_c(\mathbf{0}, b_{kun}\mathbf{w}_k^{\Phi} \bullet \mathbf{w}_u^{F_0} + \mathbf{W}^M\mathbf{h}_n^M)$$

Pour le SGSMM, il faut remplacer les formes spectrales des filtres $\mathbf{w}_k^{\Phi}$ par leur décomposition sur $\mathbf{W}^{\Gamma}$. Les paramètres à estimer pour le GSMM forment l'ensemble $\mathbf{\Theta}^{\mathrm{GSMM}} = \{\mathbf{B}, \mathbf{W}^{\Phi}, \mathbf{W}^M, \mathbf{H}^M\}$, et pour le SGSMM, il s'agit de $\mathbf{\Theta}^{\mathrm{SGSMM}} = \{\mathbf{B}, \mathbf{H}^{\Phi}, \mathbf{W}^M, \mathbf{H}^M\}$. Les coefficients d'énergie de l'instrument principal forment un tenseur $\mathbf{B}$ de taille $K \times U \times N$. Les matrices dictionnaires $\mathbf{W}^{F_0}$ pour la source de l'instrument principal et $\mathbf{W}^{\Gamma}$ pour le SGSMM sont fixées à l'avance : les détails pour la création de $\mathbf{W}^{F_0}$ sont donnés dans l'Annexe C et les informations sur la matrice $\mathbf{W}^{\Gamma}$ choisie se trouvent en Section 3.3.2.1. Les matrices de formes spectrales $\mathbf{W}^M$ et $\mathbf{W}^{\Phi}$ sont apprises directement sur le signal à analyser. La table 3.1 donne en détail ces paramètres, leur signification ainsi que les notations associées.

0.2.3 Modèle de mélange instantané

Le modèle proposé dans la section précédente est, sous certains aspects, relativement réaliste. En effet, pour l'instrument principal, pour une trame donnée, il ne peut y avoir qu'un seul état actif, en d'autres termes, une seule fréquence fondamentale avec une seule enveloppe spetrale. Cela est certes réaliste, mais pose des problèmes, notamment lors des phases d'estimation des paramètres. En effet, comme nous le verrons, le (S)GSMM nécessite un algorithme Espérance-Maximisation (EM) qui s'avère être difficile à configurer, cela à cause de la présence des états cachés.

Une alternative au (S)GSMM est donc aussi proposé dans cette thèse. En premier lieu, cet autre model permet de réduire la difficulté de la phase d'estimation, car nous n'y considérons plus de modèle à états cachés. Par ailleurs, les résultats montrent que, malgré les approximations "théoriques" faites sur le modèle de production, on peut obtenir des données qui restent cohérentes avec la définition de la mélodie et de l'instrument principal.

On se propose de remodeler la partie du modèle qui concerne l'instrument principal. La transformation opérée ici est analogue au passage qui s'est opéré entre les travaux de Benaroya et al. [2006], GSMM à l'origine, et ses travaux sur les représentations nonnégatives [Benaroya et al., 2003]. En effet, alors que pour le GSMM, la vraisemblance du signal de l'instrument principal était le "mélange" des vraisemblances conditionnelles aux états cachés, nous avons proposé de considérer que le signal audio lui-même est le mélange d'éléments sonores $\boldsymbol{\nu}_n^{ku}$ correspondant à chacune des fréquences et chacun des filtres :

$$\mathbf{v}_n = \sum_{ku} \boldsymbol{\nu}_n^{ku}$$
$$\boldsymbol{\nu}_n^{ku} \sim \mathcal{N}_c\left(\mathbf{0}, h_{kn}^{\Phi}h_{un}^{F_0}\mathrm{diag}(\mathbf{w}_k^{\Phi} \bullet \mathbf{w}_u^{F_0})\right)$$

où l'amplitude pour le couple filtre-source (k, u) a été séparée en deux contributions h_{kn}^{Φ} pour la partie filtre et $h_{un}^{F_0}$ pour la source. Il est en effet plus pratique de considérer ces deux contributions séparément pour la phase d'estimation de paramètres. De plus,

l'interprétation de ces coefficients est plus aisée et permet de déterminer plus directement les pitchs dominants recherchés. Ce modèle est dénoté IMM (*Instantaneous Mixture Model*) ou SIMM (*Smooth-filter IMM*).

Le précédent modèle pour l'accompagnement ne posant pas de problème d'estimation particulier, il peut être réutilisé pour ce nouveau modèle.

Enfin, le mélange sonore $\mathbf{X}$ vérifie, dans ce cas:

$$\mathbf{x}_n \sim \mathcal{N}_c\left(\mathbf{0}, \operatorname{diag}\left(\left(\sum_k h_{kn}^{\Phi}\mathbf{w}_k^{\Phi}\right) \bullet \left(\sum_u h_{un}^{F_0}\mathbf{w}_u^{F_0}\right) + \sum_r h_{rn}^{M}\mathbf{w}_r^{M}\right)\right)$$
$$\mathbf{x}_n \sim \mathcal{N}_c\left(\mathbf{0}, \operatorname{diag}\left(\left(\mathbf{W}^{\Phi}\mathbf{h}_n^{\Phi}\right) \bullet \left(\mathbf{W}^{F_0}\mathbf{h}_n^{F_0}\right) + \mathbf{W}^{M}\mathbf{h}_n^{M}\right)\right)$$

Dans cette dernière équation, le lien entre notre modèle et la modélisation par NMF est mis en évidence par les notations matricielles adoptées. Ce lien permet, là encore, d'élaborer des techniques d'estimation à partir des stratégies existantes pour les estimations NMF.

L'ensemble des paramètres à estimer pour l'IMM est $\mathbf{\Theta}^{\text{IMM}} = \{\mathbf{W}^{\Phi}, \mathbf{H}^{\Phi}, \mathbf{H}^{F_0}, \mathbf{W}^{M}, \mathbf{H}^{M}\}$ et pour le SIMM $\mathbf{\Theta}^{\text{SIMM}} = \{\mathbf{H}^{\Gamma}, \mathbf{H}^{\Phi}, \mathbf{H}^{F_0}, \mathbf{W}^{M}, \mathbf{H}^{M}\}$. La table 3.2 donne le détail de ces paramètres. Les matrices de dictionnaires de formes $\mathbf{W}^{F_0}$ et $\mathbf{W}^{\Gamma}$ sont fixées à l'avance, comme pour le GSMM.

0.2.4 Modèle pour l'évolution temporelle

Les deux modèles proposés précédemment visent principalement à expliquer le signal d'une trame à l'autre, indépendamment. Il est possible d'ajouter des contraintes d'évolution entre les états cachés, mais aussi de rajouter une nouvelle couche d'états correspondant aux notes qui sont supposées être jouées.

Afin d'intégrer ces deux niveaux d'évolution temporelle, il est pratique de définir **un cadre statistique commun** pour le (S)GSMM et le (S)IMM. Un cadre bayésien est particulièrement adapté dans notre cas. Notons $\mathbf{\Theta}$ l'ensemble de paramètres pouvant être ceux du (S)GSMM ou du (S)IMM. Cet ensemble peut être considéré comme une variable aléatoire, au même titre que les observations $\mathbf{X}$, ou les états cachés Z. Dans ce cas, nous nous intéressons à la probabilité jointe de ces variables. Les dépendances sont données par le graphe de la Figure 3.17, ce qui peut s'écrire :

$$p(\mathbf{X}, \mathbf{\Theta}, Z^{\Phi}, Z^{F_0}) = p(\mathbf{X}|\mathbf{\Theta})p(\mathbf{\Theta}|Z^{\Phi}, Z^{F_0})p(Z^{\Phi})p(Z^{F_0})$$

Nous pouvons aussi ajouter un état caché représentant la note de musique voulue pour l'instrument principal, E. Cet état prend ses valeurs dans l'ensemble des notes de musique (sur l'échelle occidentale), et concrètement, il est pratique de se servir de l'échelle des notes MIDI pour cela. Avec les dépendances de la Figure 3.17, on obtient :

$$p(\mathbf{X}, \mathbf{\Theta}, Z^{\Phi}, Z^{F_0}, E) = p(\mathbf{X}|\mathbf{\Theta})p(\mathbf{\Theta}|Z^{\Phi}, Z^{F_0})p(Z^{\Phi})p(Z^{F_0}|E)p(E)$$

Les observations d'une trame à l'autre sont indépendantes conditionnellement aux paramètres $\mathbf{\Theta}$, de même, les paramètres d'une trame à l'autre sont indépendants conditionnellement aux états Z^{Φ} et Z^{F_0}. Ainsi, on a :

$$p(\mathbf{X}|\mathbf{\Theta}) = \prod_n p(\mathbf{x}_n|\mathbf{\Theta}_n)$$
$$p(\mathbf{\Theta}|Z^{\Phi}, Z^{F_0}) = \prod_n p(\mathbf{\Theta}_n|Z_n^{\Phi}, Z_n^{F_0})$$

Il est à noter que l'expression de la probabilité conditionnelle de $\mathbf{\Theta}^{(\mathrm{S})\mathrm{IMM}}$ sachant Z peut être définie de sorte à ce que le modèle (S)IMM devienne un (S)GSMM. En effet, pour le (S)GSMM, les seuls paramètres pour lesquels cette probabilité est non-nulle sont ceux dont les indices correspondent aux états Z^{Φ} et Z^{F_0} :

$$p(\mathbf{\Theta}_n^{(\mathrm{S})\mathrm{IMM}}|Z_n^{\Phi}=k, Z_n^{F_0}=u) \neq 0 \Rightarrow \forall (i,j) \neq (k,u), h_{in}^{\Phi} h_{jn}^{F_0} = 0$$

Plus généralement, on pourra utiliser cette probabilité de $\mathbf{\Theta}$ conditionnellement à Z pour contraindre les paramètres à correspondre, par exemple, au couple d'état donné, mais aussi pour ajouter d'autres contraintes sur les paramètres, comme par exemple des contraintes de parcimonie comme [Mohimani et al., 2008] ou de régularité temporelle [Bertin et al., 2010].

L'évolution de la séquence Z^{Φ} est indépendante de celles de Z^{F_0} et de E, la partie filtre et la partie source étant classiquement découplées dans ce modèle source/filtre. Z^{Φ} suit un modèle de Markov. Conditionnellement à la séquence de notes E, **la probabilité de** Z^{F_0} est le produit de la probabilité d'une évolution de type markovien sur les états cachés Z^{F_0}, et de la probabilité d'avoir un état de fréquence fondamentale $Z_n^{F_0}$ à une trame n, sachant que la note est E_n. Enfin **l'évolution de la séquence des notes** E est régie par un modèle qui prend explicitement en compte les durées des notes de la séquence. Ce modèle est adaptée de [Vincent, 2004]. On peut alors écrire :

$$p(Z^{\Phi}) = p(Z_1^{\Phi}) \prod_n p(Z_n^{\Phi}|Z_{n-1}^{\Phi})$$

$$p(Z^{F_0}|E) = p(Z_1^{F_0}|E_1) \prod_n p(Z_n^{F_0}|Z_{n-1}^{F_0}, E_n, E_{n-1})$$

$$p(E_{1:n}) = p(E_n|E_{1:n-1}) p(E_{1:n-1}), \forall n \in [1, N]$$

Pour les évolutions des séquences de filtre Z^{Φ} et de source Z^{F_0}, nous utilisons essentiellement une structure de Markov, avec des probabilités *a prior* et de transition paramétrées. Ces quantités sont fixées à l'avance, fournissant des distributions cohérentes pour le "sens commun". Ainsi, nous avons fixé **l'évolution "physique" du pitch** d'une trame à la suivante, avec $Q(u,v)$ la probabilité de transition de l'état u vers l'état v, par une loi exponentielle, de paramètre α :

$$Q(u,v) \propto \exp(\alpha.\mathrm{round}(|12\delta|), \text{ où } \delta = \log_2 \mathcal{F}(v) - \log_2 \mathcal{F}(u)$$

Une telle probabilité pénalise les passages entre des pitches éloignés, par paliers de demi-tons. Cela est relativement réaliste et cohérent avec les observations faites sur la base de développement de la tâche d'estimation de F0 prépondérantes de la campagne d'évaluation MIREX. La valeur de α permet de contrôler la "force" de la contrainte. Ainsi une forte valeur entraînera des lignes mélodiques plutôt horizontales, alors qu'une valeur plus faible permettra de plus grandes fluctuations, prenant en compte par exemple des effets de type vibrato ou tremollo. Une valeur de l'ordre de grandeur de 10 permet d'obtenir des résultats corrects. Il est à noter que la valeur d'α peut varier suivant le modèle choisi (GSMM ou IMM), car les probabilités des observations sachant les états utilisées ne sont pas forcément les mêmes dans les deux cas.

De même pour **la séquence des états du filtre**, Z^{Φ}, nous avons opté pour une solution paramétrique et fixée à l'avance. Les probabilités de transition sont assez franches, avec pour le passage du filtre k au filtre l :

$$Q^{\Phi}(k,l) \propto \begin{cases} 1, \text{ si } k = l \\ \epsilon^{\Phi} << 1, \text{ si } k \neq l \end{cases}$$

La valeur d'ϵ^{Φ} petite correspond à la volonté de simuler des états sur lesquels on ne peut que rester un certain temps avant de pouvoir changer d'état. L'inconvénient des modèles de Markov caché (HMM en Anglais) vient de la difficulté à modéliser précisément la durée des épisodes statiques d'un état à l'autre. Or, typiquement, pour la partie filtre, il semble plus raisonnable de penser que la probabilité de rester sur un même filtre, donc un même phonème ou même timbre, est grande par rapport à la probabilité de changer d'état. L'évolution de la partie filtre devrait donc intuitivement être fondamentalement différente de celle de la partie source. L'approximation proposée ici permet malgré tout d'obtenir les résultats souhaités, notamment parce que la partie d'intérêt pour nos travaux est la partie source. Cela étant, si des travaux futurs devaient s'intéresser plus précisément aux filtres estimés, une meilleure modélisation de leur évolution, par exemple en adoptant un modèle similaire au modèle de durée esquissé ci-après, serait probablement nécessaire.

Enfin, la nature et l'évolution des pitches ne sont pas indépendantes de la séquence de notes qui la régisse, E. Nous proposons de réduire l'influence de la couche E sur la couche Z^{F_0} en adoptant la simplification suivante :

$$p(Z_n^{F_0}|Z_{n-1}^{F_0}, E_n, E_{n-1}) \propto p(Z_n^{F_0}|Z_{n-1}^{F_0})p(Z_n^{F_0}|E_n)$$

Cela signifie que nous supposons que l'évolution des pitches ne dépend que de la couche Z^{F_0}, avec la structure markovienne précédemment introduite, alors que la couche E influe uniquement localement sur la valeur même de l'état Z^{F_0}. En effet, la fréquence fondamentale correspondante, $\mathcal{F}(Z_n^{F_0})$, doit être "proche" de la fréquence standard correspondant à la note E_n. Cette proximité de sorte que la distribution du pitch sachant E_n suive une loi gaussienne de centre la fréquence standard de E_n, notée $\mathcal{F}^{\text{MIDI}}(E_n)$, et de variance égale à une valeur σ^{MIDI}, fixée arbitrairement dans nos expériences à environ un ton et demi :

$$p(Z_n^{F_0} = u|E_n = n^{\text{MIDI}}) \propto \exp\left(-\frac{(\log_2 \mathcal{F}(u) - \log_2 \mathcal{F}^{\text{MIDI}}(n^{\text{MIDI}}))^2}{2(\sigma^{\text{MIDI}})^2}\right)$$

Le **modèle de durée** utilisé est quant à lui le modèle segmental tel que proposé par Vincent [2004]. Grâce au formalisme adopté dans nos travaux, très proche de celui de Vincent [2004], il est possible d'intégrer le modèle original pour la couche des notes E sans modification significative. Dans un premier temps, la probabilitié de cette séquence d'états ne dépend que des durées des notes qui la composent. Il sera possible, plus tard, d'ajouter d'autres composantes, notamment la valeur des notes sur l'échelle musicale. En incluant de telles informations sur les notes, il devient possible de contraindre les notes, notamment si la tonalité est connue ou estimée, ou si les accords de l'accompagnement si disponible.

Soient $\mathcal{D}^{\text{seg}}(n)$ et $\mathcal{D}^{\text{note}}(n)$ les probabilités qu'un segment ou qu'une note dure n trames. La probabilité de la séquence de notes E s'écrit donc :

$$p(E) = \mathcal{D}^{\text{seg}}(n_1)\prod_{l=1}^{L}\mathcal{D}^{\text{note}}(d_l)\mathcal{D}^{\text{seg}}(n_{l+1} - n_l)$$

où la durée des notes est successivement notée, par exemple pour la l-ième note, d_l. Les temps de début de notes ("onsets") correspondant sont notés n_l. Les probabilités des durées sont des probabilités log-gaussiennes, comme celles proposées par Vincent [2004]. La modularité du modèle permet de définir des *a priori* différents. Si l'on dispose par exemple d'indications métronomiques, d'une estimation des temps et des mesures, des densités de probabilité avec plusieurs modes, correspondant chacun à une valeur standard de durée de note (double-croche, croche, noire, etc.) seraient sans doute plus appropriées.

0.3 Estimation des paramètres et des séquences cachées

Dans le cadre bayésien commun défini précédemment, nous désirons estimer les paramètres et séquences d'intérêt, $\mathbf{\Theta}$, Z^{Φ}, Z^{F_0} et E, étant donné le signal observé $\mathbf{X}$. Nous avons cherché à les estimer par maximum *a posteriori* (MAP) :

$$\widehat{\mathbf{\Theta}}, \widehat{Z}^{\Phi}, \widehat{Z}^{F_0}, \widehat{E} = \underset{\mathbf{\Theta}, Z^{\Phi}, Z^{F_0}, E}{\arg\max}\ p(\mathbf{X}, \mathbf{\Theta}, Z^{\Phi}, Z^{F_0}, E)$$

Cette dernière équation s'avère être difficile à résoudre directement. Nos travaux ont porté sur diverses manières d'estimer les quantités désirées. Ainsi, cinq systèmes différents sont proposés dans cette thèse. Trois d'entre eux visent à estimer la séquence de fréquences fondamentales de l'instrument principal Z^{F_0}. Un autre estime la séquence de notes jouées par l'instrument principal E. Enfin, un dernier système sépare les deux signaux modélisés, c'est-à-dire la voix principale et l'accompagnement.

Dans cette section, ces systèmes sont d'abord décrits dans les détails. Ensuite, les méthodes d'estimation utilisées, parfois par plusieurs systèmes, sont présentées : la méthode de gradient multiplicatif, issue des algorithmes de NMF classiques, pour le modèle (S)IMM, puis l'algorithme d'espérance-maximisation généralisé (GEM) nécessaire pour le modèle (S)GSMM. Enfin les méthodes de décodage des séquences sont données : l'algorithme de Viterbi pour les structures HMM puis l'algorithme de recherche par faisceaux pour la suite de notes.

0.3.1 Description des systèmes proposés

Au lieu d'estimer toutes les couches du modèles en même temps, il est possible de les estimer les unes après les autres, en ajoutant au fur et à mesure ces couches ou les dépendances correspondantes. Ce principe d'estimation sous-optimale a été appliqué dans le cas des modèles présentés dans cette thèse et a abouti à cinq systèmes différant au niveau du but visé ou du niveau d'approximation considéré.

Tout d'abord, trois systèmes sont proposés pour **estimer la séquence de fréquences fondamentales de l'instrument principal**. Ces systèmes prennent en entrée le signal sonore du mélange "instrument principal / accompagnement", et retournent en sortie la séquence $\widehat{Z}^{F_0}$ estimée. Les deux premiers systèmes ne considèrent, dans une première passe d'estimation, que les dépendances "verticales", ignorant les dépendances temporelles des séquences pour la source et le filtre. Ces dernières sont intégrés dans une seconde passe, en utilisant l'algorithme de Viterbi, notamment, pour décoder la structure HMM de ces séquences. Ces deux systèmes diffèrent par le modèle trame-à-trame sous-jacent, avec respectivement le (S)GSMM et le (S)IMM. Enfin le troisième système intègre lui toutes les dépendances des séquences dès la première passe d'estimation des paramètres $\mathbf{\Theta}^{(\mathrm{S})\mathrm{GSMM}}$. Pour ces trois systèmes, seules les couches d'observation $\mathbf{X}$, de paramètres $\mathbf{\Theta}$ et de séquences source/filtre $Z = (Z^{\Phi}, Z^{F_0})$ sont nécessaires, la couche E n'étant utilisée que pour le système suivant. Les étapes de chacun de ces systèmes sont donnés ci-dessus.

F-I **Estimation de F0 prédominante avec le (S)GSMM :**

1. Estimation des paramètres :

$$(\widehat{\mathbf{\Theta}}^{(\mathrm{S})\mathrm{GSMM}})^{(i)} = \underset{\mathbf{\Theta}^{(\mathrm{S})\mathrm{GSMM}}}{\arg\max}\ E\left[\log p(\mathbf{X}, Z^{F_0}, Z^{\Phi}; \mathbf{\Theta}^{(\mathrm{S})\mathrm{GSMM}}) | \mathbf{X}; (\mathbf{\Theta}^{(\mathrm{S})\mathrm{GSMM}})^{(i-1)}\right]$$

for $i \in [1, I]$

où I est le nombre d'itérations pour l'algorithme d'estimation.

2. Décodage de séquence :

$$\widehat{Z}^{F_0}, \widehat{Z}^{\Phi} = \arg\max_{Z^{F_0}, Z^{\Phi}} \prod_n p(\mathbf{x}_n | Z_n^{F_0}, Z_n^{\Phi}; \widehat{\mathbf{\Theta}}^{(\mathrm{S})\mathrm{GSMM}}) \times p(Z_1^{F_0}, Z_1^{\Phi}) \prod_{n>1} p(Z_n^{F_0}, Z_n^{\Phi} | Z_{n-1}^{F_0}, Z_{n-1}^{\Phi})$$

L'estimation des paramètres pour le système F-I est décrite en détail en Section 0.3.3, où l'algorithme GEM utilisé est aussi présenté. Afin de décoder la séquence Z^{F_0}, pour laquelle un HMM est utilisé, nous utilisons l'algorithme de Viterbi, appliqué en supposant que les paramètres estimés en première passe (sans structure HMM) ne sont pas significativement différents de ceux qui auraient été estimés avec la structure HMM. Le troisième système proposé a pour but d'intégrer cette structure dès la première passe d'estimation. L'algorithme de Viterbi est donné en Section 0.3.4.1.

Ce système a été publié dans [Durrieu et al., 2010], avec une participation dans deux campagnes d'évaluation MIREX en 2008 et 2009, avec des description dans les résumés longs [Durrieu et al., 2008c, 2009c].

F-II **Estimation de F0 prédominante avec le (S)IMM :**

1. Estimation de paramètres :

$$\widehat{\mathbf{\Theta}}^{(\mathrm{S})\mathrm{IMM}} = \arg\max_{\mathbf{\Theta}^{(\mathrm{S})\mathrm{IMM}}} p(\mathbf{X} | \mathbf{\Theta}^{(\mathrm{S})\mathrm{IMM}})$$

2. Décodage de séquence :

$$\widehat{Z}^{F_0} = \arg\max_{Z^{F_0}} \prod_n p(\mathbf{x}_n | \widehat{\mathbf{\Theta}}_n^{(\mathrm{S})\mathrm{IMM}}) p(\widehat{\mathbf{\Theta}}_n^{(\mathrm{S})\mathrm{IMM}} | Z_n^{F_0}) \times p(Z_1^{F_0}) \prod_{n>1} p(Z_n^{F_0} | Z_{n-1}^{F_0})$$

L'estimation des paramètres de $\mathbf{\Theta}^{(\mathrm{S})\mathrm{GSMM}}$ se fait par un algorithme de gradient multiplicatif directement adapté des techniques classiques de NMF, comme expliqué en Section 0.3.2. Cela est rendu possible par le fait que l'on ne considère que les premières dépendances, ignorant les dépendances temporelles des séquences Z^{Φ} et Z^{F_0}. Ces dépendances sont ajoutées, comme pour le système F-I, lors du décodage des séquences, utilisant encore une fois l'algorithme de Viterbi.

Il est à noter, pour l'algorithme de Viterbi, que la probabilité $p(\mathbf{x}_n | \widehat{\mathbf{\Theta}}_n^{(\mathrm{S})\mathrm{IMM}}) p(\widehat{\mathbf{\Theta}}_n^{(\mathrm{S})\mathrm{IMM}} | Z_n^{F_0} = u)$ et plus particulièrement la probabilité conditionnelle $p(\widehat{\mathbf{\Theta}}_n^{(\mathrm{S})\mathrm{IMM}} | Z_n^{F_0} = u)$ reste à être définie. Il est pratique de la définir proportionnelle au coefficient d'amplitude $h_{un}^{F_0}$. En effet, cela permet de prendre en compte l'énergie de la fréquence fondamentale associée u, tout en conservant une certaine interprétabilité du résultat. Une telle définition est, en un sens, réaliste, car on peut s'attendre à ce que la probabilité de $Z_n^{F_0} = u$ sachant les paramètres $\widehat{\mathbf{\Theta}}_n^{(\mathrm{S})\mathrm{IMM}}$ soit elle-même proportionnelle à $h_{un}^{F_0}$, ce qui est vérifié pour notre définition grâce au théorème de Bayes.

Ce système a été publié dans [Durrieu et al., 2008a, 2010], avec une participation dans deux campagnes d'évaluation MIREX en 2008 et 2009, dont les descriptions sont données dans les résumés longs [Durrieu et al., 2008c, 2009c].

F-III **Estimation de la F0 prédominante avec la structure de HMM :** les équations pour les estimations sont identiques à celles de F-I :

1. Estimation de paramètres :

$$(\widehat{\mathbf{\Theta}}^{(\mathrm{S})\mathrm{GSMM}})^{(i)} = \underset{\mathbf{\Theta}^{(\mathrm{S})\mathrm{GSMM}}}{\arg\max}\, E[\log p(\mathbf{X}, Z^{F_0}, Z^{\Phi}; \mathbf{\Theta}^{(\mathrm{S})\mathrm{GSMM}})|\mathbf{X}; (\mathbf{\Theta}^{(\mathrm{S})\mathrm{GSMM}})^{(i-1)}] \quad \text{for } i \in [1, I]$$

2. Décodage de séquence :

$$\widehat{Z}^{F_0}, \widehat{Z}^{\Phi} = \underset{Z^{F_0}, Z^{\Phi}}{\arg\max} \prod_n p(\mathbf{x}_n|Z_n^{F_0}, Z_n^{\Phi}; \widehat{\mathbf{\Theta}}^{(\mathrm{S})\mathrm{GSMM}}) \times p(Z_1^{F_0}, Z_1^{\Phi}) \prod_{n>1} p(Z_n^{F_0}, Z_n^{\Phi}|Z_{n-1}^{F_0}, Z_{n-1}^{\Phi})$$

La réelle différence entre F-I et F-III réside en pratique dans le fait que pour F-I, l'estimation des paramètres s'effectue sans la structure HMM de dépendance temporelle pour les séquences. Pour F-III, les HMMs sont intégrés dès cette estimation, les paramètres ainsi estimés sont donc plus en accord avec le modèle global. Techniquement, la différence est visible notamment par le fait que, dans le critère à maximiser, les probabilités *a posteriori* des états sont conditionnés, pour F-III, à l'ensemble des trames du signal observé, $p(Z_n^{\Phi} = k, Z_n^{F_0} = u|\mathbf{X})$, alors que pour F-I, la condition porte uniquement sur la trame en cours, $p(Z_n^{\Phi} = k, Z_n^{F_0} = u|\mathbf{x}_n)$.

Par ailleurs, nous avons mis au point un système qui permet d'**estimer à la fois la séquence de fréquences fondamentales prépondérante et la séquence de notes correspondante**, visant ainsi une **transcription musicale**. Ce système, publié notamment dans [Weil et al., 2009b], se base sur le système F-II pour une première estimation de fréquences prépondérantes candidates. Ensuite, en ajoutant les dépendances temporelles caractéristique de la couche E, ainsi que les liens entre la couche Z^{F_0} et E, la séquence E qui maximise la vraisemblance jointe est estimée grâce à l'algorithme de recherche en faisceaux de Vincent [2004]. Ce système est dénommé MUS-I et suit la procédure suivante :

MUS-I **Transcription "musicale" de la mélodie principale avec le modèle IMM :**

1. Estimation des paramètres :

$$\widehat{\mathbf{\Theta}}^{(\mathrm{S})\mathrm{IMM}} = \underset{\mathbf{\Theta}^{(\mathrm{S})\mathrm{IMM}}}{\arg\max}\, p(\mathbf{X}|\mathbf{\Theta}^{(\mathrm{S})\mathrm{IMM}})$$

2. Décodage de séquence Z^{F_0} et sélection de candidats :

$$\widehat{Z}^{F_0} = \underset{Z^{F_0}}{\arg\max} \prod_n p(\mathbf{x}_n|\widehat{\mathbf{\Theta}}_n^{(\mathrm{S})\mathrm{IMM}}) p(\widehat{\mathbf{\Theta}}_n^{(\mathrm{S})\mathrm{IMM}}|Z_n^{F_0}) \times p(Z_1^{F_0}) \prod_{n>1} p(Z_n^{F_0}|Z_{n-1}^{F_0})$$

3. Décodage de la séquence de notes :

$$\widehat{E}, \widehat{Z}^{F_0} = \underset{E, Z^{F_0}}{\arg\max}\, p(\mathbf{X}|\widehat{\mathbf{\Theta}}^{(\mathrm{S})\mathrm{IMM}}) p(\widehat{\mathbf{\Theta}}^{(\mathrm{S})\mathrm{IMM}}|Z^{F_0}) p(Z^{F_0}|E) p(E)$$

Enfin, le dernier système, SEP-I, permet la division du signal audio entrant en deux signaux audio, le premier correspondant à l'instrument principal dont la mélodie a été préalablement déterminée, et le second signal correspond au résiduel, c'est-à-dire l'accompagnement dans le cas des signaux audio concernés. Encore une fois, le système F-II est utilisé en prétraitement, ce qui procure la séquence de fréquences fondamentales attribuée à la source désirée. Une nouvelle estimation de paramètres est alors opérée, de sorte à améliorer les estimations faites en première passe, notamment grâce à la connaissance de la mélodie ainsi estimée. Plusieurs passes d'estimations de paramètres sont possibles, avec l'ajout graduel de différentes informations, comme l'ajout d'un modèle de non-voisement, tel que nous l'avons traité dans [Durrieu et al., 2009b]. Enfin, les signaux séparés sont obtenus d'abord par filtrage de Wiener adaptatif, grâce à des masques appliqués directement sur la TFCT du mélange, comme proposé par Benaroya et al. [2006], puis par une opération d'addition-recouvrement pour laquelle la condition de reconstruction parfaite est vérifiée par un choix adéquat de paramètres de configuration.

SEP-I **Séparation de l'instrument principal et de l'accompagnement :**

1. Première passe d'estimation de paramètres :

$$\bar{\boldsymbol{\Theta}}^{(\mathrm{S})\mathrm{IMM}} = \underset{\boldsymbol{\Theta}^{(\mathrm{S})\mathrm{IMM}}}{\arg\max}\, p(\mathbf{X}|\boldsymbol{\Theta}^{(\mathrm{S})\mathrm{IMM}})$$

2. Décodage de séquence (estimation de la mélodie) :

$$\widehat{Z}^{F_0} = \underset{Z^{F_0}}{\arg\max} \prod_n p(\mathbf{x}_n|\bar{\boldsymbol{\Theta}}_n^{(\mathrm{S})\mathrm{IMM}}) p(\bar{\boldsymbol{\Theta}}_n^{(\mathrm{S})\mathrm{IMM}}|Z_n^{F_0}) \times p(Z_1^{F_0}) \prod_{n>1} p(Z_n^{F_0}|Z_{n-1}^{F_0})$$

3. Seconde passe d'estimation des paramètres :

$\widehat{\boldsymbol{\Theta}}^{(\mathrm{S})\mathrm{IMM}} = \underset{\boldsymbol{\Theta}^{(\mathrm{S})\mathrm{IMM}}}{\arg\max}\, p(\mathbf{X}|\boldsymbol{\Theta}^{(\mathrm{S})\mathrm{IMM}})$ avec pour initialisation des amplitudes $\widetilde{\mathbf{H}}^{F_0}$

4. Calcul des TFCTs respectives, par filtrage de Wiener adaptatif (donné ci-dessous pour le SIMM) :

$$\widehat{\mathbf{V}} = \frac{\mathbf{W}^{\Gamma}\widehat{\mathbf{H}}^{\Gamma}\widehat{\mathbf{H}}^{\Phi} \bullet \mathbf{W}^{F_0}\widehat{\mathbf{H}}^{F_0}}{\mathbf{W}^{\Gamma}\widehat{\mathbf{H}}^{\Gamma}\widehat{\mathbf{H}}^{\Phi} \bullet \mathbf{W}^{F_0}\widehat{\mathbf{H}}^{F_0} + \widehat{\mathbf{W}}^{M}\widehat{\mathbf{H}}^{M}} \bullet \mathbf{X}$$

$$\text{et } \widehat{\mathbf{M}} = \frac{\widehat{\mathbf{W}}^{M}\widehat{\mathbf{H}}^{M}}{\mathbf{W}^{\Gamma}\widehat{\mathbf{H}}^{\Gamma}\widehat{\mathbf{H}}^{\Phi} \bullet \mathbf{W}^{F_0}\widehat{\mathbf{H}}^{F_0} + \widehat{\mathbf{W}}^{M}\widehat{\mathbf{H}}^{M}} \bullet \mathbf{X}$$

La seconde passe d'estimation s'effectue avec une initialisation donnée par la mélodie estimée précédemment. Ainsi, la matrice des amplitudes pour la partie source de l'instrument principal, $\mathbf{H}^{F_0}$, est initialisée par la matrice $\widetilde{\mathbf{H}}^{F_0}$ qui est telle que :

$$\widetilde{h}_{un}^{F_0} = 0 \text{ if } |12\log_2 \mathcal{F}(u) - 12\log_2 \mathcal{F}(\widehat{Z}_n^{F_0})| > \frac{1}{4}$$

Une telle initialisation garantit que les coefficients ne correspondant pas à la fréquence fondamentale estimée pour l'instrument principal sont "désactivés" et que ces coefficients, grâce aux règles de mises à jour multiplicatives, resteront à 0.

0.3.2 Méthode de gradient multiplicatif pour le (S)IMM

Dans cette section, nous rappelons le principe de gradient multiplicatif tel qu'utilisé par Lee and Seung [2001] pour la NMF. Son application dans notre cas est ensuite donnée, avec l'algorithme d'estimation au maximum de vraisemblance. Dans cette section, le jeu de paramètre $\boldsymbol{\Theta}^{(\mathrm{S})\mathrm{IMM}}$ est noté $\boldsymbol{\Theta}$.

Pour les systèmes F-II, MUS-I et SEP-I, il est nécessaire d'estimer les paramètres du modèle (S)IMM, sans les dépendances temporelles régissant les évolutions des séquences Z^{Φ} et Z^{F_0}. Par ailleurs, dans cette section, nous ne supposons pas d'autres probabilités *a priori* sur $\boldsymbol{\Theta}$, ce qui est équivalent à supposer que $p(\boldsymbol{\Theta}|Z^{\Phi}, Z^{F_0})$ est non informative (proportionnelle à 1).

Nous désirons donc finalement simplement estimer, dans un premier temps, le jeu de paramètres $\boldsymbol{\Theta}$ qui maximise la probabilité conditionnelle $p(\mathbf{X}|\boldsymbol{\Theta})$, en d'autres termes, le jeu de paramètres $\widehat{\boldsymbol{\Theta}}$ qui "explique" le mieux les observations, au sens du critère suivant :

$$C_{\mathrm{IMM}}(\boldsymbol{\Theta}) = \log p(\mathbf{X}|\boldsymbol{\Theta})$$
$$C_{\mathrm{IMM}}(\boldsymbol{\Theta}) = \sum_{f,n} \log \frac{|x_{fn}|}{\pi s_{fn}^{(\mathrm{S})\mathrm{IMM}}} - \frac{|x_{fn}|^2}{s_{fn}^{(\mathrm{S})\mathrm{IMM}}}$$

$s_{fn}^{(\mathrm{S})\mathrm{IMM}}$ est la variance paramétrée par $\boldsymbol{\Theta}$, telle que :

$$\mathbf{S}^{(\mathrm{S})\mathrm{IMM}} = \left(\mathbf{W}^{\Phi}\mathbf{H}^{\Phi}\right) \bullet \left(\mathbf{W}^{F_0}\mathbf{H}^{F_0}\right) + \mathbf{W}^{M}\mathbf{H}^{M}$$

Févotte et al. [2009b] ont montré qu'un tel formalisme était équivalent à estimer les paramètres $\boldsymbol{\Theta}$ de telle sorte que la variance paramétrée, $\mathbf{S}^{(\mathrm{S})\mathrm{IMM}}$, soit la plus proche possible du spectre de puissance $|\mathbf{X}|^2$, au sens de la divergence d'Itakura-Saito. De plus, le modèle choisi, avec les notations matricielles ci-dessus, est très proche du formalisme de la NMF, d'où le choix de l'utilisation des méthodes déjà existante pour la NMF avec divergence d'Itakura-Saito pour résoudre ce problème d'estimation. On notera qu'il est possible d'ajouter des *a priori* sur les paramètres, ce qui ajoute un terme dans le critère à maximiser. Un exemple de tel ajout d'information sur les paramètres est défini en Section 3.4.3 et développé en Section 5.2.3, permettant un certain contrôle sur l'estimation des amplitudes associées à la partie source de l'instrument principal.

La solution d'estimation la plus simple et qui a fourni des résultats acceptables est la méthode du gradient multiplicatif. En dérivant le critère C^{IMM}, paramètre par paramètre, on se rend compte que le résultat est la soustraction de deux parties positives relativement caractéristique. Si l'on note cette décomposition du gradient de la sorte, pour un paramètre θ quelconque de $\boldsymbol{\Theta}$:

$$\frac{\partial C_{\mathrm{IMM}}(\boldsymbol{\Theta}))}{\partial \theta} = \underbrace{-\left(\sum_{fn} \frac{\partial s_{fn}^{(\mathrm{S})\mathrm{IMM}}(\theta)}{\partial \theta} \frac{1}{s_{fn}^{(\mathrm{S})\mathrm{IMM}}(\theta)}\right)}_{\nabla^-} + \underbrace{\left(\sum_{fn} \frac{\partial s_{fn}^{(\mathrm{S})\mathrm{IMM}}(\theta)}{\partial \theta} \frac{|x_{fn}|^2}{s_{fn}^{(\mathrm{S})\mathrm{IMM}}(\theta)^2}\right)}_{\nabla^+}$$

où ∇^+ et ∇^- sont des quantités positives. Alors une règle de mise à jour multiplicative du paramètre $\theta^{(i)}$ est possible et permettra d'orienter $\theta^{(i+1)}$ vers une nouvelle valeur du paramètre qui réduira la divergence d'Itakura-Saito, et donc augmentera le critère désiré :

$$\theta^{(i+1)} \leftarrow \theta^{(i)} \left(\frac{\nabla^+}{\nabla^-}\right)^{\omega}$$

Dans cette équation, on notera aussi la présence de la puissance ω, qui permet de contrôler le pas d'avancement du gradient multiplicatif. Cette puissance est limitée à des valeurs réelles strictement entre 0 et 2 [Badeau et al., 2009].

En utilisant le principe du gradient multiplicatif, on peut facilement obtenir les formules de mise à jour de chacun des paramètres, pris les uns après les autres. On se rend compte qu'il est possible, comme pour la NMF usuelle, de mettre à jour les matrices entières à chaque itération de l'algorithme. Les détails sont donnés dans l'Algorithme 0.1, avec plus de détails sur les calculs en Annexe B.1.

Algorithm 0.1 Règles de mise à jour pour (S)IMM:
Estimation de $\boldsymbol{\Theta} = \{\mathbf{W}^{\Phi}, \mathbf{H}^{\Phi}, \mathbf{H}^{F_0}, \mathbf{W}^{M}, \mathbf{H}^{M}\}$
ou de $\boldsymbol{\Theta} = \{\mathbf{H}^{\Gamma}, \mathbf{H}^{\Phi}, \mathbf{H}^{F_0}, \mathbf{W}^{M}, \mathbf{H}^{M}\}$

for $i \in [1, I]$ **do**

- Paramètres de la partie source pour l'instrument principal :

$$\mathbf{H}^{F_0} \leftarrow \mathbf{H}^{F_0} \bullet \frac{(\mathbf{W}^{F_0})^T \mathbf{P}^{F_0}}{(\mathbf{W}^{F_0})^T \mathbf{Q}^{F_0}}$$

avec $\left\{ \begin{array}{rcl} \mathbf{P}^{F_0} & = & |\mathbf{X}|^2 \bullet (\mathbf{W}^{\Phi}\mathbf{H}^{\Phi})/(\mathbf{S}^{\mathrm{IMM}})^2 \\ \mathbf{Q}^{F_0} & = & (\mathbf{W}^{\Phi}\mathbf{H}^{\Phi})/\mathbf{S}^{\mathrm{IMM}} \end{array} \right.$

- Paramètres de la partie filtre pour l'instrument principal :

$$\mathbf{H}^{\Phi} \leftarrow \mathbf{H}^{\Phi} \bullet \frac{(\mathbf{W}^{\Phi})^T \mathbf{P}^{\Phi}}{(\mathbf{W}^{\Phi})^T \mathbf{Q}^{\Phi}}$$

$$\mathbf{W}^{\Phi} \leftarrow \mathbf{W}^{\Phi} \bullet \frac{\mathbf{P}^{\Phi}(\mathbf{H}^{\Phi})^T}{\mathbf{Q}^{\Phi}(\mathbf{H}_{\Phi})^T}$$

$$\mathbf{H}^{\Gamma} \leftarrow \mathbf{H}^{\Gamma} \bullet \frac{(\mathbf{W}^{\Gamma})^T \mathbf{P}^{\Phi}(\mathbf{H}^{\Phi})^T}{(\mathbf{W}^{\Gamma})^T \mathbf{Q}^{\Phi}(\mathbf{H}^{\Phi})^T}$$

avec $\left\{ \begin{array}{rcl} \mathbf{P}^{\Phi} & = & |\mathbf{X}|^2 \bullet (\mathbf{W}^{F_0}\mathbf{H}^{F_0})/(\mathbf{S}^{\mathrm{IMM}})^2 \\ \mathbf{Q}^{\Phi} & = & (\mathbf{W}^{F_0}\mathbf{H}^{F_0})/\mathbf{S}^{\mathrm{IMM}} \end{array} \right.$

- Paramètres pour l'acompagnement :

$$\mathbf{H}^{M} \leftarrow \mathbf{H}^{M} \bullet \frac{(\mathbf{W}^{M})^T (|\mathbf{X}|^2/(\mathbf{S}^{\mathrm{IMM}})^2)}{(\mathbf{W}^{M})^T (1/\mathbf{S}^{\mathrm{IMM}})}$$

$$\mathbf{W}^{M} \leftarrow \mathbf{W}^{M} \bullet \frac{(|\mathbf{X}|^2/(\mathbf{S}^{\mathrm{IMM}})^2)(\mathbf{H}^{M})^T}{(1/\mathbf{S}^{\mathrm{IMM}})(\mathbf{H}^{M})^T}$$

end for

0.3.3 Algorithme GEM pour le (S)GSMM

Pour le (S)GSMM, l'algorithme d'estimation est plus compliqué que celui utilisé pour le (S)IMM. En effet, la présence intrinsèque d'états cachés dans le modèle rend nécessaire une estimation du type espérance-maximisation (EM). Nous définissons d'abord le critère à maximiser, avant de donner l'algorithme d'estimation des paramètres. Dans cette section, le jeu de paramètres $\boldsymbol{\Theta}^{(\mathrm{S})\mathrm{GSMM}}$ est noté $\boldsymbol{\Theta}$.

Comme pour le modèle (S)IMM, l'estimation des paramètres dans le système F-I est effectuée en supposant que les dépendances temporelles sont ignorées. Encore une fois, aucun *a priori* sur les paramètres n'est donné. Par conséquent, par simplicité, nous considérons pour cette section le critère de maximum de vraisemblance, défini comme suit :

$$C_{\text{GSMM}}(\mathbf{\Theta}, \mathbf{\Theta}^{(i-1)}) = E\left[\log p(\mathbf{X}, Z^{F_0}, Z^{\Phi}; \mathbf{\Theta}) | \mathbf{X}; \mathbf{\Theta}^{(i-1)}\right]$$

Le principe de l'algorithme EM est alors de trouver, à chaque itération i, étant donné une estimation $\mathbf{\Theta}^{(i-1)}$ des paramètres, une nouvelle estimation des paramètres $\mathbf{\Theta}^{(i)}$ telle que le critère ci-dessus soit maximisé :

$$\mathbf{\Theta}^{(i)} | C_{\text{GSMM}}(\mathbf{\Theta}^{(i)}, \mathbf{\Theta}^{(i-1)}) \geq C_{\text{GSMM}}(\mathbf{\Theta}, \mathbf{\Theta}^{(i-1)}) \, \forall \mathbf{\Theta}$$

Cependant, comme pour le modèle (S)IMM, chercher les zéros de la dérivée du critère n'admet pas de solution analytique simple, et il faut approcher ce résultat par un algorithme de type EM généralisé (GEM), où le nouveau jeu de paramètres est tel que le critère ne décroisse pas :

$$\mathbf{\Theta}^{(i)} | C_{\text{GSMM}}(\mathbf{\Theta}^{(i)}, \mathbf{\Theta}^{(i-1)}) \geq C_{\text{GSMM}}(\mathbf{\Theta}^{(i-1)}, \mathbf{\Theta}^{(i-1)})$$

Afin de résoudre ce problème, il est utile de réécrire le critère de sorte à ce que les étapes E (espérance) et M (maximisation) de l'algorithme GEM soient réalisables le plus simplement possible. Ainsi, en utilisant notamment l'astuce habituelle pour le critère EM avec les mélanges de Gaussiennes, on obtient le critère suivant :

$$C_{\text{GSMM}}(\mathbf{\Theta}, \mathbf{\Theta}^{(i-1)}) = \sum_{n,k,u} \left[\sum_{f} \left(\log \frac{|x_{fn}|}{\pi s_{fn}^{(\text{S})\text{GSMM}|ku}} - \frac{|x_{fn}|^2}{s_{fn}^{(\text{S})\text{GSMM}|ku}} \right) + \log \pi_{ku} \right] \times p(k, u | \mathbf{x}_n; \mathbf{\Theta}^{(i-1)}) - \lambda \left(\sum_{k,u} \pi_{ku} - 1 \right) + \text{CST}$$

Dans le critère ci-dessus, on a noté $s_{fn}^{(\text{S})\text{GSMM}|ku}$ la variance pour l'état $(Z^{\Phi}, Z^{F_0}) = (k, u)$, au bin fréquentiel f et à la trame n, avec $\mathbf{s}_n^{(\text{S})\text{GSMM}|ku}$ égale au vecteur $b_{kun}\mathbf{w}_k^{\Phi} \bullet \mathbf{w}_u^{F_0} + \mathbf{W}^M \mathbf{h}_n^M$. Le terme "CST" est un terme indépendant des paramètres à estimer.

Avec un tel critère, l'algorithme GEM se résume à deux étapes relativement simples : la première étape, l'étape E, consiste à calculer les probabilités *a posteriori* des états cachés (k, u), $p(k, u | \mathbf{x}_n; \mathbf{\Theta}^{(i-1)})$. La seconde étape, l'étape M, consiste à mettre à jour les paramètres désirés de sorte à faire croître le critère C_{GSMM}. Pour cette étape, les dérivées du critère aboutissent à des expressions très similaires à celles que nous avons déjà données pour l'estimation des paramètres du modèle IMM. Nous pouvons donc utiliser une méthode de gradient multiplicatif au sein de l'étape M, avec des formules obtenues de manière analogue à celles de l'algorithme d'estimation des paramètres de l'IMM. Ces formules et la procédure d'estimation GEM pour le (S)GSMM sont données dans l'Algorithme 0.2.

0.3.4 Décodage de séquences

0.3.4.1 Algorithme de Viterbi

L'algorithme de Viterbi est un algorithme permettant de trouver la séquence d'états cachés la plus probable ayant généré une séquence d'observation. La structure de dépendances des

Algorithm 0.2 Algorithme GEM pour l'estimation des paramètres de (S)GSMM : estimation de $\mathbf{\Theta}$, égal à $\mathbf{\Theta}^{\text{GSMM}} = \{\mathbf{B}, \mathbf{W}^{\Phi}, \mathbf{H}^{M}, \mathbf{W}^{M}\}$ ou $\mathbf{\Theta}^{\text{SGSMM}} = \{\mathbf{B}, \mathbf{H}^{\Gamma}, \mathbf{H}^{M}, \mathbf{W}^{M}\}$

for $i \in [1, I]$ **do**

- $\forall k, u, n,\ \ b_{kun} \leftarrow b_{kun} \dfrac{p^B_{kun}}{q^B_{kun}}$, où $\left\{ \begin{array}{rcl} p^B_{kun} & = & \displaystyle\sum_f \dfrac{w^{\Phi}_{fk} w^{F_0}_{fu} |x_{fn}|^2}{(s^{(\text{S})\text{GSMM}|ku}_{fn})^2} \\ q^B_{kun} & = & \displaystyle\sum_f \dfrac{w^{\Phi}_{fk} w^{F_0}_{fu}}{s^{(\text{S})\text{GSMM}|ku}_{fn}} \end{array} \right.$

Étape E : calcul de $\gamma_n^{(i-1)}(k,u) = p(k,u|\mathbf{x}_n; \mathbf{\Theta}^{(i-1)})$ avec l'Algorithme B.1.
Étape M : mise à jour des paramètres :

- (GSMM) $\forall f, k,\ \ w^{\Phi}_{fk} \leftarrow w^{\Phi}_{fk} \dfrac{p^{\Phi}_{fk}}{q^{\Phi}_{fk}}$, où $\left\{ \begin{array}{rcl} p^{\Phi}_{fk} & = & \displaystyle\sum_{u,n} \gamma_n^{(i-1)}(k,u) \times \dfrac{b_{kun} w^{F_0}_{fu} |x_{fn}|^2}{(s^{(\text{S})\text{GSMM}|ku}_{fn})^2} \\ q^{\Phi}_{fk} & = & \displaystyle\sum_{u,n} \gamma_n^{(i-1)}(k,u) \dfrac{b_{kun} w^{F_0}_{fu}}{s^{(\text{S})\text{GSMM}|ku}_{fn}} \end{array} \right.$
- (SGSMM) $\forall p, k,\ \ h^{\Gamma}_{pk} \leftarrow h^{\Gamma}_{pk} \dfrac{p^{\Gamma}_{pk}}{q^{\Gamma}_{pk}}$, où $\left\{ \begin{array}{rcl} p^{\Gamma}_{pk} & = & \displaystyle\sum_{f,u,n} \gamma_n^{(i-1)}(k,u) \times \dfrac{b_{kun} w^{\Gamma}_{fp} w^{F_0}_{fu} |x_{fn}|^2}{(s^{\text{SGSMM}|ku}_{fn})^2} \\ q^{\Gamma}_{pk} & = & \displaystyle\sum_{f,u,n} \gamma_n^{(i-1)}(k,u) \dfrac{b_{kun} w^{\Gamma}_{fp} w^{F_0}_{fu}}{s^{\text{SGSMM}|ku}_{fn}} \end{array} \right.$
- $\forall r, n,\ \ h^{M}_{rn} \leftarrow h^{M}_{rn} \dfrac{p^{H}_{rn}}{q^{H}_{rn}}$, où $\left\{ \begin{array}{rcl} p^{H}_{rn} & = & \displaystyle\sum_{k,u,f} \gamma_n^{(i-1)}(k,u) \dfrac{w^{M}_{fr} |x_{fn}|^2}{(s^{(\text{S})\text{GSMM}|ku}_{fn})^2} \\ q^{H}_{rn} & = & \displaystyle\sum_{k,u,f} \gamma_n^{(i-1)}(k,u) \dfrac{w^{M}_{fr}}{s^{(\text{S})\text{GSMM}|ku}_{fn}} \end{array} \right.$
- $\forall f, r,\ \ w^{M}_{fr} \leftarrow w^{M}_{fr} \dfrac{p^{W}_{fr}}{q^{W}_{fr}}$, où $\left\{ \begin{array}{rcl} p^{W}_{fr} & = & \displaystyle\sum_{k,u,n} \gamma_n^{(i-1)}(k,u) \dfrac{h^{M}_{rn} |x_{fn}|^2}{(s^{(\text{S})\text{GSMM}|ku}_{fn})^2} \\ q^{W}_{fr} & = & \displaystyle\sum_{k,u,n} \gamma_n^{(i-1)}(k,u) \dfrac{h^{M}_{rn}}{s^{(\text{S})\text{GSMM}|k,u}_{fn}} \end{array} \right.$

end for

états cachés doit suivre un HMM, comme les couches Z^{Φ} et Z^{F_0} de nos modèles de signaux. L'algorithme est d'abord donné pour un HMM générique, puis les différents aménagements nécessaires pour chaque application (et chaque système) sont précisés.

Nous disposons de la séquence d'observation $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ et nous désirons connaître la séquence (cachée) $Z = \{Z_1, \ldots, Z_N\}$:

$$\widehat{Z} = \arg\max_{Z} p(\mathbf{X}, Z)$$

La séquence Z suit un HMM, avec pour probabilités de transition $p(Z_n = v | Z_{n-1} = u) = q(u, v)$. On suppose par simplicité que la variable Z_n est à valeur dans $[1, U]$. La probabilité jointe s'écrit alors :

$$\begin{aligned} p(\mathbf{X}, Z) &= p(\mathbf{X}|Z)p(Z) \\ &= \prod_n p(\mathbf{x}_n|Z_n) \times p(Z_1) \prod_{n>1} p(Z_n|Z_{n-1}) \\ &= p(Z_1)p(\mathbf{x}_1|Z_1) \prod_{n>1} p(\mathbf{x}_n|Z_n)p(Z_n|Z_{n-1}) \end{aligned}$$

Plutôt que de calculer cette probabilité pour toutes les U^N séquences possibles, ce qui est en pratique impossible, l'algorithme proposé par Viterbi [1967] réduit le nombre de chemins à parcourir pour trouver le chemin optimal. En effet, introduisons la quantité suivante :

$$\delta_{un} = \max_{Z_{1:n-1}} p(\mathbf{x}_{1:n}, Z_{1:n-1}, Z_n = u)$$

On se rend facilement compte que, d'une part, cette quantité est liée au maximum de vraisemblance désiré :

$$\max_u \delta_{uN} = \max_{Z_N = u} \max_{Z_{1:N-1}} p(\mathbf{x}_{1:N}, Z_{1:N-1}, Z_N = u) = \max_Z p(\mathbf{X}, Z)$$

Si l'on peut calculer les δ_{uN}, $\forall u$, alors on peut déterminer quel est le dernier état de la séquence cachée désirée. En fait, une seconde constatation sur cette quantité δ_{un} est que l'on peut déterminer, pour une trame donné n, toutes les valeurs $\delta_{u(n+1)}$, $\forall u$, et ce uniquement grâce à δ_{vn}, $\forall v$, les probabilités de transition $q(v, u)$ et les probabilités conditionnelles $p(\mathbf{x}_{n+1}|Z_{n+1} = u)$, $\forall u$. En effet, on a la relation suivante :

$$\delta_{u(n+1)} = \left(\max_v \delta_{vn} q(v, u) \right) p(\mathbf{x}_{n+1}|Z_{n+1} = u)$$

L'algorithme de Viterbi consiste donc à calculer itérativement les quantités δ_{un}, $\forall u$, pour des valeurs de n croissantes. L'algorithme correspondant est décrit en détail dans l'Algorithme 0.3. On a essentiellement besoin de conserver deux quantités par trame et par état : la probabilité du chemin aboutissant à l'état u à la trame n, δ_{un}, mais aussi l'état à la trame $n-1$ correspondant à ce chemin, ψ_{un}.

Enfin, pour pouvoir appliquer l'algorithme de Viterbi, dans les différents systèmes où nous l'avons utilisé, il convient de considérer les notes suivantes :

F-I La séquence désirée est $Z = (Z^{F_0}, Z^{\Phi})$.

F-II Pour décoder la séquence Z^{F_0}, la probabilité conditionnelle est supposée être proportionnelle au coefficient d'amplitude de $\mathbf{H}^{F_0}$.

F-III Bien que la contrainte temporelle de la structure HMM ait été implémentée directement dans l'estimation de paramètres, il reste nécessaire d'utiliser cet algorithme pour déterminer le chemin le plus probable.

Algorithm 0.3 Algorithme de Viterbi

Initialisation :
for $u \in [1, U]$ **do**

$$\delta_{u1} = p(Z_1 = u)p(\mathbf{x}_1|Z_1 = u)$$
$$\psi_{u1} = 0$$

end for
Itération :
for n de 2 à N **do**
  for $u \in [1, U]$ **do**

$$\delta_{un} = \left(\max_v \delta_{v(n-1)} q(v, u)\right) p(\mathbf{x}_n|Z_n = u)$$
$$\psi_{un} = \arg\max_v \delta_{v(n-1)} q(v, u)$$

  end for
end for
Terminaison :

$$\widehat{Z}_N = \arg\max_u \delta_{uN}$$

Rétro-propagation :
for n de N à 2 **do**

$$\widehat{Z}_{n-1} = \psi_{\widehat{Z}_n n}$$

end for

0.3.4.2 Algorithme de recherche par faisceaux

Afin de modéliser la séquence de notes E jouée par l'instrument principal, nous avons utilisé des distributions dépendant des durées des notes. Malheureusement, cela entraîne l'impossibilité d'utiliser un algorithme de type Viterbi. En effet, on pourrait transformer ce formalisme en un HMM, mais en pratique, cela entraînerait un grand nombre d'états cachés à gérer, ce qui ne rendrait donc pas l'algorithme réellement réalisable. Au lieu d'effectuer une recherche exhaustive parmi tous les chemins E possibles, nous suivons l'algorithme de recherche par faisceaux, adoptée par Vincent [2004]. Nous l'avons quelque peu remanié de sorte à ce qu'il prenne en compte les spécificités de notre modèle.

Le principe de cet algorithme reste similaire à celui utilisé pour l'algorithme de Viterbi. En effet, on peut mettre en évidence le même type de lien entre les probabilités calculées à la trame n et à la trame $n+1$. De fait, si l'on se fixe un certain nombre de chemins, qu'on les fait évoluer de la première trame à la dernière, en ajoutant, gardant ou soustrayant les notes et en testant les combinaisons les plus probables, alors il est possible d'approcher la solution désirée, tout en ayant testé un nombre de chemins raisonnable.

La procédure d'estimation est donnée dans l'Algorithme 0.4. Dans le système MUS-I, la séquence de notes E est estimée après avoir estimé les paramètres $\widehat{\mathbf{\Theta}}$ et la séquence de fréquences fondamentales optimale (au sens de système F-II) $\widehat{Z}^{F_0}$, sans prendre en compte la couche des notes. Cette séquence estimée permet de définir des notes candidates pour chaque trame, E_n^C.

Algorithm 0.4 Estimation de la séquence de notes E pour le système MUS-I

Initialisation des chemins
for toutes les notes candidates E_1^C **do**
  Trouver le meilleur $\widehat{Z}_1^{F_0}$.
  Calculer $p(\mathbf{x}_1, \widehat{\mathbf{\Theta}}_1, \widehat{Z}_1^{F_0}, \widehat{E}_1)$.
end for
Trier et éliminer les chemins les moins probables.
Extension des chemins
for $n \in [2, N]$ **do**
  for tous les chemins conservés : **do**
    Rallonger le chemin avec l'une de ces opérations : continuation, suppression et remplacement. Le remplacement et les nouvelles notes doivent être prises de l'ensemble des notes candidates E_n^C.
    for chaque chemin étendu : **do**
      if $E_n \neq 0$ **then**
        Trouver le meilleur $\widehat{Z}_n^{F_0}$.
      end if
      Calculer $p(\mathbf{x}_{1:n}, \widehat{\mathbf{\Theta}}_{1:n}, \widehat{Z}_{1:n}^{F_0}, \widehat{E}_{1:n})$, avec $p(\mathbf{x}_{1:n-1}, \widehat{\mathbf{\Theta}}_{1:n-1}, \widehat{Z}_{1:n-1}^{F_0}, \widehat{E}_{1:n-1})$.
    end for
  end for
  Trier et éliminer les chemins les moins probables..
end for
Terminaison
$\widehat{E}$ est sélectionné comme étant celui qui maximise $p(\mathbf{X}, \widehat{\mathbf{\Theta}}, \widehat{Z}^{F_0}, E)$.

0.4 Applications : Extraction de la mélodie principale

Les systèmes F-I et F-II ont été évalués dans le cadre de campagnes d'évaluation internationale, pour la tâche d'extraction de la mélodie principale. Les résultats obtenus lors du MIREX Audio Melody Extraction (AME) de 2008 sont donnés dans la Table 6.2 et ceux de MIREX AME 2009 sont donnés dans la Table 6.4.

Les résultats de nos algorithmes pour MIREX 2008 et 2009 montrent que nos systèmes ont de bonnes performances pour la plupart des bases de données utilisées, sauf pour la base de données MIR-1K, pour lequel le mélange à -5dB ne permet plus à nos algorithmes de distinguer la mélodie *principale*. Par ailleurs, pour la base de données MIR-1K, une classification voix chantée / son instrumental serait plus appropriée à un certain stade de décision, étant donné la nature des morceaux (Karaoke) et le contenu de ceux-ci, où parfois la voix chantée est doublée à l'octave par d'autres sons. Cela peut perturber encore plus un système qui ne détecterait pas les différences de production entre le chant et les autres instruments.

D'une manière générale, on constate que les résultats obtenus par nos algorithmes sont quand même meilleurs sur les sous-ensembles avec voix chantée que sur les sous-ensembles purement instrumentaux (voire synthétique). En effet, pour les morceaux instrumentaux, le modèle de l'accompagnement, équivalent à de la NMF, permet de prendre en compte la majorité du contenu présent dans le mélange, incluant le potentiel instrument principal.

Le système F-III n'a pas été formellement comparé aux deux premiers. Quelques résultats préliminaires tendent à montrer que l'apport de la structure HMM lors de l'estimation des paramètres n'est pas aussi bénéfique que l'on aurait pu l'espérer. En effet, durant l'algorithme GEM correspondant, il faut calculer des probabilités *a posteriori* qui dépendent de tout le signal. Ces probabilités reflètent donc, en un sens, combien le modèle source/filtre avec structure HMM est proche des données observées. Dans le cadre de l'algorithme GEM pour le (S)GSMM, sans la structure HMM, la probabilité d'un état (k, u) *a posteriori* à la trame n n'est conditionnée que par l'observation à cette même trame n. Dans une certaine mesure, cette probabilité ne donne une mesure d'adéquation du modèle source/filtre que pour la trame donnée. De ce fait, si le modèle source/filtre choisi n'est pas complètement adapté aux observations, les erreurs de modélisation sont prises en compte plus globalement avec le HMM qu'avec le GSMM, ce qui peut expliquer les résultats moins bons obtenus par le HMM.

Malgré des performances en deça de nos attentes, le système F-III et le modèle d'estimation incluant le HMM dans le GSMM sont des pistes intéressantes à approfondir et à analyser. En effet, une telle approche permet de réduire les approximations faites lors de la phase d'estimation. La dégradation des performances que nous observons vient probablement d'un problème de modélisation, et non seulement d'un poblème d'implémentation ou d'algorithme.

Enfin, **le système MUS-I** a été évalué dans notre article [Weil et al., 2009b] et au cours d'une campagne d'évaluation interne au projet Quaero. Pour les premières évaluations, des signaux synthétiques ont été générés. Sur ces signaux, notre algorithme obtient de bons résultats, avec des valeurs de précision et de rappel entre 60% et 70%.

Pour la deuxième évaluation, des signaux plus réalistes ont été utilisés. Les résultats ainsi obtenus sont moins bons que ce à quoi l'on pouvait s'attendre, ne dépassant pas les 15% de précision et de rappel. Il y a plusieurs raisons possibles à cette baisse de performance. Tout d'abord, il est intéressant de noter que sur cette même base de données, F-I et F-II éprouvent aussi une certaine baisse de performances. Les morceaux traités

sont plus longs, et contiennent plus de passage sans chanteur : dans ces passages, nos algorithmes transcrivent des notes correspondant à l'instrument principal en cours, qu'il s'agisse d'un solo de guitare ou d'un riff quelconque. Cela entraîne une baisse inévitable dans les résultats. Un tel problème ne peut être résolu qu'en ajoutant une détection du type d'instrument principal à nos systèmes. Par ailleurs, un potentiel problème de justesse a été identifié : si le chanteur chante trop en dessous des fréquences standard (par exemple A4 = 440Hz), alors les notes estimées risquent de ne pas correspondre aux annotations, bien qu'il est probable que l'erreur ne soit en fait qu'une simple translation de la ligne mélodique d'un demi-ton.

Par ailleurs, sur les extraits chantés, on constate que la variation de hauteur (en Hz) de la ligne mélodique ne varie pas toujours en directe relation avec la ligne des notes, comme lors d'un vibrato. Cela accentue la difficulté du problème, qui pourrait être partiellement traité si d'avantage de connaissance musicale est intégrée dans le système, comme par exemple les notions de tonalité ou d'accord.

0.5 Applications : Séparation de l'instrument principal

Le système de séparation SEP-I a donné lieu à deux articles de conférence : Durrieu et al. [2009a] et Durrieu et al. [2009b], ainsi qu'à deux participations à des évaluations : SiSEC 2008 et Quaero 2009.

Dans [Durrieu et al., 2009a], SEP-I a été testé sur des morceaux mono-canaux, avec l'algorithme expliqué précédemment. Les résultats montrent d'abord que, connaissant la mélodie, SEP-I est capable d'approcher une séparation idéale en terme de SDR. Avec estimation automatique de la mélodie, les résultats sont plus mitigés, mais restent comparables (favorablement) aux autres travaux sur l'amélioration ou l'atténuation de la voix principale.

Dans [Durrieu et al., 2009b], nous avons proposé une extension stéréo aux algorithmes mono présents dans [Durrieu et al., 2009a]. Par ailleurs, les mécanismes de filtres lisses et d'estimation de la partie non-voisée de la partie principale sont explicités. En résumé, l'extension à la stéréo revient à estimer les paramètres conjointement sur les deux canaux, tout en supposant que les signaux correspondant sont indépendants statistiquement l'un de l'autre. Le lissage des filtres est imposé structurellement, par la décomposition de ceux-ci sur une famille de fonctions lisses, alors que l'intégration de la partie non-voisée correspond à l'ajout d'un élément "bruit" dans la matrice $\mathbf{W}^{F_0}$. Les résultats obtenus montrent un gain certain entre les résultats de l'algorithme mono et ceux de l'algorithme stéréo. Par ailleurs, même si l'ajout du lissage des filtres n'aboutit pas à une amélioration des résultats, l'ajout du non-voisé, lui, permet dans certains cas d'améliorer les résultats, en termes de mesures objectives, mais aussi de manière assez nette à l'écoute des fichiers séparés. Le système SEP-I stéréo a participé à l'évaluation SiSEC 2008, où nos résultats ont atteint un second rang sur la moyenne des résultats. Il est à noter que les algorithmes ayant obtenus les meilleurs résultats sont ceux qui détectaient la mélodie préalablement à la séparation en elle-même.

Enfin, l'évaluation Quaero 2009 consistait à analyser des morceaux longs, avec découpe possible en petits morceaux. Malheureusement, pour des raisons techniques, nous n'avons pas pu évaluer les performances sur l'ensemble de test. Cela étant, sur la base de développement, nous avons obtenus de bons résultats, avec des gain en terme de SDR et SIR important pour la partie voix principale, et un peu moins grand pour la partie accompagnement.

0.6 Conclusions et perspectives

Nous avons proposé deux modèles génériques pour les signaux de mélange de type "voix chantée + accompagnement". L'adéquation de ces modèles avec les signaux réels est étudié dans le cadre de deux applications : la transcription de la mélodie en séquence de fréquences fondamentales et la séparation de l'instrument jouant cette mélodie du reste des instruments du mélange.

Le premier modèle proposé pour l'instrument principal est un modèle de mélange de Gaussiennes amplifiées (MMGA, GSMM en Anglais). Les états cachés de ce modèle exprime explicitement la dépendance du signal d'intérêt à la fréquence fondamentale, ce qui permet d'établir un lien direct entre l'estimation de la séquence optimale d'états cachés ayant généré les observations et la séquence de fréquences fondamentales mélodiques sous-jacentes. Tout cela est rendu possible par l'adaptation du modèle de production source/filtre au cadre statistique de séparation de sources de Benaroya et al. [2006]. Le second modèle est, sous certains aspects, une généralisation du modèle GSMM précédent. En effet, le signal est alors supposé être la combinaison de tous les états cachés du GSMM, d'où la dénomination de modèle à mélange instantané (IMM).

Pour chaque modèle, l'accompagnment est modélisé par une somme de signaux gaussiens, indépendants. Il s'avère que l'estimation des paramètres qui encodent ces signaux est équivalente à un problème de factorisation en matrices positives (NMF) : le spectre de puissance est décomposé sur une base de modèles spectraux, avec pour mesure d'erreur de reconstruction la divergence d'Itakura-Saito. Le parallèle entre le cadre adopté dans cette thèse et les méthodologies propres à la littérature sur la NMF est essentiel pour l'élaboration des algorithmes d'estimation que nous proposons.

Pour les deux modèles, GSMM et IMM, les contraintes temporelles d'évolution des états sont incluses par deux couches d'états : une couche dite "physique" et une autre dite "musicologique". La couche physique prend essentiellement en charge la séquence de fréquences fondamentales, en la contraignant à être relativement lisse. La seconde couche contraint les notes de la mélodie à avoir des durées réalistes d'un point de vue musical. Par ailleurs, des améliorations de ces modèles d'origine sont proposés. Tout d'abord, les filtres de la partie principale sont contraints par construction à être régulier, en les décomposant sur une base de fonctions lisses. De plus, en incorporant un élément "non-voisé" dans la base de spectres de la partie source de l'instrument principal, une séparation de ce dernier et de l'accompagnement plus complète est possible.

Nous proposons cinq systèmes, F-I, F-II, F-III, MUS-I et SEP-I. Les trois premiers systèmes estiment la séquence de fréquences fondamentales, fournissant une fréquence F0 par trame. MUS-I permet d'obtenir la séquence de notes correspondant à la mélodie et SEP-I sépare le signal d'entrée en deux signaux audio, l'estimée de l'instrument principal et celle de l'accompagnement.

Les différentes expériences montrent que F-I et F-II fournissent de bons résultats. Pour cette raison, F-II, qui est par ailleurs plus simple et rapide à implémenter que F-I, sert de pré-traitement aux systèmes MUS-I et SEP-I. F-I et F-II ont été évalués au sein de campagnes d'évaluation internationales, MIREX 2008 et 2009, ce qui a montré qu'ils étaient au niveau de l'état de l'art. F-III est basé sur une estimation du modèle GSMM avec les contraintes temporelles markoviennes. Des évaluations plus précises et systématiques doivent encore être menées sur F-III, malgré les faibles scores obtenus par les tests préliminaires. Comme on pourrait s'attendre à ce que F-III ait de meilleures performances que F-I, ces nouveaux tests devraient permettre de mieux analyser et identifier les erreurs de F-III et

les corriger, si possible.

MUS-I obtient des résultats encourageant sur des fichiers synthétiques, mais semblent plus faibles pour des signaux réels. Plusieurs raisons sont possibles à cela. l'adéquation entre la séquence de F0 et la séquence de note ne semble pas aussi évidente que cela n'en a l'air, loin d'être déterministe. Ainsi un modèle plus complexe, éventuellement avec plus de données musicales comme la tonalité, le tempo, devrait permettre de surmonter les difficultés rencontrées actuellement par MUS-I.

Enfin, les résultats de notre système de séparation SEP-I, dans sa version stéréo, ont obtenus parmi les meilleurs résultats à l'évaluation internationale SiSEC 2008. Les résultats de cette campagne d'évaluation montrent que le principe d'estimation de la mélodie préalable à la séparation elle-même, permet d'obtenir des résultats qui ont dépassé l'état de l'art d'alors. Cela valide donc notre méthode sur cet aspect.

Des améliorations multiples des modèles sont possibles. Nous avons toujours à l'esprit l'envie d'élaboration, *in fine*, un système qui n'aurait pas besoin d'approximations durant la phase d'estimation. L'estimation conjointe de toutes les quantités et paramètres définis pour chaque modèle est un but qui devrait permettre, si ce n'est d'obtenir de meilleurs résultats, au moins de valider le modèle dans son ensemble. D'autres types de contraintes pourront alors être ajoutées, pour encore mieux correspondre au signal : des contraintes d'ordre musical, avec l'ajout de la tonalité, du tempo, des rythmes, pour orienter la valeur des notes, la justesse et les onsets/offsets vers des valeurs en accord avec tout l'environnement (l'accompagnement). Le développement d'un modèle d'accompagnement paraît aussi envisageable et souhaitable, tant la simplicité de celui-ci peut, à terme, nuire à l'estimation des paramètres en général, car sans doute trop redondant avec le modèle source/filtre de la partie principale. La base de la partie source de l'instrument principal pourrait inclure des éléments correspondant à des signaux dont la fréquence fondamentale n'est pas constante, par exemple avec des variations linéaires de cette fréquence. Enfin la modélisation du silence de l'instrument principal, liée au problème de discrimination entre celui-ci et les instruments de l'accompagnement, devrait faire l'objet d'une étude plus approfondie.

Notations

Acronyms

AME	Audio Melody Extraction
DFT	Discrete Fourier Transform
DTFT	Discrete Time Fourier Transform
EM	Expectation-Maximization
F-I, F-II, F-III	The three systems proposed for frame-wise estimation of the melody
FT	Fourier Transform
F0, f_0	Fundamental frequency
GEM	Generalized EM
GMM	Gaussian Mixture Model
GSMM	Gaussian Scaled Mixture Model
HM-GSMM	Hidden Markov-GSMM, HMM for the lead instrument
HMM	Hidden Markov Model
IF	Instantaneous Frequency (spectrogram)
IMM	Instantaneous Mixture Model
ISMIR	International Society for Music Information Retrieval
MIREX	Music Information Retrieval Evaluation eXchange
MUS-I	System proposed to estimate the sequence of notes of the melody
NMF	Nonnegative Matrix Factorization
OLA	OverLap-Add procedure
PDF	Probability Density Function
PSD	Power Spectrum Density
QbH	Query-by-Humming
SEP-I	Lead instrument separation system.
SGSMM	Smooth filters-GSMM
(S)GSMM	used when it applies to either GSMM or SGSMM
SiSEC	Signal Separation Evaluation Campaign
SIMM	Smooth filters-IMM
(S)IMM	used when it applies to either IMM or SIMM
V-IMM	"Voiced-IMM", variation of SEP-I, only includes voiced parts in lead instrument
VU-IMM	"Voiced and Unvoiced - IMM", variation of SEP-I, includes both voiced and unvoiced parts
w.r.t.	with respect to
w.s.s.	Wide sense stationary

Parameters and variables

$\gamma_n^{(i-1)}(k,u)$	$\gamma_n^{(i-1)}(k,u) = p(k,u\|\mathbf{x}_n;\boldsymbol{\Theta}^{(i-1)})$, for the (S)GSMM
$\boldsymbol{\nu}_n^{ku}$	Gaussian component for source u and filter k, see Equation (3.36)
$\boldsymbol{\Theta}$	Parameter set, for the (S)GSMM or the (S)IMM model, depending on the context
$\boldsymbol{\Theta}^{\text{(S)GSMM}}$	Parameter set, for the (S)GSMM. See Table 3.1
$\boldsymbol{\Theta}^{\text{(S)IMM}}$	Parameter set, for the (S)IMM. See Table 3.2
σ^{MIDI}	Parameter shaping the note to frequency constraint in Equation (3.29)
$\mathbf{B}$	Amplitude tensor for the lead instrument GSMM
$E = E_{1:N}$	Sequence of note states for the leading instrument
F	Number of frequency bins for the STFTs
$\mathbf{H}^{\Gamma}$	Coefficients for the decomposition of $\mathbf{W}^{\Phi}$ on $\mathbf{W}^{\Gamma}$
$\mathbf{H}^{\Phi}$	Amplitude coefficients for the lead instrument filter part
$\mathbf{H}^{F_0}$	Amplitude coefficients for the lead instrument source part
$\mathbf{H}^{M}$	Amplitudes for the accompaniment components
K	Number of basis elements (columns) in matrix $\mathbf{W}^{\Phi}$
$\mathbf{M}$	The (complex) STFT matrix for the accompaniment M
N	Number of frames of the STFTs
$[n_{\min}^{\text{MIDI}}, n_{\max}^{\text{MIDI}}]$	Minimum and maximum values for the candidate MIDI notes for E
R	Number of components for the accompaniment M
U	Number of elements in the dictionary matrix $\mathbf{W}^{F_0}$
U_{st}	Number of elements per semitone in the dictionary matrix $\mathbf{W}^{F_0}$
$\mathbf{V}$	The (complex) STFT matrix of the lead instrument V
$\mathbf{W}^{\Gamma}$	Dictionary of smooth elementary filters
$\mathbf{W}^{\Phi}$	Matrix of spectral shapes for the lead intrument filter part
$\mathbf{W}^{F_0}$	Dictionary of spectral combs for the lead instrument source part
$\mathbf{W}^{M}$	Matrix of spectral shapes for the components of the accompaniment M
$\mathbf{X}$	The (complex) STFT matrix of the observed mixture X
$Z^{\Phi} = Z_{1:N}^{\Phi}$	Sequence of states for the leading instrument filter part
$Z^{F_0} = Z_{1:N}^{F_0}$	Sequence of states for the leading instrument source part

Functions

d_{EUC}	Scalar Euclidean (EUC) distance, appears in Section 4.3
d_{IS}	Scalar Itakura-Saito (IS) divergence, defined Section 4.2
d_{KL}	Scalar Kullback-Leibler (KL) divergence, appears in Section 4.3
D_{IS}	IS divergence between two matrices, Section 4.2, Definition 9
$\mathcal{F}$	Mapping function from a source element number to a fundamental frequency, defined in Equation (3.10)
$\mathcal{F}^{\mathrm{MIDI}}$	Mapping function from a MIDI code number to a fundamental frequency, defined in Equation (3.27)
$\mathcal{G}(\alpha, \beta)$	Gamma distribution, with shape parameter α and scale parameter β, defined in Section A.2
$\Gamma(\alpha)$	Gamma function, defined in Section A.2
$\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$	Real Gaussian distribution, with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$
$\mathbf{N}(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$	Evaluation of the real Gaussian distribution at vector $\mathbf{x}$, with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$
$\mathcal{N}_c(\boldsymbol{\mu}, \boldsymbol{\Sigma})$	Complex Gaussian distribution, with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, see Section A.1.1
$\mathbf{N}_c(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$	Evaluation of the complex Gaussian distribution at vector $\mathbf{x}$, with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$

Operators and matrix notations

$\bullet$	Element-wise or Hadamard product between matrices
FT[.]	Fourier transform operator
$\mathbb{1}_A$	Indicator function for an event A : $\mathbb{1}_A = \begin{cases} 1 & \text{if } A \text{ is true ;} \\ 0 & \text{otherwise.} \end{cases}$
$\mathbf{A}$, $\mathbf{a}_j$, a_{ij}	Matrix $\mathbf{A}$, the vector at its column j and the element at its row i and column j
$\mathbf{A}^T$	Transpose of matrix $\mathbf{A}$
$[\mathbf{A}]_{ij}$	Applied to a matrix $\mathbf{A}$, represents the element at the i^{th} row, and j^{th} column of $\mathbf{A}$, *i.e.* a_{ij}
$A_{n:m}$	For a sequence A: $A_{n:m} = \{A_i\}_{i=n\ldots m}$

Chapter 1

Introduction

Imagine you want a song. You can remember the drums, the bass, the riffs, but not the lyrics, the name of the artist or the title. Well, you cannot *google* it. Luckily, the musical knowledge of your colleague next room is only second to his addiction for coffee. Now how could he help? Would he recognize your tune if you mimic that drum solo *"poom tchak poom tchak"*, or that guitar riff that goes *"geeding geeding"*? In most cases, you'll instead find yourself singing the main melody, which you may have identified as the most characteristic part of the song, with the least ambiguity. Apart from a few exceptions[1], the melody will indeed be the feature of choice for the retrieval of songs. Knowing a *geek* is obviously a good thing, but hunting him down near the coffee machine might become rather cumbersome in the long run. What if your machine and your favorite search engine could provide such a retrieval service?

Technically speaking, the ongoing development of information technologies makes it necessary to find ways of indexing and processing a wide range of data types. Textual information processing used to be the main focus of Internet-based search engines. A recent grow of interest for multimedia content, images, music and videos, can be observed. In this context, we focus on musical content, and more specifically on music signal processing. In particular, our work aims at addressing both the problem of main melody extraction, that is the predominant fundamental frequency estimation and melody note tracking, as well as the problem of lead instrument separation from the accompaniment.

In this introduction, we first give a general presentation of music signal processing. Then the motivations for the melody estimation and the lead instrument separation are discussed. At last, we present the contributions of our works and give the organization of this thesis.

1.1 Automatic music signal processing

Music signal processing has been a growing field of research for many years. Many related applications have been proposed: music instrument classification ([Essid et al., 2006a,b] or [Joder et al., 2009]), for instance, which aims at determining, given a music excerpt, which instruments are playing, tempo estimation and beat tracking ([Scheirer, 1998], [Cemgil and Kappen, 2003], [Ellis, 2007] or [Alonso et al., 2007]) or chord sequence estimation ([Papadopoulos and Peeters, 2008], [Bello and Pickens, 2005] or [Oudre et al.,

[1] such as the guitar opening of "Smoke on the Water", or the bass line from the opening credits of TV series "Mission: Impossible".

2009]). These applications are interesting for musical transcription purposes, but also for more general indexing and retrieval applications. The lower level (multiple) fundamental frequency estimation ([Klapuri, 2001] or [Emiya et al., 2009]) allows to decompose the sound into elementary atoms which may form "molecules" when grouped together in time ([Leveau, 2007]). Ultimately, these musical objects are identifiable as musical notes, and provide, in combination with the tempo, the rhythm and so forth, the *Grail* of music signal research: a musical score.

Of course, such an application is admittedly quite limited as it mainly targets musicians. A musical score, or anything related or converging towards a musical score, can however also be considered as a content-based indexation of the corresponding music excerpt, hence opening the results of music signal processing to a wider audience. Could I find some song that sounds like the ones I like? What are the other versions of that song? I don't know the lyrics or the title of that song, but maybe my computer could find it somewhere if I sang its melody (only slightly out of key)? This type of questions are addressed by the Music Information Retrieval (MIR) research field.

In textual information retrieval, in order to retrieve some documents, a query that represents what is desired must be provided. The machine then returns the documents that relevantly match with the query. Instead of the keywords which are fed into the Internet search engines, MIR research focuses on queries of another type, such as another music excerpt or a sung melody. However, a straight comparison of the waveforms does not always make sense: music signal processing aims at extracting the relevant information from these waveforms, projecting the signal onto another space where the comparison is more meaningful, mainly from a perceptual point of view.

On a higher level, this research topic may also lead to a better understanding of how music is produced, how it is felt and also how humans listen to and perceive music signals. To a certain extent, it also allows to have glimpses of how our brain works and addresses questions such as: what is music similarity? what is a tempo, a note or a rhythm? Music signal processing research is still young, and the computational issues music signal processing was experiencing not so long ago tend to slowly fall, allowing us to explore further horizons. Reaching reliable musical score generators or content-based music search machines however still seems a distant goal. The aforementioned tasks may be seen as milestones on the path towards these higher level, and more general, cognitive applications.

1.2 Main melody estimation

Among the elementary tasks that are sought after, the estimation and transcription of the melody of a song is a well studied topic. It is sometimes difficult to describe everything that appears within a song, but there are often particular objects, such as a melody, a chord progression or a rhythmc pattern, that characterize it. These "salient" objects are of decisive importance in the context of retrieval, since one may expect that the query will coincide with either of these cues. One such object is the main melody, which typically is that particular melody line that we sing or whistle. Such a feature can be theoretically used in applications such as indexing, Query-by-Humming (QbH) or cover version detection.

The melody extraction task consists in estimating the sequence of frequencies or notes of the melody, from polyphonic music signals. A very interesting assumption about this task, proposed by Goto [2000], is that we can extract a useful piece of information from the audio mixture, while avoiding to estimate, almost discarding, the other objects that are

present in the audio signal. In our case, this amounts to extracting a predominant melody line without estimating the accompaniment notes. Goto et al. [2002] also coined the term "Real World Computing" (RWC) audio, when they introduced the RWC database, which emphasizes the aim of such a task: we may not be able, for now, to fully understand or detect what is happening in complex audio mixtures. However, what if we try to describe just one part of the signal? How well can we do that? To what extent does this make sense?

There are no obvious answers to these questions. In our case, they may also raise other issues: what is a melody? what is *the* melody? How do we assess the success of such a detection problem? In the present work, it is believed that some melody concept can be extracted directly from the signal. We therefore avoid to use, for instance, any cognitive modelling or learning step, although, of course, some knowledge, mostly about the physical production aspects of the audio signals, needs to be included in the proposed model.

1.3 *De-soloing*: leading instrument separation

Music signal processing and Blind Audio Source Separation (BASS) have often been associated in recent works as in [Vincent, 2006] or [Gillet and Richard, 2008]. It is indeed widely agreed that many problems such as instrument classification, musical transcription or lyrics recognition may become much easier if the separated sources are available. In such a case, at best, a multiple fundamental frequency estimation for example becomes a "mere" monophonic fundamental frequency estimation. Unfortunately, the results from source separation algorithms, especially for real world signals, may still introduce artifacts and errors which cannot be compensated by the following processing steps. Some works have also highlighted the possibility of enhancing source separation using musical scores or any other musical annotations. This indeed provides an intuitively useful feature to help localize the different sources in time and in frequency.

In the context of the present work, the separation of the lead instrument, playing the estimated main melody, is also a proper way of assessing how well the estimation of the melody was. This separation task is sometimes referred to as "de-soloing" or "singing voice separation", especially when dealing with specific signals with a singer as lead instrument. It is easy to think of a way of using the melody estimated by some algorithm in order to separate the corresponding instrument, and inversely, using some monophonic pitch estimation algorithm on the separated leading voice signal provides the desired main melody. However, would it be possible to consider both these tasks in a unified framework? Would it be possible to jointly perform them? The present work aims at designing signal models that would enable such processing.

1.4 Contributions

During this phd thesis, we have developed a source/filter model for the singing voice and successfully used it within a statistical framework in the extraction of the main melody and the separation of the leading instrument, namely the instrument playing the main melody.

In particular, we have used a **source/filter model** for the singing voice together with a spectrum decomposition for the remaining accompaniment part, **within a unified statistical framework**. Following different assumptions on the signal, we have derived two specific frameworks for the singing voice. First, the **Gaussian Scaled Mixture Model**

(GSMM), as in [Benaroya et al., 2006], assumes that, given a dictionary of elementary sources, the voice signal is only produced by one active source per frame. Second, a so-called **"Instantaneous Mixture Model" (IMM)** is proposed. It assumes that the voice is produced by an instantaneous mixture of all the sources of the dictionary. Although not realistic from a production viewpoint, the IMM is more flexible than the GSMM and the correct source can be detected as a post-processing.

A clear improvement compared with works by Benaroya et al. [2006] and Ozerov et al. [2007], who inspired this work, in the specific case of single-channel leading voice / accompaniment separation, has been brought by the proposed source model. An advantage of our models may lie in the parameterization of the leading voice signal, which explicitly depends on the fundamental frequency. Our approaches can therefore be **unsupervised** while most previous works on the topic need to be supervised. The proposed production model, more complex and possibly closer to a realistic parameterization than former works, has also contributed to the improvement. Furthermore, compared with works by Vincent [2004], the proposed model is not limited to a set of musical notes mapped onto the Western musical scale. By using a larger dictionary of spectral shapes, encoded by their fundamental frequencies, **the model is even able to closely fit complex signals with great variabilities such as singing voice**.

The proposed asymmetrical models, where the leading voice and the accompaniment have distinct models, allows to **work independently on different aspects of each contribution**. More specifically, one can easily add prior knowledge to one model part without modifying the other part. It notably allows to control the complexity of the model: in our case, we reduced the complexity of the accompaniment part, in order to refine the model for the leading part.

In addition to the models, we have derived **two main parameter estimation algorithms**, first an EM algorithm in order to estimate the parameters for the GSMM and then a multiplicative gradient method for the IMM model, which is inspired by one of the most popular "Non-Negative Matrix Factorisation (NMF)" estimation method. These algorithms and the five systems that are derived from them show that, in spite of the complexity of the proposed models, approximate solutions can be found, with results and performances that are at the state of the art for each application. We also provide some hints and preliminary results on algorithms that aim at implementing the estimation with gradual release of the approximations made to the original models.

1.5 Organization

In this thesis, we first recall the motivation of this work: the estimation, transcription and separation applications for which the models were developed. The state-of-the-art techniques are also described and discussed in Chapter 2.

In Chapter 3, we then introduce the proposed signal models, both for the singing voice and the accompaniment part. The approximations that allow us to infer the parameters and address the different applications are then described and discussed.

After the signal models are presented, we give some results proving the equivalence between our framework and NMF estimation methods in Chapter 4. The results for that chapter explain how we could derive the algorithms given in Chapter 5.

The objective of the production models is not to generate audio signals, they are not meant to synthetize songs. We desire to use them to analyze music excerpts. In order to do so, five systems are proposed in Chapter 5. The needed algorithms that fit the

parameters of our models as well as possible to the audio data are also described, for each of the models, along with the sequence tracking algorithms that implement the temporal dependencies introduced in Chapter 3.

In Chapter 6, the application of these algorithms are discussed in the targetted tasks, namely the main melody estimation and the leading instrument separation.

At last, we summarize the contributions of this work in Chapter 7, where we also give some perspectives for future studies.

At the end of this thesis, the first appendices A and B provide some more details, respectively on the chosen complex Gaussian distribution and on the derivation of the algorithms of Chapter 5.

Chapter 2

State of the art

We are interested in this work in designing an audio signal model that can be used for two particular tasks: the estimation of the main melody and the audio separation of the main instrument and the accompaniment. These tasks are related to each other, as we will show in this chapter.

The first motivation for our work is the main melody fundamental frequency estimation from polyphonic music mixtures. The second application, leading instrument and background music separation, has emerged as being intrinsically linked with the problem at hand. If we have a solution to either of the problems, there intuitively is an easy way to address the other problem.

In Section 2.1, the definition of the "main melody" is discussed. In the subsequent sections, the different targetted applications, namely the estimation of the main melody and the separation of the leading instrument, are described and a review of the existing methods that address these applications is given.

2.1 What is the "main melody"?

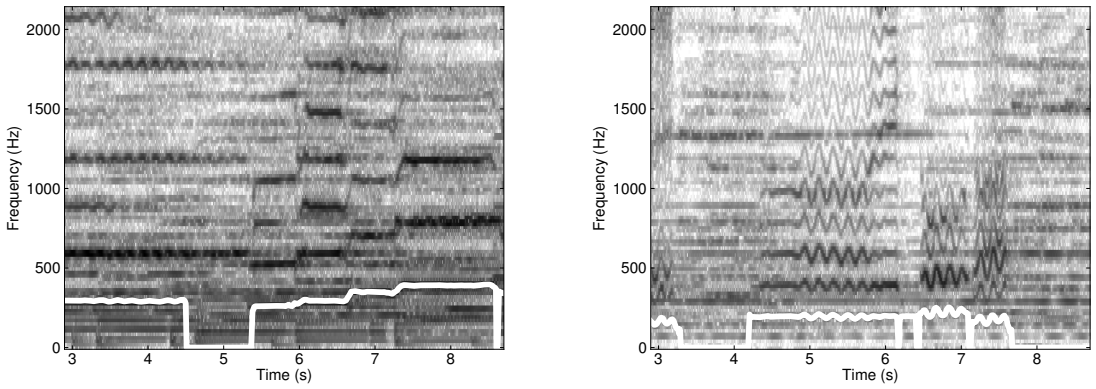
The original motivation for the present work is one of the tasks proposed at the Music Information Retrieval Evaluation eXchange (MIREX) evaluation campaign, namely the "Audio Melody Extraction" (AME) task. Its online definition reads[1]:

Goal: To extract the melody line from polyphonic audio.

Participants to the evaluation can therefore rather freely define the concrete characteristics and assumptions that allow them to discriminate between the notes of the melody and the background music. The organisers of the campaign have provided an annotated database so as to give to the participants an overview of what indeed is expected from their algorithms: more than using a formal definition, this task is defined by the available database. We could venture to say that one of the goals of the task is also to define the "melody line" as objectively as possible in order to be able to estimate it automatically. A description of the different databases for MIREX is provided in Appendix D.1.

Examples from the provided training database are given on Figure 2.1. The ground-truth for the desired main melody are also represented. The considered audio signals are all constituted of one instrument playing the melody, the leading instrument, along with some musical accompaniment. The most common style within these databases is "popular

[1] `http://www.music-ir.org/mirex/2009/index.php/Audio_Melody_Extraction`



(a) Daisy1.wav (b) Opera_male5.wav

Figure 2.1: Short-time Fourier transform (STFT) of 2 excerpts from the ADC2004 database. The ground-truth melody line is drawn as solid line over the STFT.

music", with many "pop-songs", but some jazz examples, vocal and instrumental ones, or some opera excerpts are also included. The main melody is therefore a concept which is hard to define. The main melody can be played by a musical instrument, a singer or even by a synthesizer. A common definition from the provided excerpts is not obvious. However, for most of the excerpts, there seems to be no ambiguity about what the main melody is.

The characteristics of the melody which are invariant across all these excerpts therefore have to be identified before one can build a model to detect it. A definition for the main melody is first discussed. Then we will rule out the signals that are not considered as containing a main melody. At last, some limitations and extensions are discussed.

2.1.1 A definition for the main melody

The definition given by Paiva [2006] in his PhD thesis, also on melody extraction, seems satisfying with regards to many aspects. It indeed reads:

Definition 4 (Predominant melody [Paiva, 2006]:) *Melody is the dominant individual pitched line in a musical ensemble.*$\diamond$

From this definition, as in [Paiva, 2006], the different elements can be further explained:

- *musical ensemble*: this expression first recalls the general environment. The processed signals are played by one or several musical instruments, which sets the potentially polyphonic background for the melody estimation.
- *pitched*: unpitched percussive sounds are thus excluded from candidates for the main melody.
- *individual*: the melody is assumed to be monophonic, played by one instrument at a time, in order to keep a certain coherence of the timbre. One may also add the assumption that the lead instrument is the same within one audio excerpt, as is done in [Paiva, 2006].

- *line*: this may highlight, to a certain extent, that the melody line is expected to exhibit some kind of smoothness.

- *dominant*: the leading character of the main melody is a difficult concept to formalize. It touches to the very nature of the desired auditory object: it mostly relies on a subjective assessment. However, some general, common sense rules can be observed. Physically, in order for the lead part to actually lead the rest of the audio mixture, the main instrument needs to play somehow louder, or with a pitch range, dynamics that make it more remarkable than the accompaniment.

Many works have also discussed on the definition of the melody, or the "main melody", such as [Gómez, 2002] or [Ponce de León et al., 2008]. The main melody can be defined on a more perceptual point of view, but a more neutral definition was preferred in the proposed work. Indeed, introducing perception in the melody definition also means that some subjective aspects of the underlying concept have to be incorporated. An objective definition relying solely on the signal, from a production point of view (and not a perception point of view), may as well lead to relevant results, as is shown in Chapter 6.

In the present work, the chosen definition for the melody also gives the characteristics of the main melody and what the proposed models should ultimately take into account. The above definition can therefore be seen under the following lights, which are also to be compared with the assumptions on the "predominant-F0" as described by Goto [2000]. These remarks also guided the choices made for our models:

- *musical ensemble*: the accompaniment may be polyphonic as well as inexistant. It is also not precisely characterized: the instruments playing this accompaniment part can be diverse, polyphonic, percussive, or even be the same as the lead instrument. The variability of the accompaniment has to be modelled and is addressed by the proposed choice in Section 3.3.2.2.

- *pitched*: the proposed models in Chapter 3 mainly focus on human voice as lead instrument, especially for the production model. In a more general fashion, the lead instrument can be assumed to own the following features: monophonic, harmonic, with a limited range for the fundamental frequency. Note that we are defining the melody, in this section, and not the instrument that is playing it. When the separation of that instrument is desired, one should also add some assumptions about the sound production process. For a singer voice, it may be necessary to add the unvoiced part, namely the consonants, as is done in our signal model, in order to obtain a better separation.

- *individual*: the uniqueness of the lead instrument is also assumed in the proposed model. However, this hypothesis may be only remotely modelled, as will be shown later.

- *line*: the continuity of the melody line can be modelled with respect to two aspects. The physical aspect reflects the assumption that the sound was produced by some physical system, a musical instrument or a human. The melody pitch line is therefore constrained to be smooth: the bigger the frequency interval, the harder it is for the practitioner. On a larger temporal scale, namely on a musical scale, this melody should also exhibit some musical note continuity. The difference between these two "constraints" are that the former physical constraint applies on shorter terms, but

with a stronger bind: it commands the fundamental frequency F0 level, while the second commands the note level, which is looser in frequency, but with longer temporal bounds.

- *dominant*: as in [Goto, 2000], the main predominance assumption of the proposed work comes from the energy dominance of the lead instrument over the accompaniment instruments. This does not mean that the lead instrument should actually have a higher energy than the accompaniment: the lead-to-accompaniment ratio may be lower than 1, with a relatively identifiable melody, for instance when the accompaniment is mainly composed of percussive instruments, while for some examples, this ratio may be over 1, but with a less clear melody line.

Note at last that while the above discussion aims at defining the melody *a posteriori*, one should not forget that for many of the considered styles, the melody appears explicitly right in the original music scores: the so-called "lead-sheet", which is the musical score format in which many pop-songs are published, gives in detail the melody line, that is the notes, the rhythms and tempi, the key, while the accompaniment is often indicated by a reduced form, that is the harmonisation, the drum style, if any, and so forth. The melody is therefore directly linked to the way the music was written and composed.

2.1.2 Main melody: counter-examples

In this section, some cases for which a "main melody" can not be defined without ambiguity are discussed.

First of all, in general, non-melodic percussive sounds such as those of the drums or the like, which are mainly meant to give the rhythm are discarded as "main instrument". As a consequence, the proposed models will not aim at fitting this kind of sounds, even though percussive sounds may have pitched components, such as the toms of a drum, the congas or any bell sound, for which the Glockenspiel is a good example. This assumption however also rules out instrument such as the xylophone or the vibraphone which are often used to play the melody, notably in Jazz music. As is discussed later, the proposed models mainly focus on singer voices or instruments that obey to a similar sound production process.

Sometimes, it is also difficult to identify a clear leading melody from a polyphonic mixture. Indeed, in cases like duets, chorals or some fugues, none of the voices or instruments really dominates. In some songs, the singer is supported by so-called backing vocals, which may consist in harmonization of the lead melody line. It then becomes hard to disconnect the lead from the other harmonies, since the goal of such techniques generally is to change the timbre of the lead itself, rather than to provide a mere support or background for the singer. As an example, on the song "pop1.wav", from the ADC2004 database, the first lyrics ("Michelle") is sung by several male singers, each voice giving a note of the chord. Without knowing the original song by The Beatles, it would be rather difficult to decide which of these notes should be considered as belonging to the lead melody.

At last, whithin a song, the singer may also stop singing. In these "silences", another instrument often either plays in turn the theme or improvises a solistic part. Should this instrument be transcribed as playing the main melody, or should it be left as playing some "fill-in"? This question may be answered in several ways. Many existing main melody transcription systems, as will be seen later in this section, include some classification stage, in order to identify whether the lead instrument is a singer or not, for instance.

2.1.3 Scope of this work

Not all polyphonic music signals possess a melody line. The proposed models do therefore not pretend to be useful in any musical situation. However, most of the commercial recordings, as for popular songs, do possess a melody. This research first aims at indexing files in order to provide a semantically relevant feature for retrieval systems. This activity has a meaning mainly for big databases, and especially for commercial recordings, with obvious economical consequences.

The MIREX databases are also mainly consistuted of songs. The lead instrument is therefore mostly a human singer, especially from the MIREX2005 database. Although this is the case in most popular songs, the definition of the main melody for the present work was not limited to a specific instrument. Indeed, choosing to transcribe as lead instrument only the human voice ([Sutton et al., 2006] or [Hsu et al., 2009]) rules out desirable signals such as instrumental jazz, or instrumental improvisations in rock music, and so forth. Our definition of the melody therefore allows to deal with this type of signals.

In some cases, not explicitly ruled out by the previously given definition, the annotated melody was harder to track for the systems proposed in our work. It was the case for a rap song, from the Musical Audio Source Separation (MASS) database [Vinyes, 2008], from the Music Technology Group (MTG), for which the separated lead instrument and accompaniment are rather disappointing. Rap is a style where the "singer" is actually more speaking than singing. To this extent, one may wonder whether such songs should or should not be considered for evaluating melody extraction systems and separation systems based on the melody line.

At last, the signals of interest are digital audio signals, mostly monophonic, *i.e.* with only one channel. An extension of our models to stereophonic signals is discussed in Section 6.2, as published in [Durrieu et al., 2009b]. Interestingly, multi-channel information has been very rarely used, if at all, for the task of main melody extraction, as is obvious from the review done in Section 2.2. The general assumption for the main melody, when one does not define it with the instrument that plays it, does not rely on spatial considerations. It is however surprising that this problem has never explicitly been addressed with source separation techniques. As browsed in Section 2.3, many source separation techniques rely on spatial information to estimate the different sources, but recent research work have also led to systems that can successfully process single-channel signals, in [Ozerov et al., 2007] for instance. The models proposed in this thesis find their roots in the models for these mono-channel source separation systems.

2.2 Main melody estimation

In this section, the motivation for main melody estimation is first discussed. The objectives and the applications for the results are also given. The estimation can be made in two granularities, with different applications in mind: first, the fundamental frequency estimation of the main melody, which describes the melody as precisely as possible, without any quantification, and second a more coarse yet musical result, based notably on the temporal quantification of the melody into notes. Both these tasks are discussed and reviewed in the following sections.

2.2.1 Main melody extraction: historical objectives and applications

One of the original works on this topic was done by Goto [2000], who described the task as a *predominant-F0 estimation* task [Goto, 2000]. The author introduced it for its relevance compared to the existing (multiple) fundamental frequency estimation tasks. Indeed, the state-of-the-art algorithms for fundamental frequency estimation, as proposed by de Cheveigné and Kawahara [2002], for monopitch estimation, or Klapuri [2008], Christensen and Jakobsson [2009] or Marolt [2004] for multi-pitch estimation, are restricted to audio signals with a somewhat low polyphony (less than 10 concurrent pitches). For the former methods, the monophonic assumption leads to systems that do not scale well when applied to polyphonic music. The latter methods are well suited for very specific types of signals such as piano music, chamber music or the like, but may not be adapted to our application, where we consider signals with a melody accompanied by a rather dense polyphonic accompaniment. The multi-pitch algorithms indeed usually aim at describing all the pitched content of the audio signal. However, for the considered genres and styles, such as popular music or jazz music, the sounds may be so complex that even well tuned multi-pitch algorithms may fail. We may also not desire to describe completely the accompaniment, since we are often interested only in the chord sequence that is underlying the accompaniment.

The approach by Goto [2000] is therefore first motivated by the need for a way of indexing these complex audio signals for applications such as query-by-humming (QbH) [Pauws, 2002] or [Ryynänen and Klapuri, 2008a], query-by-example (QbE), or cover version detection ([Ellis and Poliner, 2007], [Serrà et al., 2008] or [Foucard et al., 2010]). The principle is then that one does not need to describe the whole processed signal, leaving the accompaniment aside while focussing on the main melody line, as represented by a succession of fundamental frequencies. For almost 10 years now, the techniques proposed by several works ([Paiva et al., 2005], [Ellis and Poliner, 2006], [Goto, 2005], and an overview of several of these methods in [Poliner et al., 2007]) have improved so as to prove that such an approach is possible and may lead to results that are ready to be used in the targetted applications.

The task seen as a mere predominant F0 tracking may be unsatisfying, especially from a musicological point of view. Indeed, as reflected on the discussions for the Audio Melody Extraction (AME) task, at Music Information Retrieval Evaluation eXchange (MIREX) in 2005[2], the estimation of the melody may also be considered from a transcription point of view, hence finally outputting a musical score for the melody, or at least musical notes, with boundaries - onsetting and offsetting time, also respectively referred to as the "onset" and the "offset" - and a note label. Both these dimensions can be viewed as quantized versions of the F0 estimation, with respect to respectively time and frequency. Brute force quantization may not be the most satisfying method. Few works have however been proposed to tackle this problem: Ryynänen and Klapuri [2008b] is, to the best of our knowledge, the only publication providing such a transcription result for main melody estimation.

To sum up, there are two tasks related to the main melody estimation. A first task is the frame-wise predominant F0 estimation. The second task is the note-wise melody transcription which requires a (musically) quantized result. In the following sections, the frame-wise F0 estimation algorithms are first presented and the note-wise melody estimation algorithms are then explained. Each time, the advantages and the short-comings of

[2] online at: `http://www.music-ir.org/mirex/2005/index.php/Audio_Melody_Extraction`

the methods are discussed, and the contributions of our work are described.

2.2.2 Frame-wise fundamental frequency estimation of the main melody

As previously mentioned, one of the pioneering works on this topic has been carried out by Goto [2000], with a more complete description in [Goto, 2004]. Since then, many works have been proposed, with various methods ranging from auditory scene analysis to classification schemes. Many of those have been participating to the "Audio Melody Extraction" (AME) task at the MIREX evaluation, starting from the Melody Extraction Contest during the Audio Description Contest 2004 (ADC2004), held during the International Conference on Music Information Retrieval (ISMIR) in 2004 [Gómez et al., 2006] or the MIREX evaluation in 2005 [Poliner et al., 2007] up to the last edition of the evaluation during MIREX 2009[3].

Definition 5 (Frame-wise predominant F0 estimation:) *A system aiming at addressing the frame-wise F0 estimation application is expected to take as input the digital audio signal, process it, and provide a sequence of fundamental frequencies, along with the corresponding time stamps associated with the analysis frames. Concretely, the result may be written in an output file, with the following format for the n^{th} line:*

```
<Time of frame n (s)> <Tabulation> <F0 at frame n (Hz)>
```

The sequence of fundamental frequencies must correspond to the melody played by the lead instrument.$\diamond$

2.2.2.1 Existing approaches

Most of the main melody extraction systems that have been proposed so far exhibit some similar algorithm flow, with typically two main steps: an F0 candidate selection step, followed by a predominant F0 tracking step. The difference between these systems lies in the conceptual principle that motivates the extraction strategy. For the first step, the music signal is generally mapped onto a dimension which represents the F0 dimension. The second step may involve several principles, ranging from physical smoothness to higher semantic level models. To illustrate how these steps were implemented in previously published system, we propose the following brief descriptions. Because of its similarity with the present work, the system proposed by Goto [2004] is described in more details.

Goto [2004], for his PreFEst (Predominant-F0 Estimation) system, first computes a time-frequency representation of the music signal, using the Instantaneous-Frequency (IF) spectrogram proposed by Abe and Honda [2006]. Then the IF spectrogram is band-pass filtered, depending whether the system should retrieve the bass line or the melody line. At last a probability density function (PDF) is defined as follows: each frame of the filtered IF spectrogram is normalized, becoming a PDF $\widehat{p}^{(n)}(f)$, at frequency bin f and frame n.[4]

Then the PDF is decomposed with the assumption that it is the weighted sum of several PDFs, $\widehat{p}(f|f_0)$, which are themselves mixtures of Gaussian distributions with means harmonically related, parameterized by the fundamental frequency f_0. An example of what the elementary PDFs might look like is given on Figure 2.2. The global PDF for frame n

[3] Online: `http://www.music-ir.org/mirex/2009/index.php/Audio_Melody_Extraction_Results`

[4] A similar representation is adopted in [Raj et al., 2007], for audio signal separation, where the reader can find an interesting interpretation for such a statistical framework.

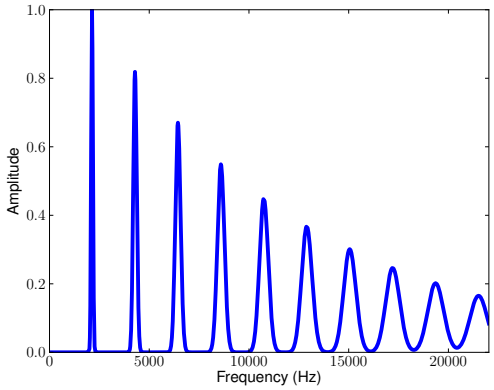


Figure 2.2: Example of a spectral comb PDF $\widehat{p}(f|f_0)$, generated as explained in Goto [2004]

then approximated such that:

$$\widehat{p}^{(n)}(f) \approx \int_{f_0} \omega_{f_0}^{(n)} \widehat{p}(f|f_0) df_0 \tag{2.1}$$

where $\omega_{f_0}^{(n)}$ is the weight associated with fundamental frequency f_0, at frame n. The method proposed in [Goto, 2004] then estimates the parameters for the tone models that give the spectral shape in $\widehat{p}(f|f_0)$ as well as the weights $\omega_{f_0}^{(n)}$. These weights can thereafter be considered as the aforementioned mapping from the signal to the F0 dimension. These parameters indeed give, for each frame n and fundamental frequency f_0, the relative strength of f_0 within frame n. The weights and all the relevant parameters are estimated through an Expectation-Maximization (EM) algorithm [Dempster et al., 1977], maximizing a "mean log-likelihood" defined as the integral over the frequencies of the product between the "observed" PDF $\widehat{p}^{(n)}(f)$ and the logarithm of the parametric PDF $\sum_{f_0} \omega_{f_0}^{(n)} \widehat{p}(f|f_0)$:[5]

$$\sum_f \widehat{p}^{(n)}(f) \log \left(\sum_{f_0} \omega_{f_0}^{(n)} \widehat{p}(f|f_0) \right) \tag{2.2}$$

It is interesting to note the similarity between this estimation problem and works that have been done in Non-negative Matrix Factorization (NMF), notably by Lee and Seung [1999], where the objective function is actually equivalent to the opposite of the Kullback-Leibler divergence, as explained in [Lee and Seung, 2001]. Indeed, all the quantities in $\sum_{f_0} \omega_{f_0}^{(n)} \widehat{p}(f|f_0)$ are non-negative, especially the weights. Furthermore, the similarity between the criterion Equation (2.2) and the objective function of Lee and Seung [1999] is striking. Note however that, to some extent, the criterion Equation (2.2) may suffer from the fact that it is unbounded: for any given values of the weights, there will always be a weight set, such that the criterion is higher. This is addressed by adding priors on these parameters, such as the normalization of these weights [Goto, 2004].

The melody tracking is then held on the obtained "matrix" of the $\omega_{f_0}^{(n)}$ weights by a *multi-agent structure* [Goto, 2004]. Several tracks are created and terminated using rules

[5] Note that all the notations adopted here are not the same as the ones used in [Goto, 2004], and the general formulation was also summarized, while trying to respect the author's method principles.

that depend on the values of the $\omega_{f_0}^{(n)}$ along the tracks, much like a Viterbi smoothing algorithm [Viterbi, 1967], which would follow several tracks, instead of one optimal one.

An advantage of this method, as explained in [Goto, 2004], comes from the fact that the mixture parameters $\omega_{f_0}^{(n)}$ are jointly estimated, hence providing a salience function for each frame which is smoothly derived from the signal. It is more specifically to be compared to previous works on multiple F0 estimation, some of which being iterative estimation-substraction algorithms ([de Cheveigné, 1993] or [Klapuri, 2001]). In these works, a hard decision is made when substracting. Such a process may lead to inaccurate estimations on the resulting residuals. Being able to jointly estimate all the constributions that constitute the signal is believed to be a desirable feature when building a model. Note at last that some other systems have been using the first stage of PreFEst as pre-processing, such as in [Marolt, 2005] or in [Fujihara et al., 2006].

Many works have also taken into account some perceptual assumptions made about human auditory system. Ryynänen and Klapuri [2005], notably, use on a first step the multipitch algorithm by Klapuri [2001], to compute a pitch salience function, along with features reflecting some onsetting properties. The front-end step in [Klapuri, 2001] is intended to mimick how the the human ear treats the audio signal, further processed with perceptual principles derived from [Meddis, 1986]. The salience of the pitches in a given range is computed thanks to an equation close to a subharmonic summation (SHS) [Hermes, 1988], a technique also used by many other systems ([Cao and Li, 2008], [Hsu et al., 2009] or [Wendelboe, 2009]). The salience, as well as other features describing the onsetting character of the corresponding f_0 and computed from the signal, form a vector of observation. The observation is assumed to depend on some underlying hidden state sequence of note, coined as "note-events". The evolution of the observation within each note is modelled through a hidden Markov model (HMM), while the evolution of the sequence between the notes depends on a supervised musicological layer. The note-event model provides a very interesting framework, theoretically sound, and with an easy interpretation of the different parameters involved in the model. Another positive aspect of the note-event model is the fact that the instrument which plays the melody and the accompaniment are described by different models, such that the discrimination of the main melody is also potentially grounded on another feature space than only the pitch and the energy ones.

However, this method as well as many others, notably those using auditory based representations does not lead to a straightforward solution for source separation. Usually, this is circumvented using sinusoidal model [Ryynänen et al., 2008], at the cost of sub-optimality of the estimates, since the whole process, namely the transcription and the separation, can not be done but sequentially, first transcribing the melody and then removing the corresponding source from the mixture. The signal model proposed in this thesis aims at allowing a joint estimation of both the melody and the corresponding lead instrument separated signal. The spectral models that are used originate from some of the source separation techniques that are browsed in Section 2.3.

It is worth at last to discuss the classification based approach studied by Ellis and Poliner [2006]. A Support Vector Machine (SVM) classifier is used to detect which note was played. The result is also a frame-wise F0 sequence output, although the quantification onto the Western musical scale is actually also provided at the same time. The features that were tested are mainly derived from the short-time Fourier transform (STFT) of the signal, with different strategies to normalize them and obtain better results. This approach is motivated by the classification aspect of the task. Indeed, the human auditory system along with the brain can be considered as a pattern learning and pattern matching machine.

The advantage of such an approach is that very few prior assumptions are provided to the system, such that the data on which it is trained actually defines the desired labels. To a certain extent, the systems by Marolt [2004] or Ryynänen and Klapuri [2008b] are also involving classifiers, be it artificial neural networks or hidden Markov processes. However, the system designed by Ellis and Poliner [2006] still requires the least assumptions for estimating the F0 sequence, since the other aforementioned systems assume some sort of structure for the main melody.

2.2.2.2 Discussion and position of the thesis work

Our approach includes several original contributions when compared with the other systems. First, specific and different models are used for each component (leading instrument versus accompaniment) of the music mixture to take into account their specificities and their production process. Indeed, since this study focuses on signals for which the predominant instrument usually is a singer, there is a particular interest to exploit the physical characteristics of the production of the human voice compared to any other instrument as in Sutton et al. [2006]. It is then proposed to represent the leading voice by a specific source/filter model that is sufficiently flexible to capture the variability of the singing voice in terms of pitch range and timbre (or more specifically the produced vowel). The resulting decomposition is presented in Chapter 3, and the models bear similarities with works by Goto [2004], at least in the interpretation given to the estimated parameters. However an interesting originality of the models lies in the use of **"physically-inspired" basis functions**, especially for the source part of the lead voice, while the other methods mainly give approximated parametric spectral comb: the Gaussian mixtures of Goto [2004] may indeed not fit, from a generative point of view, the chosen IF representation. Our choice of the time-frequency representation also better suits source separation purposes.

Second, unlike many existing systems, **the accompaniment is explicitly modelled** in our framework. It is assumed to include instruments that exhibit more stable pitch lines compared to a singer and/or a more repetitive content (same notes or chords played by the same instrument, drum events which may remains rather stable in a given piece and so forth). To exploit this relative pitch stability and temporal repetitive structure, the model for the accompaniment is closely related to **Non-negative Matrix Factorization (NMF) with the Itakura-Saito (IS) divergence** [Févotte et al., 2009a]. Our systems further discriminate between the leading instrument and the accompaniment by assuming that the energy of the former is most of the time higher than that of the latter.

Third, the leading voice is modelled in a statistical framework in which two different generative models are proposed. Both of them include the previously mentioned source/filter parameterization. The first model is a source/filter Gaussian Scaled Mixture Model (GSMM) Benaroya et al. [2006] while the second one is a more general Instantaneous Mixture Model (IMM). The generative model is essentially inspired by single-channel blind source separation approaches presented in Benaroya et al. [2006] and Ozerov et al. [2007]. **We can therefore also proceed to the actual separation of the estimated solo part and background part**, within the same framework as the estimation itself. An overview of the existing methods that address similar tasks is given in Section 2.3.

At last, many aspects from previous works seemed important for the elaboration of our models. For instance, as Goto [2004], we believe that a decomposition that estimates all at once the different contributions of each considered note is more satisfying than the alternative method which consists in estimating and substracting iteratively these contributions,

much like Matching Pursuit algorithms. The models proposed in Sections 3.3.2 and 3.4.1 therefore aim at allowing a joint decomposition. The distinction between the models for the lead instrument and the accompaniment have also been investigated by Ryynänen and Klapuri [2005] or Hsu et al. [2009]. These works may however be more related to a classification problem, where we propose models that allow to describe, potentially in great details, both the lead voice and the background music.

2.2.3 Note-wise approaches

The results of our methods demonstrate the relevance of the frame-wise approach in a number of applications. A further step towards a transcription in terms of musical score is to quantize the pitch along the (Western) musical scale and to provide the time instants for the start (onset) and the end (offset) of each note.

Definition 6 (Note-wise melody estimation:) *A system aiming at providing a note-wise transcription of the melody is expected to return the MIDI code numbers, onsets and offsets corresponding to each of the notes that appear in the melody. A line of the output file may therefore look like the following:*

```
<onset (s)> <Tabulation> <offset (s)> <Tabulation> <MIDI code number>
```

The desired notes are the notes played by the lead instrument.◇

There are not so many works on this particular topic. No international evaluation has been set up yet. However, some works have been proposed, such as [Paiva, 2006], [Ryynänen and Klapuri, 2006] or [Ryynänen and Klapuri, 2008b]. In the former, the author uses many rules and heuristics to cluster the different melodic F0 streams into note streams, and then detects the true notes from the spurious ones. The latter approach, as described previously relies on a "note-event" model which explicitly labels the notes, and within these notes, the evolution of the pitches is observed.

More works have been done in the somewhat more difficult task of polyphonic transcription [Ryynänen and Klapuri, 2005, Marolt, 2004, Emiya et al., 2009, Bertin et al., 2010]. This task is more difficult in the sense that the systems are expected to output several concurrent notes, while the main melody extraction task only requires one note at a time. For the latter problem, one may desire the melody of a rather complicated audio signal, on which most polyphonic music transcription systems would probably fail. It would be tempting to take advantage of this litterature on polyphonic music transcription to proceed to the main melody extraction: combining the output of these algorithms to results of melody estimation on symbolic data ([Rizo et al., 2006] or [Ponce de León et al., 2008] for instance) could lead to fairly good results. However, stream discrimination, namely estimating which instrument played which notes in the resulting output, is generally needed. Although with symbolic data the tracks are usually well separated, it is still an open problem when dealing with audio signals, which has only been partly addressed in works by Leveau [2007] or Duan et al. [2009].

At last, one should note that very few works have aimed at transcribing the resulting melody into what one could call "human readable" music score. Indeed, for a computer to understand, reproduce or process for similarity tasks these resulting melodies, the notes with their MIDI labels and time boundaries is usually sufficient. However, for a (human) musician, it may be challenging to play the resulting score, when metric indications is omitted. Cemgil [2004] proposes a statistical frame-work within which it is possible to

design penalty functions on the "complexity" of the duration and rhythm quantization result ([Cemgil et al., 1999]), in a context where the onsets of the notes are assumed to be known. It is worth noting that the model proposed in [Cemgil, 2004] could also be extended to estimate the whole chain of musical quantities, namely the fundamental frequencies, the notes, their rhythms and so forth. The model proposed here also provides such a flexibility and the different proposed systems aim at progressively including the whole chain within only one estimation phase.

2.3 Source separation, leading instrument separation

Having estimated the desired melody line, an interesting application is to separate the instrument playing the melody, the "lead instrument", from the rest of the mixture, the "accompaniment". Such an application has the advantage that it allows to subjectively assess the detected melody line: did we get the desired target instrument? It also gives some insight concerning the model and how much it fits to the data, notably using the (objective) evaluation tools from the source separation field.

In the following, generic source separation algorithms are first presented. Systems more specifically oriented towards musical audio source separation are then discussed.

2.3.1 Source separation

For the instantaneous linear mixture case, which this study is limited to, the source separation task is defined as in the following definition.

Definition 7 (Source separation (instantaneous linear mixture):) *Let* $\mathbf{s}$ *be the source random vector of size* I *(the number of sources). The* $J \times I$ *mixture matrix* $\mathbf{A}$ *is defined such that the observation vector* $\mathbf{x}$ *of size* J *(the number of sensors) writes:*

$$\mathbf{x} = \mathbf{A}\mathbf{s} \tag{2.3}$$

A source separation system aims at estimating the sources $\widehat{\mathbf{s}}$ *knowing the observation* $\mathbf{x}$.$\diamond$

The aim of source separation is to extract separated contributions (or sources) from a mixture by exploiting their differences in terms of spatial location and/or time-frequency (or timbral) content. It is common to categorize the source separation problem according to the difference between the number of available sensors or channels J and the number of sources or contributions I.

- $I \geq J$: over-determined case (determined case, when $I = J$)
  To address this problem, many works have been proposed. Famous techniques to "invert" the system and estimate the mixing parameters include the Principal Component Analysis (PCA) [Pearson, 1901], or the Independent Component Analysis (ICA) [Jutten and Herault, 1991] and [Comon et al., 1994]. All these techniques somewhat rely on low level characteristics of the sources, such as their mutual independence or their spatial positions with respect to the sensor array.

  Non-negative versions of ICA [Plumbley, 2003] and [Abdallah and Plumbley, 2004] allow to describe signals for which there is a structural non-negativity constraint on either the mixing matrix or the sources themselves, such as power spectra for audio signals. In such a decomposition, without constraint, a negative power spectrum, that is to say a destructive spectrum, would be hard to interprete and might confuse a pattern recognition algorithm, especially if it aims at mimicking human perception.

- $J < I$: under-determined case
  In many cases, the only available observation $\mathbf{x}$ is limited. As human beings, we only have 2 eyes and 2 ears. This means that, using only 2 sensors for vision or audition, the human "machine" is able to separate, or at least detect or focus on different sources (or objects) that compose what he sees or hears. Following this observation, a source separation system could be expected to be able to decode very difficult signals such as stereo audio signals ($J = 2$), or even more difficult ones such as mono-channel audio signals or images ($J = 1$). Many works have been done in the field of mono-channel audio signal separation, as will be discussed in Section 2.3.2.

There is a wide diversity of applications for source separation in general. In audio signals, to separate speech signals from musical sounds [Benaroya et al., 2006], separate the singer [Ozerov et al., 2007] or separate different concurrent speakers [Weiss and Ellis, 2010, Le Roux et al., 2007]. Source separation can also be used in medical applications, as in neuro-science [Mørup et al., 2008]. Some data can also provide, after processing, interesting results in astro-physics [Cardoso et al., 2008].

2.3.2 Audio and music source separation

In the field of audio source separation, many reviews have already been proposed by Vincent [2004] or Virtanen [2006], for instance. In this section, a particular focus on techniques that were proposed to address the specific task of separating a specific source, typically a human voice, from an "undesired" background audio environment, be it noise, for speech enhancement applications, or music, for singing voice extraction. This study more specifically focuses on monaural lead instrument separation, and the following overview aims at browsing what has been done in that particular field or in closely related fields.

2.3.2.1 Existing systems

A number of audio source separation approaches such as [Benaroya et al., 2006] or [Ozerov et al., 2007] rely on supervised techniques to extract the vocal part from any other musical background. They introduce a statistical and flexible framework. The sources are specified and classified by their spectral characteristics. They are then separated using a Wiener time-frequency mask. Approaches like [Lagrange et al., 2008], [Li and Wang, 2007] or [Ryynänen et al., 2008] rely on sinusoidal models and unsupervised techniques to label several groups of sinusoids as belonging to either of the expected sources. In [Davy and Godsill, 2003, Davy et al., 2006], sinusoidal models are also used, but the estimation of the different parameters is done within a Bayesian framework. The use of a sinusoidal model however usually impairs the subjective quality of the results, creating very typical artefacts, especially in high frequency components of the signals.

The overview from Section 2.2.2 is related to the latter type of approach. Indeed, works by [Li and Wang, 2007] or [Ryynänen et al., 2008] rely on a pre-processing that detects the main melody - or more specifically the *singing melody* - in order to proceed to the actual separation using, respectively a time-frequency binary mask Roweis [2001] and sinusoidal modelling. Another approach proposed by Han and Raphael [2007] for "desoloing" monaural music signals, also based on an ideal time-frequency binary masking, is quite related to the work proposed in this thesis. However, the technique proposed in the aforementioned paper is limited by the need of having a musical score already aligned on the audio. Furthermore, the technique seems to provide very good "desoloing" performances,

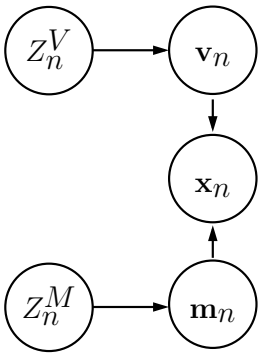


Figure 2.3: [Benaroya et al., 2006]: Graphical model for the observation layer, first layer dependency for the mixture. The Fourier vectors for the voice $\mathbf{v}_n$ and the music $\mathbf{m}_n$ are respectively generated through the states Z_n^V and Z_n^M. The mixture vector $\mathbf{x}_n$ is the sum of $\mathbf{v}_n$ and $\mathbf{v}_n$, and thus only depends on these vectors. The only observed variable is $\mathbf{x}_n$.

but does not lead to satisfying results in extracting the lead instrument, which admittedly was not the goal of the corresponding application. Nevertheless, all these works fall into the general topic of musically-informed source separation.

The other type of approaches is more specifically embodied by the model proposed by Benaroya et al. [2006], which aims at separating speech from background music. In order to do so, each contribution, speech and music, is modelled by a **Gaussian Scaled Mixture Model (GSMM)**. The chosen representation is the short-time Fourier transform (STFT), such that each state of the GSMM, for each contribution, characterizes a specific spectral shape. The model in [Benaroya et al., 2006] can be represented by a graphical model, as shown on Figure 2.3. It assumes that the mixture $\mathbf{x}_n$, at frame n, is the sum of the two contributions to be separated, the voice $\mathbf{v}_n$ and the music $\mathbf{m}_n$. Each contribution is controlled by its own state in the GSMM, respectively Z_n^V and Z_n^M. This technique however requires some supervision, as no prior or production model is included in the GSMM.

One of the interesting contributions that Benaroya et al. [2006] brought to audio source separation is the use of Wiener filtering, even for cases where the signals are not completely stationary, but *locally stationary*. Works dating back to McAulay and Malpass [1980] or Ephraim and Malah [1984] also use similar techniques, also in the time-frequency domain. They however only consider the simpler case where the second contribution is some stationary noise. Another comparison could be drawn with techniques using a binary masking of the spectrogram: Roweis [2001] or Jourjine et al. [2000], with the DUET system, separate sources that assumedly do not overlap in the time-frequency domain. Such approaches are necessary when one does not estimate the respective energies of each contribution. Benaroya et al. [2006], in a way, extended their works, and showed that the estimation of both the spectral shapes of the speech and the concurrent background music is possible and may lead to good separation results. One should however keep in mind that transformations on time-frequency representations such as the STFT may lead to "inconsistent" STFTs as discussed in [Benaroya, 2003] and addressed by Griffin and Lim [1984] or Le Roux et al. [2008].

This separation technique was further developed by Ozerov et al. [2007], applied to singing voice separation from the accompaniment. The system first detects the frames

with and without vocal part, and then estimates the different contributions using the same formalism as in [Benaroya et al., 2006]. The obtained results show how important it is to be able to discriminate the two contributions, and that a good frame classification leads to better spectral estimations. The approach by Weiss and Ellis [2010], aiming at separating two speech signals, uses a similar framework, and the authors propose in [Weiss and Ellis, 2009] an efficient separation algorithm based on hidden Markov models (HMM) of the signals.

Many audio source separation algorithms using Non-negative Matrix Factorisation (NMF) have also been proposed. NMF was made popular in image processing by Lee and Seung [2001] and since then has been used in many other fields, such as audio processing, but also for brain image decomposition ([Mørup et al., 2008] or [Cichocki, 2004] and [Cichocki, 2002]). The parallel between the image and spectrogram models is clear in [Smaragdis and Brown, 2003], [Virtanen, 2007], [FitzGerald et al., 2008] and [Févotte et al., 2009a] to cite but a few. A similar parameter estimation technique can also be found in [Benaroya et al., 2003] and [Benaroya et al., 2006]. The source/filter model proposed in this thesis was also studied for very similar purposes, within an NMF context, in [Virtanen and Klapuri, 2006].

At last, it is important to note the similarity between our work and the musical source separation algorithm proposed by Vincent [2004]. Indeed, a statistical spectral model is proposed therein, in order to enable several tasks within a unified framework, among which musical transcription and source separation. It is based on pre-trained spectral shapes, with additional corrective spectra that aim at addressing several issues such as vibrato or timbre variations. The chosen representation is the logarithm of the power spectrogram, expressed on a logarithmic frequency scale. The logarithmic scales for the power and the frequencies are motivated from a psychoacoustical point of view.

2.3.2.2 Position of the thesis work

Our algorithms take advantage from both the musically inspired approach and the more straight-forward source separation approach. The melody line of the lead instrument is used as in [Ryynänen et al., 2008], but the separation is held within a statistical framework adapted from Benaroya et al. [2006]. Our methods are **unsupervised**, and thus differ from the supervised techniques of Benaroya et al. [2006] and Ozerov et al. [2007]. Furthermore, the framework may allow to perform all the tasks, melody estimation and separation, jointly, instead of sequentially as done in [Ryynänen et al., 2008] or [Heittola et al., 2009]. We have been working towards this joint estimation, although the preliminary results do not yet provide the expected better performances.

Our work also overcomes some indeterminacies inherent to the source/filter model from Virtanen and Klapuri [2006], notably by setting in advance the dictionary of spectral shapes for the source part, at the cost of restricting the type of lead instrument that is considered. Our choice of signal model, namely a source/filter to represent the lead instrument, provides another advantage over the separation systems by Ozerov et al. [2007] (or Benaroya et al. [2006]) and Vincent [2004]. Indeed, compared to [Ozerov et al., 2007], our decomposition can be semantically interpreted, as will be seen in Section 3.3.2.1: the parameters for the source part of the lead mostly represent its fundamental frequency F0 while the filter part bears the spectral shape or formant information (which we do not explicit use in this thesis). In [Ozerov et al., 2007], the states of the GSMM do not have an explicit link to the content of the signal, and their interpretation relies on the supervision

stage and on the data that were provided during the training.

Furthermore, our source/filter model, and its meaning in terms of F0 and formants, provides an intermediate level of description which is missing in [Vincent, 2004]. Indeed, Vincent [2004] models the signal directly thanks to musical note level, with an intermediate layer of descriptors which are learnt, such that, as in [Ozerov et al., 2007], their interpretation is not obvious, although possible. The link between the parameters of our model and notions such as the F0 or the formants makes it easier to understand how the estimation is done, whether it succeeded or not and how to design constraints on these parameters, as will be discussed in Section 3.4.3, with an example of constraint on the parameters corresponding to the energy of each F0.

At last, extensions to stereophonic mixtures, hence taking advantage of both spatial and frequency structures, are possible and have been studied in [Ozerov and Févotte, 2010], [Arberet et al., 2010] and in [Vincent, 2004], but also in this thesis, as developed in Section 6.2.4.2.

Chapter 3

Signal Model

In this chapter, the proposed signal model for the desired applications, namely main melody estimation and separation, is developed and discussed.

The generic signal processing framework for the proposed model is first given in Section 3.1. In Section 3.2, the statistical signal processing framework is further developed. In Section 3.3, the primary parametric models for both the leading voice (or main instrument, playing the main melody) and the accompaniment are proposed. The first aim of this model is to be as realistic as possible, given the assumptions on the signals. Since this leads to a complicated estimation problem, another model for the leading instrument is proposed in Section 3.4. This new model essentially leads to faster estimation algorithms. Although not as realistic as the first model, the interpretations of the estimated parameters are satisfying.

At last, a summary for each of the models is given in Section 3.5. These summaries are meant to provide a global view of the models through the equations that define the different dependencies and evolutions developed separately in Sections 3.3 and 3.4.

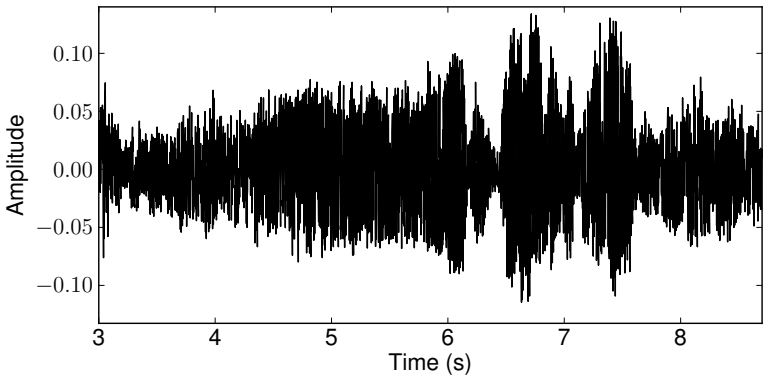
3.1 Modelling the spectrum of the audio signals

As for many previous works in music transcription as well as in audio source separation, the proposed models rely on a time-frequency representation of the signal, which roughly mimics auditory perceptual mechanisms.

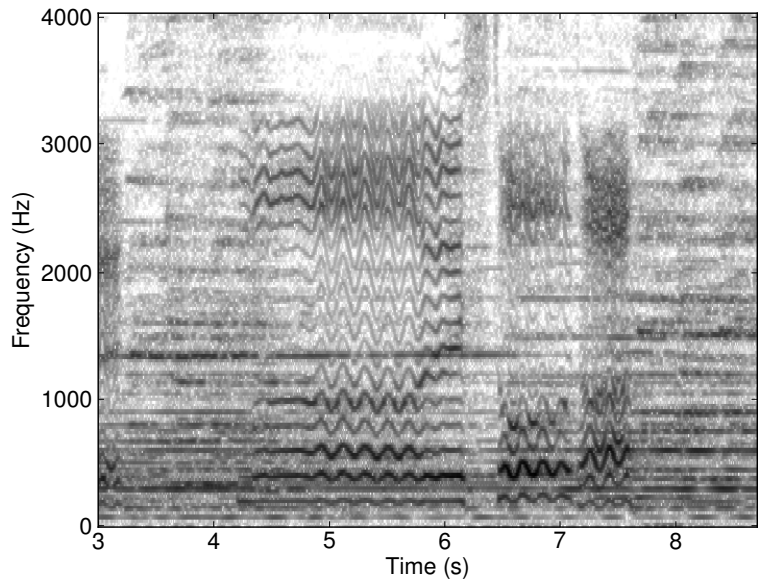
For the purpose of the applications at hand, the **short-time Fourier transform (STFT)** is well adapted. Let y be a discrete time-domain signal. We further assume that y is a single-channel signal[1]. y is first segmented into frames of a given length L, with a constant hopsize L_{hop}. The total number of frames is denoted by N. For frame $n \in [1, N]$, the Fourier transform $\mathbf{y}_n$ of size $2F$ is computed. The element at frequency bin $f \in [1, F]$ is denoted as y_{fn}. The Fourier tranforms for all the frames are stacked as column vectors in the $F \times N$ STFT matrix $\mathbf{Y}$.

An example of such a representation is given on Figure 3.1. The represented excerpt, "opera_male5.wav", from the ADC2004 database (see Section 6.1.1.4 for a description of the databases). On the figure, the power of the STFT is represented with different degrees of gray from light to dark, ranging from low to high power values, on a decibel scale.

[1] An extension for stereo signals is given in Section 6.2.4.2. However, since our method does not rely on spatial information, but rather on the spectral properties of the desired sound objects, for better clarity, in this section, we only present the theory for single-channel signals.



(a) Waveform of the file "opera_male5.wav"



(b) STFT of the above waveform

Figure 3.1: STFT example: excerpt from ADC2004 database, "opera_male5.wav". Darker colors correspond to higher energy, proportional to the squared magnitude of the STFT (its "power"), in dB. The analysis window length is 46.44ms, and the overlap ratio is 87.5 %, or, equivalently, the hopsize between the analysis windows is 5.8ms.

Many works in music transcription use perceptually and musically motivated time-frequency transforms such as the so called "constant Q transform" (CQT) [Brown, 1991] or discrete wavelet transforms (DWT) [Mallat, 2008]. These representations allow to adjust the frequency resolution, in order to obtain, for the CQT, a better frequency resolution in low frequency components and a better time resolution in high frequency ones. Rather than directly considering the signal itself, many works in fundamental frequency estimation also consider how the human ear processes the information. Such a perception based approach is motivated by the assumption that the desired quantities, namely the pitch sequences, are perceptually defined concepts, and should therefore be processed by some system mimicking the human auditory functions. One of the most illustrative examples of such systems is probably the system by Large and Kolen [1994], who used oscillators to simulate the way biological cells are behaving, in the case of meter estimation. Similar biologically inspired systems have also been proposed, by Scheirer [1998], for beat tracking, or by Marolt [2004], for piano music transcription.[2]

The use of the STFT can however be advantageous for several reasons, notably in the case of source separation. First, the STFT is invertible. Using an overlap-add (OLA) procedure allows to obtain, with mild conditions on the parameterization of the hopsize and the analysis and synthesis windows, a perfect reconstruction of the signal . Some strategies exist to "invert" the CQT and the DWT, but they all lead to approximate solutions (such as in [Fitzgerald et al., 2006] or in [Slaney et al., 1994]), which may introduce some bias in the (objective) performance criteria. Another reason is that the proposed model does not require a better precision in low frequencies than in higher frequencies, as will be discussed in Section 5.1.

For the present work, the chosen parameters to compute the STFTs are as follows:

- **Window length:** around 46.44ms. At a sampling rate of 44100Hz, this correspond to a window length of $L = 2F = 2048$ samples. Such a size of analysis windows allows a good trade-off between time resolution and frequency resolution.

- **Window type:** for the main melody transcription application, the analysis window is the Hann window. The formula for a Hann window $\mathbf{w}$ of size $2F$ is, for $t \in [0, 2F-1]$:

$$w_t = 0.5 \left(1 - \cos \frac{\pi t}{F}\right) \tag{3.1}$$

  The frequency resolution, in terms of standardized frequency, is $\frac{1}{F}$.

  For the source separation application, the analysis weighting window is the cosine window (also known as sinebell window). The formula of such a window $\mathbf{w}$ of size $2F$ is, for $t \in [0, 2F-1]$:

$$w_t = \sin \frac{\pi t}{2F} \tag{3.2}$$

[2] This last reference is most interesting, as the system proposed therein is almost only designed after ear perception models and connectionist models: the signal is processed through a perceptual filter bank, whose outputs are further enhanced in accordance with perceptual principles [Meddis, 1986]. Many layers of post-processing are then used, essentially neural networks, to obtain frequency activity and note detections, which adds to the degree of similarity between the system and the auditory and cognitive model for human hearing.

The frequency resolution, in terms of standardized frequency, is $\frac{3}{4F}$. For source separation, the perfect reconstruction from the STFT is possible using an overlap-add (OLA) procedure, with the same window as synthesis window, and an overlap ratio of 50 %.

- **Hop size:** Various hopsizes have been tested. At the Music Information Retrieval Evaluation eXchange (MIREX) campaigns, the Audio Melody Extraction (AME) task, the provided ground-truth was given at frames spaced by 5.8ms (ADC 2004, equivalent to 256 samples at a sampling rate of 44100Hz) and 10ms (MIREX 2005, 441 samples at 44100Hz). The hopsize of 5.8ms better corresponds to source separation expectations, as it provides an overlap ratio of 87.5 % (7/8), while with the 10ms, it becomes a ratio that is less convenient for reconstruction (1607/2048), although not impossible. The hopsize of 5.8ms was therefore preferred in most of the presented algorithms.

  As for the perfect reconstruction, as discussed above, the regular overlap ratio for the sinebell window is 50%. However, it is interesting to consider that the overlap ratio of 87.5% actually corresponds to computing 4 STFTs with 50% of overlap between the frames. The perfect reconstruction therefore also holds here, after a normalization by a factor 4 of the output of the OLA procedure.

3.2 Gaussian Signals

The statistical model initially discussed in [Benaroya, 2003] and later in Ozerov [2006] for application in audio source separation involves a generic statistical framework where the spectral shapes of the different contributions of the mixture are explicitly modelled. This framework still leaves room for further modelling of these spectral shapes, as is proposed in this work. Similar statistical signal models can be found for speech enhancement [Ephraim and Malah, 1984] or even cosmic microwave background separation [Cardoso et al., 2008].

For a given signal y, the n^{th} frame of the STFT $\mathbf{y}_n$ is modelled as a complex random variable, following a proper multivariate complex Gaussian distribution, defined in Appendix A.1.1:

$$\mathbf{y}_n \sim \mathcal{N}_c(\boldsymbol{\mu}_n^y, \boldsymbol{\Sigma}_n^y)$$

We further assume that the vectors are centered ($\boldsymbol{\mu}_n^y = 0$) and that their covariance matrix $\boldsymbol{\Sigma}_n^y$ is diagonal.

The first assumption can be easily understood for its consequence in the time domain: indeed, if the Fourier transform $\mathbf{y}_n$ of frame n of y is centered: the mean of the n^{th} frame of y is 0. This assumption is therefore compliant with a general assumption that audio signals are centered.

The second assumption is equivalent to assuming the independence of the frequency bins of the Fourier transform: for wide sense stationary (w.s.s.) signals, this assumption holds. Let $\{x_t\}_{t\in\mathbb{R}}$ be a w.s.s. process, centered, then the auto-covariance function r^X of x only depends on the lag such that

$$r^X(\tau) = E[x_{t+\tau} x_t^*] \tag{3.3}$$

and the Power Spectral Density (PSD) $\mathbf{s}^X$ can be defined such that $s_f^X = \text{FT}\left[r^X(\tau)\right]_f$. Then, the following proposition holds:

Proposition 1 (Auto-covariance of a w.s.s. signal) *The Fourier transform $\tilde{x}$ of x verifies:*

$$E[\tilde{x}_{f+\xi}\tilde{x}_f^*] = \delta_\xi s_f^X \tag{3.4}$$

The proof of Proposition 1 is given in Section A.1.2.

This result motivates the assumption of diagonal covariance matrix in our Gaussian framework. In practice, since the analysis is done on limited durations, there always is a windowing effect such that the Fourier transform at neighbouring frequency bins exhibits a certain correlation. Strictly speaking, the covariance matrix of the discrete Fourier transform (DFT) of our signals, considered as a random vector, is not diagonal. However, we can make this approximation which alleviates the estimation algorithms and which also provides a simple Wiener estimator (see Section 6.2.2). Neglecting the windowing effect does not seem to have many consequences in the performance of our systems. For the n^{th} frame of y, the covariance matrix is therefore assumed diagonal, with on the diagonal the PSD vectors. This diagonal is equivalently called the covariance diagonal, the spectral shape or the PSD within the present document.

$$\mathbf{\Sigma}_n^y = \text{diag}(\mathbf{s}_n^y) \tag{3.5}$$

This Gaussian assumption for musical sounds may seem far-fetched. Indeed, these audio signals are usually partly composed of sinusoids, which means that the bins of the Fourier transform corresponding to the frequencies of these sinusoids have a rather deterministic repartition on the complex plane. In other terms, in the proposed framework, the real part and imaginary part of the Fourier transform for the frequency bins corresponding to sinusoids have a deterministic relation such that their squared sum should be equal to the squared amplitude of the sinusoid, a result which is not compliant with the assumption of a mean vector equal to $\mathbf{0}$. However, in Appendix A.1.1, we recall that assuming such a Gaussian distribution on complex random variables is equivalent to assuming a Rayleigh distribution on the modulus and a uniform distribution of the phase. The mode of the Rayleigh distribution is equal to the square-root of half the variance parameter, which is non-null. The modulus of the FT, seen as a random variable, is therefore not centered, which is intuitively easier to understand for the considered audio signals.

3.3 Primary model for a "voice plus music" polyphonic signal

In the present work, the input signal, or observed signal x, is a musical sound mixture, with a clear "leading instrument" v, with a potential accompaniment part m played by some other instruments.

We assume that the STFT of the mixture signal $\mathbf{X}$ is the instantaneous mixture of the STFTs of the two contributions: the singing voice $\mathbf{V}$ and the background musical accompaniment $\mathbf{M}$:

$$\mathbf{X} = \mathbf{V} + \mathbf{M} \tag{3.6}$$

We also assume that, for each frame n, $\mathbf{v}_n$ and $\mathbf{m}_n$ are centered proper Gaussian:

$$\mathbf{v}_n \sim \mathcal{N}_c(\mathbf{0}, \text{diag}(\mathbf{s}_n^V)) \tag{3.7}$$

$$\mathbf{m}_n \sim \mathcal{N}_c(\mathbf{0}, \text{diag}(\mathbf{s}_n^M)) \tag{3.8}$$

These two vectors are also assumed independent one from the other. As the sum of two independent Gaussian vectors, the Fourier transform of the mixture at frame n, $\mathbf{x}_n$, is also a proper Gaussian vector:

$$\begin{aligned}\mathbf{x}_n &\sim \mathcal{N}_c(\mathbf{0}, \mathrm{diag}(\mathbf{s}_n^X)) \\ &\sim \mathcal{N}_c(\mathbf{0}, \mathrm{diag}(\mathbf{s}_n^V + \mathbf{s}_n^M))\end{aligned} \tag{3.9}$$

The models developed in this PhD thesis aim in particular at parameterizing the PSD vectors $\mathbf{s}_n^V$ and $\mathbf{s}_n^M$.

From Equation (3.9), one can also note that the independence assumption leads to an intuitively relevant property of sound sources. Indeed, as will be seen in Chapters 4 and 5, the problem at hand is to find a proper model for the variances $\mathbf{s}_n^V$ and $\mathbf{s}_n^M$ such that their sum approximates the power spectrum of the signal: $\mathbf{s}_n^V + \mathbf{s}_n^M \approx |\mathbf{x}_n|^2$. To comply with the positive-definiteness of the covariance matrices, we need $s_{fn}^V > 0$ and $s_{fn}^M > 0$ for all $f \in [1, F]$. This first means that these variances are homogeneous to the individual power spectra of the singing voice and of the accompaniment. Second, the power of the mixture can therefore beseen as the sum of the powers of the contributions. This in turn boils down to neglecting the *destructive* effects the respective waveforms can have one on the other.

This may of course not be realistic in certain circumstances. Let us assume that we use this statistical model to fit two mixed sinusoidal signals, with the same frequency. The resulting energy, in the frequency channel which encloses the frequency of the signals, actually also depends on the phase difference between the two signal. However, since the musical instruments that play in the mixture x are potentially played by different musicians, this phase difference is hard to model or predict. Assuming the additivity of the power spectrum within the Gaussian framework allows to implicitly take into account these phenomena: the chosen complex proper Gaussian distribution leads to a uniform distribution of the phase of the STFT (see Appendix A.1). The phase of the signals is therefore more loosely modelled, such that the case mentioned above (*i.e.* the typical overlapping partial problem) is smeared. The Gaussian distribution also allows for a certain margin around the power that is expected (the sum over all the contributions) for the actual power, which can be 0 (out of phase destructive effect) up to the sum of the powers of the contributions (the partials are in phase).

The non-negative additivity of sound sources has been widely used, especially in MIR related applications such as musical transcription [Bertin et al., 2010] or [Smaragdis and Brown, 2003], but also by works on audio source separation also assuming this additivity [Benaroya et al., 2006] or [Virtanen, 2007].

3.3.1 Graphical generative model

As stated previously, at frame n, the mixture Fourier vector $\mathbf{x}_n$ is generated by $\mathbf{v}_n$ and $\mathbf{m}_n$. The relation is deterministic, since $\mathbf{x}_n$ is the sum of $\mathbf{v}_n$ and $\mathbf{m}_n$. The corresponding graphical representation is given in Figure 3.2. This is the first description step, and only shows the instantaneous mixture assumption.

Ideally, the goal of this work is to transcribe the melody both as fundamental frequency and at musical note levels. In the graphical model, these levels should therefore also appear. This can be done by simply adding these layers to Figure 3.2. A first hidden layer of fundamental frequencies "$F_0(n)$" controls the leading voice signal Fourier tranform $\mathbf{v}_n$, and on top of this layer, the musical note hidden states "$E(n)$" controls the "F_0" layer.

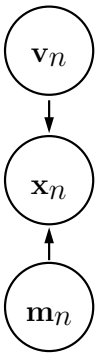


Figure 3.2: Graphical model for the observation layer, first layer dependency for the mixture. This graph is the simplest relations that are assumed between the mixture at frame n, $\mathbf{x}_n$, and the lead instrument and accompaniment, respectively $\mathbf{v}_n$ and $\mathbf{m}_n$.

To be realistic, we need to add temporal dependencies, as demonstrated on Figure 3.3. Indeed, from a physical point of view, we expect the leading voice to be played by a single instrument, which implies that the fundamental frequencies can not be an arbitrarily discontinuous "melody line", especially when the chosen time-frequency representation consists of overlapping windows. A frame and its following frame therefore share a significant amount of information. A smoothing scheme is therefore needed, that constrains the evolution of the F_0 layer from one frame to the other. In this work, we investigate a Markov chain model for this layer. The details on these temporal constraints are given in Section 3.3.3.

For the musical note layer, a more complex musicological evolution scheme is needed. The evolution is indeed not bound to the frame level, but to time durations or even higher description levels such as the tempo of the music or the note rhythms. The dependency structure therefore needs to link the state $E(n)$ at frame n with all the previous states $E(\nu)$, with $\nu < n$, as schematically shown on Figure 3.3, and further detailed in Section 3.3.4.

In the following sections we further develop the dependencies between the different variables. First, the frame level models notably exhibit the link between the fundamental frequency F_0 level and the Fourier transform level $\mathbf{v}_n$. More generally, the models for $\mathbf{v}_n$ and $\mathbf{m}_n$ are detailed, and refined graphical models are given. We then further explain the physical constraints assumed for the F_0 layer and the musicological E layer.

3.3.2 Frame level generative models

A generative (or production) model is used for both the leading voice and the accompaniment contributions. The leading voice is modelled using a source/filter model adapted from speech processing techniques, and which conveniently suits the given framework. The accompaniment model is the Gaussian composite model proposed by Benaroya et al. [2003] for audio source separation applications.

3.3.2.1 Source/filter model for the singing voice

The first purpose of the present work is to transcribe the main melody into fundamental frequencies and thereafter into musical notes. The pitched aspect of the leading or singing instrument is therefore an important cue for the proposed model. Mostly used in speech

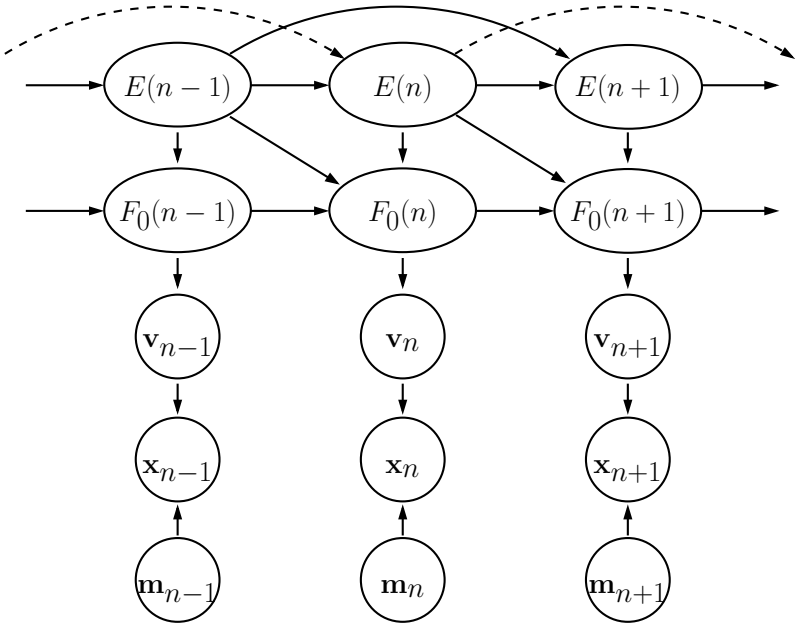


Figure 3.3: Schematic graphical model for the observation layer, with desired temporal dependencies. The mixture $\mathbf{X}$ is the sum of $\mathbf{V}$ and $\mathbf{M}$, $\mathbf{v}_n$ depends on the fundamental frequency of the lead instrument, "$F_0(n)$", which in turn depends on the note $E(n)$, but also on its previous value $F_0(n-1)$ and $E(n-1)$. At last, the note $E(n)$ depends on its whole past $E(1,\ldots,n-1)$.

Figure 3.4: Source/Filter model: the glottal source excitation e is filtered by filter g, leading in the time domain to the convolution $g * e$.

processing, the source/filter model assumes that the source part can be separated from the filter part in a vocal signal [Fant, 1970]: the source part is essentially characterized by the pitch or fundamental frequency of the signal while the filter part is linked to the global spectral envelope of the signal. In speech production, the source corresponds to the signal e emitted by the (glottal) source, and the filter component is due to the frequency response of the vocal tract, acting as a filter g for the source signal e, as shown in Figure 3.4.

e is assumed to be a wide sense stationary (w.s.s.) random process, with PSD vector $\mathbf{w}^{F_0}$. For the Fourier frequency bin f, the PSD value is then given by $w_f^{F_0}$. The resulting source/filter signal has a PSD equal to the term by term (or Hadamard) product between the PSD $\mathbf{w}^{F_0}$ of the excitation e and the power $\mathbf{w}^{\Phi}$ of the frequency response of the filter part[3]:

$$|\mathbf{g}|^2 \bullet \mathbf{w}^{F_0} = \mathbf{w}^{\Phi} \bullet \mathbf{w}^{F_0}$$

[3]The superscript F_0 (respectively Φ) will be used for elements related to the source (respectively filter) part of the leading instrument. For the source part, this emphasizes its link with the fundamental frequency.

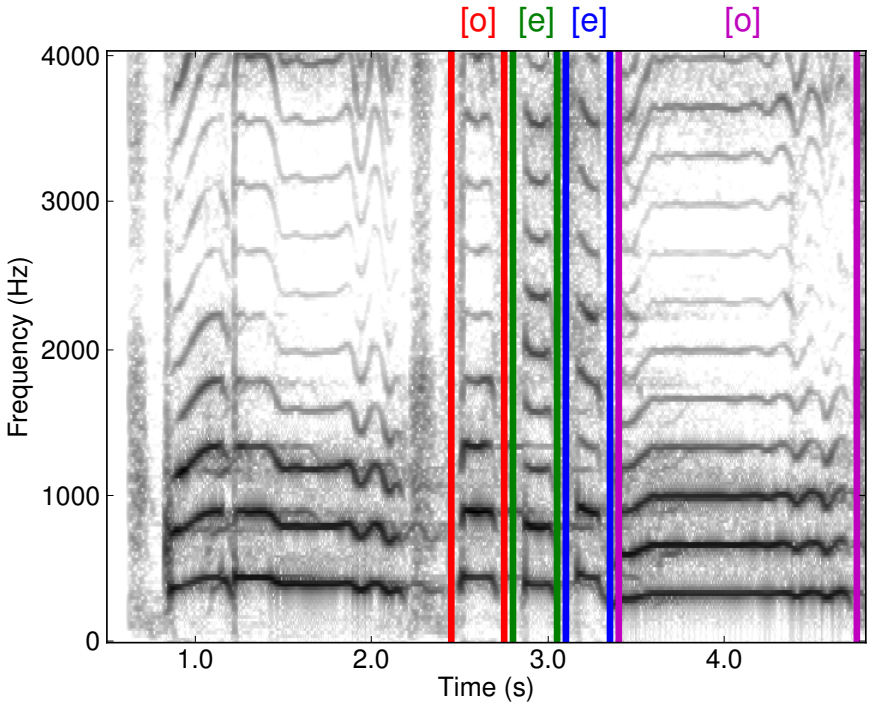


Figure 3.5: STFT of an excerpt of a song by Tamy ([Vinyes, 2008]). The fundamental frequency is easily spotted, with the comb structures of the STFT. The vowel [o] is characterized in this picture by a quite strong energy under 1500Hz, and [e] exhibits more energy than [o] in the band $[1500, 3500]$.

where $\bullet$ denotes the Hadamard product and $\mathbf{g}$ the frequency response of the filter g [4].

However, at each frame n, the source and filter components may differ from the other frames. The variability of the leading instrument both spans timbral variability as well as intonation fluctuation. Indeed, a human can sing different lyrics, different vowels and consonants, which are related to different vocal tract positions. The intonation is the ability to play or sing different pitches. These variabilities are visible on Figure 3.5, on which the labels of the sung vowels were added. The pitch line evolves independently from the spectral envelopes, since the two repeated vowels [o] and [e] are sung with different fundamental frequencies. On the other hand, one may expect sensibly the same fundamental frequencies for the first [o] and the second [e]. In order to model these types of variability, for 3 vowels and 3 notes, the GSMM model of Benaroya et al. [2006] would need $3 \times 3 = 9$ spectral shapes. To exploit both dimensions of phonation, *i.e.* the pitch and the spectral envelope, the leading voice signal can be assumed to have been generated conditionally upon two kinds of hidden states: the filter state Z^{Φ} (for the vowels) and the source state Z^{F_0} (for the fundamental frequencies). In such a framework, the number of spectral shapes therefore becomes, in total for this example, $3 + 3 = 6$. This is an important issue for practical implementations: with the source/filter modelling, the number of necessary spectral shapes can be reduced while still including a rather wide range of possible spectra.

[4] Note that the notation $\mathbf{g}$ was introduced mainly to highlight the link between the time domain filter and the parameter $\mathbf{w}^{\Phi}$. This frequency response will not be used as such in the remainder of this document.

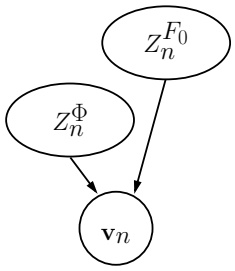


Figure 3.6: Graphical model for the frame-wise model generating the leading voice signal. The lead instrument signal $\mathbf{v}_n$ is generated by two hidden states, Z^{Φ} the filter part state, and Z^{F_0} the hidden state for the fundamental frequency.

Figure 3.6 shows the dependency graph of the leading voice model for a single frame. The states Z^{Φ} and Z^{F_0} are independent one from the other. The temporal evolution of these states is further studied in Section 3.3.3. Conditionally upon these states, the vectors $\mathbf{v}_n$ are independent from the values of the neighbouring vectors.

The filter part is assumed to be restricted to a limited range of possible spectral shapes. Let K be the number of these shapes. The k^{th} shape is denoted $\mathbf{w}_k^{\Phi}$ and all the K vectors form a $F \times K$ matrix $\mathbf{W}^{\Phi}$. Z_n^{Φ} can therefore take values in $[1, K]$. The matrix $\mathbf{W}^{\Phi}$ could be either learnt from the processed signal (unsupervised framework) or learnt from a dataset so as to catch a specific type of leading instrument, such as a male singer (supervised case). The first unsupervised framework was chosen in our work, since we defined the melody as being played by any possible instrument, provided it is harmonic and generally follows a source/filter production model.

Since we are mainly interested in the pitched content of the lead instrument, the voiced sections of the source parts are modelled in priority. The source part is therefore characterized by the fundamental frequency f_0 of the corresponding generated signal. Let U be the total number of allowed fundamental frequencies. $Z_n^{F_0}$ takes values in $[1, U]$. For $u \in [1, U]$, a mapping $\mathcal{F}(u)$ is defined from $[1, U]$ to a given set of fundamental frequencies. This set can typically span a more or less dense range of frequencies between a minimum fundamental frequency and a maximum frequency. The u^{th} source spectrum is denoted $\mathbf{w}_u^{F_0}$, all the spectra forming a $F \times U$ dictionary $\mathbf{W}^{F_0}$. This dictionary is fixed before the estimation. Modifying the spectra in $\mathbf{W}^{F_0}$ may also modify the actual fundamental frequency of the corresponding time domain signal, that is why one should avoid re-adjusting them. It was also decided to fix this dictionary instead of estimating the potential F0s, such that we do not need to compute these spectra at each estimation.

Figures 3.7 and 3.8 respectively display the time domain source signal (or glottal flow, top pane) and the corresponding power spectrum (bottom pane) for two different fundamental frequencies, using the chosen KLGLOTT88 glottal source model [Klatt and Klatt, 1990]: $f_0 = 183$Hz, and $f_0 = 1210$Hz. Details on the KLGLOTT88 model can be found in Appendix C. On Figure 3.9, a fixed spectral comb dictionary is represented. The fundamental frequency range is, in Hz, $[100, 800]$. Here, the mapping $\mathcal{F}$ verifies:

$$\mathcal{F}(u) = f_0^{\min} \times 2^{\frac{u-1}{12U_{\text{st}}}}, \forall u \in [1, U] \tag{3.10}$$

where $f_0^{\min} = 100$Hz and $U_{\text{st}} = 4$ is the number of elements in $\mathbf{W}^{F_0}$ whose fundamental frequencies are within one semitone. Using the formula (3.10), the fundamental frequency

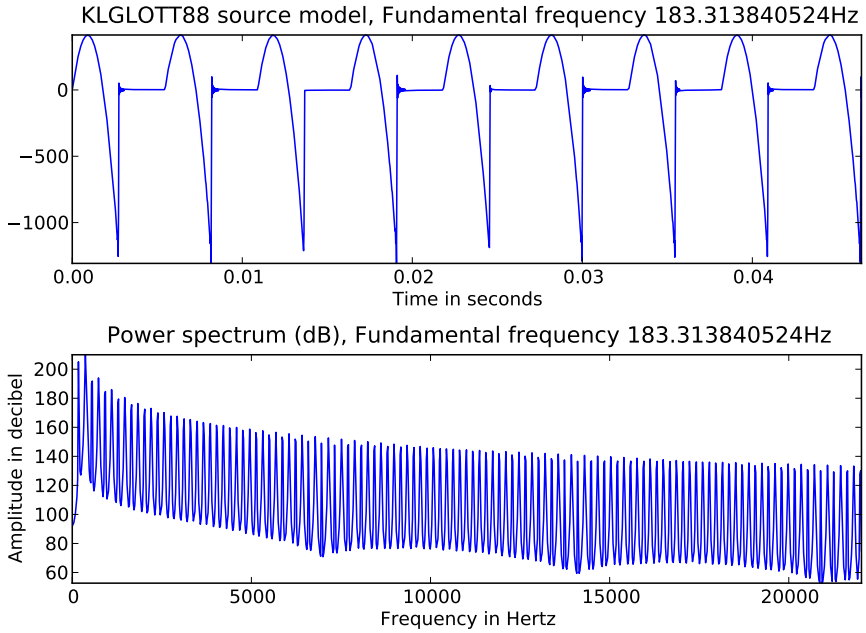


Figure 3.7: Source signal and corresponding "spectral comb" generated by the KL-GLOTT88 model, $f_0 \approx 183$Hz.

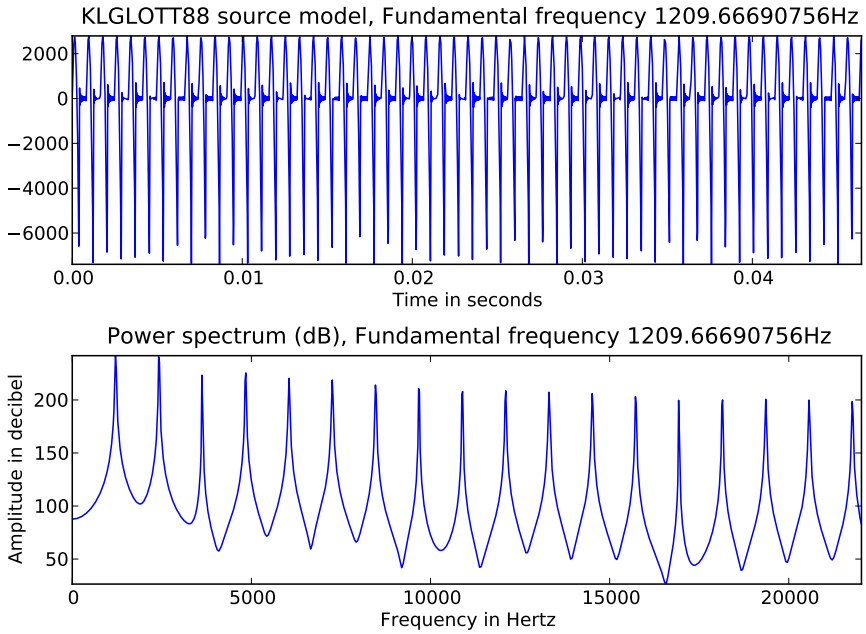


Figure 3.8: Source signal and corresponding "spectral comb" generated by the KL-GLOTT88 model, $f_0 \approx 1210$Hz.

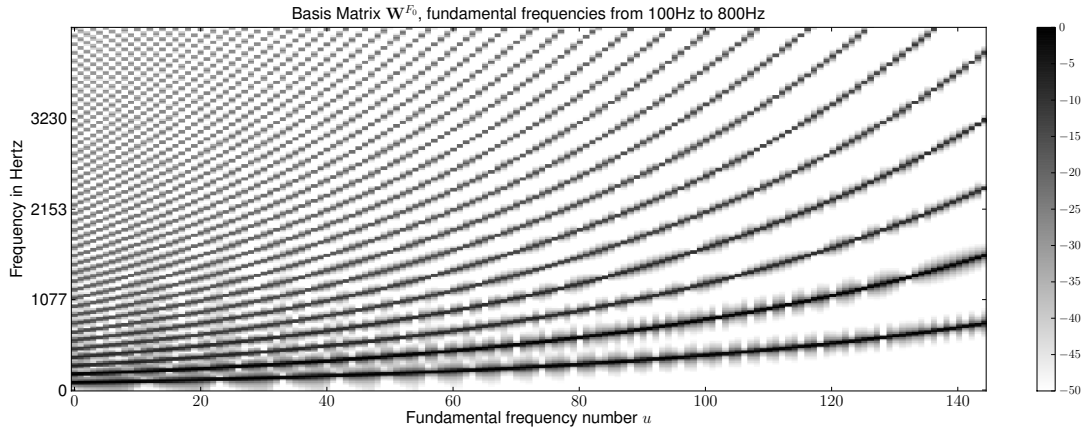


Figure 3.9: Dictionary matrix $\mathbf{W}^{F_0}$. Darker colors correspond to higher energies, in dB. The fundamental frequencies range from 100Hz to 800Hz. The number of elements per semi-tone, U_{st}, should be high enough such that high frequency lobes slightly overlap from one element to the other, in order to be able to fit the signal, and more concretely follow partials throught the frames.

range thus corresponds to a subdivision of the Western musical scale. In this study, several parameters were tested in various situations, with values for U_{st} varying from 4 to 16. A value of 4 usually gives enough spectral combs providing very good results in F0 estimation, as shown in Section 6.1.1. However, for separation tasks, it may be more insteresting to use higher values, like 8 or even 16.

Following [Benaroya et al., 2006], the leading instrument is modelled as a **Gaussian Scaled Mixture Model (GSMM)**. However, the states are in our framework the state couples $Z_n = (k, u) \in [1, K] \times [1, U]$. With the Gaussian assumption, the STFT of V is assumed to follow the conditional density, at frame n:

$$\mathbf{v}_n | \{Z_n = (k, u)\} \sim \mathcal{N}_c(0, b_{kun} \text{diag}(\mathbf{w}_k^{\Phi} \bullet \mathbf{w}_u^{F_0})) \tag{3.11}$$

where b_{kun} is the amplitude coefficient corresponding to state (k, u). The vectors $\mathbf{w}_k^{\Phi}$ and $\mathbf{w}_u^{F_0}$ are both normalized, $\forall u$ and k, in order to avoid indeterminacies when estimating the necessary parameters. This implies the use of these amplitude factors which allow to fit the normalized spectral shapes to the power STFT of the signal.

The observation likelihood is the weighted mixture of all the conditional probabilities:

$$p(\mathbf{v}_n) = \sum_{k,u} \pi_{ku} p(\mathbf{v}_n | Z_n = (k, u))$$

where π_{ku} is the *prior* probability of state (k, u). The above equation is then equivalent to the following convention:

$$\mathbf{v}_n \sim \sum_{k,u} \pi_{ku} \mathcal{N}_c(0, b_{kun} \text{diag}(\mathbf{w}_k^{\Phi} \bullet \mathbf{w}_u^{F_0}))$$

The generative process of the GSMM is schematically drawn on Figure 3.10. This model can thereafter be interpreted as follows: for frame n, each source u is filtered by

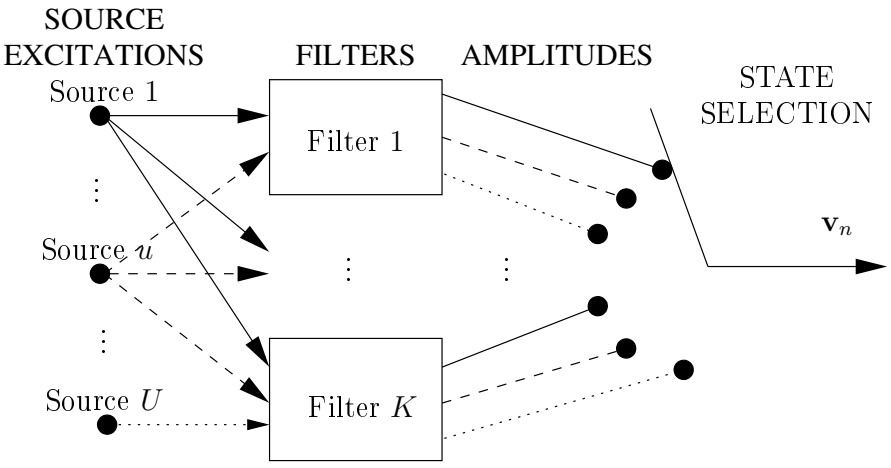


Figure 3.10: Schematic principle of the generative GSMM for the main instrument part. Each source u is filtered by each filter k. For frame n, the signal is then multiplied by a given amplitude and a "state selector" then chooses the active state.

each filter k. The result is multiplied by the amplitude coefficient $\sqrt{b_{kun}}$ [5]. At last, the selector symbolizes the process of drawing the active pair (k, u) with the *prior* probabilities π_{ku}. The signal of the chosen source/filter combination constitutes the lead instrument signal to be added to the accompaniment contribution.

The mixture likelihood in Equation (3.9) can be re-written to take into account the leading voice dependencies to the states Z:

$$\mathbf{x}_n \sim \sum_{k,u} \pi_{ku} \mathcal{N}_c(\mathbf{0}, \mathrm{diag}(b_{kun} \mathbf{w}_k^{\Phi} \bullet \mathbf{w}_u^{F_0} + \mathbf{s}_n^M)) \tag{3.12}$$

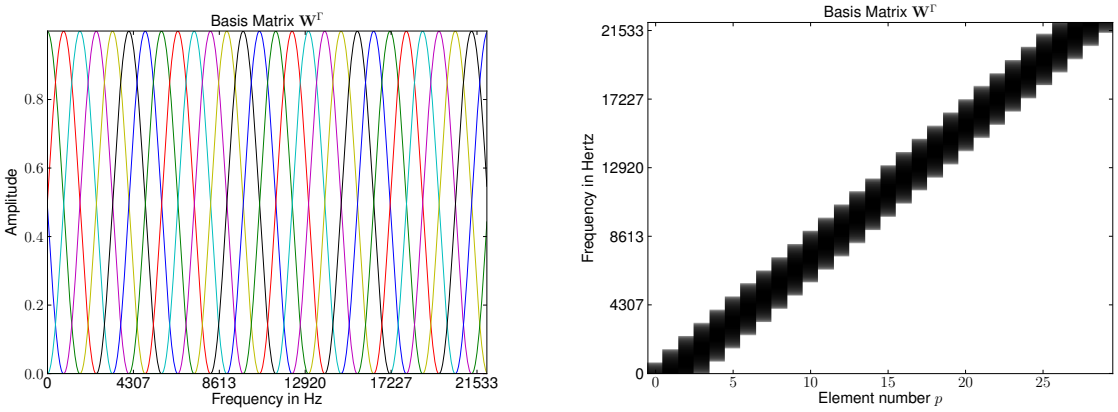
We denote the leading instrument's variance at frame n, conditionally upon the state $Z_n = (k, u)$, $\mathbf{s}_n^{V,\mathrm{GSMM}|ku} = b_{kun} \mathbf{w}_k^{\Phi} \bullet \mathbf{w}_u^{F_0}$, such that the likelihood of the observation conditionally upon the state (k, u) is given by:

$$\mathbf{x}_n | \{Z_n = (k, u)\} \sim \mathcal{N}_c(\mathbf{0}, \mathrm{diag}(\mathbf{s}_n^{V,\mathrm{GSMM}|ku} + \mathbf{s}_n^M))$$

Let $\mathbf{B}$ be the $K \times U \times N$ tensor whose entries are b_{kun}, and $\mathbf{\Theta}^{V,\mathrm{GSMM}} = \{\mathbf{B}, \mathbf{W}^{\Phi}\}$ the set of parameters that needs to be estimated from the observed signal, for the proposed leading voice GSMM.

In the proposed framework, an estimation of the filter matrix $\mathbf{W}^{\Phi}$ directly from the data is desired. However, without any constraint on these spectral envelopes, the interpretation of the obtained matrix may be impossible, and later use on other application worthless. An alternative to this drawback is to introduce a smoothness constraint on the filters in the GSMM, obtaining a new parameterization for the model, then denoted **Smooth filters-GSMM (SGSMM)**. For a given $k \in [1, K]$, $\mathbf{w}_k^{\Phi}$ is assumed to be a non-negative linear combination of smooth spectral shapes $\mathbf{w}_p^{\Gamma}$, $p \in [1, P]$, which form a dictionary matrix $\mathbf{W}^{\Gamma}$. The chosen dictionary is constituted of P several overlapping Hann windows, as shown on Figure 3.11(a). As can be seen on Figure 3.11(b), such a family of functions is a collection of band-pass filters. One could also use other types of filters such as filters whose frequency

[5] The amplitude coefficient b_{kun} is applied in the power spectrum domain, hence the square root for the time domain.



(a) Dictionary of overlapping Hann functions.

(b) "Matrix" view.

Figure 3.11: Dictionary matrix $\mathbf{W}^{\Gamma}$. To enforce a smooth structure to the elements of filter matrix $\mathbf{W}^{\Phi}$, they are modelled as combinations of the $P = 30$ smooth elements of $\mathbf{W}^{\Gamma}$.

supports are "uniformly spaced subbands on the Equivalent Rectangular Bandwidth (ERB) scale" [Vincent et al., 2008]. An example of a combination of the elements of $\mathbf{W}^{\Gamma}$ is given on Figure 3.12.

The spectral shape matrix $\mathbf{W}^{\Phi}$ is decomposed onto $\mathbf{W}^{\Gamma}$ such that $\mathbf{W}^{\Phi} = \mathbf{W}^{\Gamma}\mathbf{H}^{\Gamma}$, where $\mathbf{H}^{\Gamma}$ is the $P \times K$ (non-negative) amplitude coefficient matrix: for $k \in [1, K]$,

$$\mathbf{w}_k^{\Phi} = \mathbf{W}^{\Gamma}\mathbf{h}_k^{\Gamma} = \sum_p \mathbf{w}_p^{\Gamma} h_{pk}^{\Gamma} \tag{3.13}$$

The set of parameters corresponding to the SGSMM is defined as $\boldsymbol{\Theta}^{V,\mathrm{SGSMM}} = \{\mathbf{B}, \mathbf{H}^{\Gamma}\}$. Note that $\mathbf{W}^{\Gamma}$ does not belong to $\boldsymbol{\Theta}^{V,\mathrm{SGSMM}}$, as it is fixed and not estimated. The number P of elements in $\mathbf{W}^{\Gamma}$ allows to control the regularity of the obtained filters.

At last, for the source separation purpose, an extra element can be inserted in $\mathbf{W}^{F_0}$ in order to model unvoiced parts of the lead instrument. This is mainly used in the source separation system presented in [Durrieu et al., 2009b], which is further detailed in Section 6.2.4. This new basis element $\mathbf{w}_{U+1}^{F_0}$ is set to a uniform value for all the frequencies: it then models the source part for unvoiced sounds as if it were some white noise. It seems better to estimate this unvoiced part in an additional round of parameter estimation, as explained in Section 6.2.4, in order to avoid catching too many other "noisy" components, e.g. drums, which do not correspond to the main (melodic) instrument. Once the filters corresponding to the voiced part of the lead instrument are well estimated, then a new round of estimation including the unvoiced element $\mathbf{w}_{U+1}^{F_0}$ can be done, potentially also avoiding to change the previously estimated filters. The underlying assumption is that the voiced and unvoiced parts of the lead instrument are generated by the same source/filter process, with the same filters.

The parameters involved in these models, the GSMM and the SGSMM, are summarized in Table 3.1, in Section 3.3.2.3. The only difference between the GSMM and SGSMM concerns the filter parameterization which leads to merely minor changes in the estimation theory derived later. In the proposed algorithms, this amounts to choosing to directly

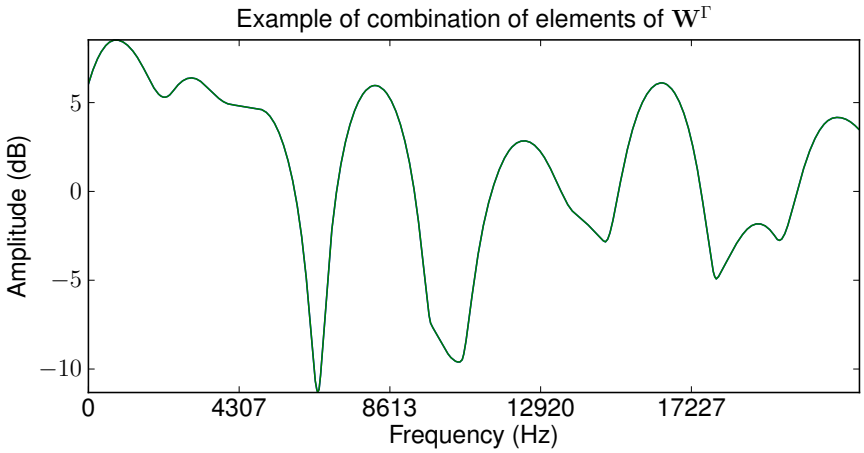


Figure 3.12: An example of a combination of elements of $\mathbf{W}^{\Gamma}$, $P = 30$. The higher P, the less smooth the spectral shape.

estimate $\mathbf{W}^{\Phi}$ (GSMM) or the matrix $\mathbf{H}^{\Gamma}$ (SGSMM). The estimations for both matrices are given in the related algorithms.

3.3.2.2 Instantaneous mixture for the accompaniment

The musical background is assumed to be composed of a wide variety of instruments, such as a guitar, a bass guitar, a piano or drums. In order to take this variability into account, the chosen model is less constrained than the one used for the leading voice.

To this effect, the model proposed by Benaroya et al. [2003] particularly fits our application. Indeed, in contrast with the GSMM proposed by the same authors [Benaroya et al., 2006], the accompaniment is assumed to be the instantaneous mixture of several components, characterized by their spectral shapes.

The vector $\mathbf{m}_n$ of the Fourier transform of the accompaniment part at frame n is the sum of R mutually independent component vectors $\mathbf{m}_n^r$, with $r \in [1, R]$: $\mathbf{m}_n = \sum_r \mathbf{m}_n^r$. For a given component number r, the spectral shape is always the same, $\mathbf{w}_r^M$, across all the frames of the STFT. An amplitude factor allows to adapt the (normalized) energy of $\mathbf{w}_r^M$ to the actual energy for each frame n: $h_{rn}^M > 0$. $\mathbf{m}_n$ therefore follows, for a given frame n:

$$\mathbf{m}_n^r \sim \mathcal{N}_c(0, h_{rn}^M \mathrm{diag}(\mathbf{w}_r^M)) \tag{3.14}$$

As the sum of the Gaussians, $\mathbf{m}_n$ is also Gaussian such that:

$$\begin{aligned} \mathbf{m}_n &\sim \mathcal{N}_c(0, \sum_r \mathrm{diag}(h_{rn}^M \mathbf{w}_r^M)) \\ \mathbf{m}_n &\sim \mathcal{N}_c(0, \mathrm{diag}(\mathbf{W}^M \mathbf{h}_n^M)) \end{aligned} \tag{3.15}$$

The set of parameters to be estimated for the accompaniment part is denoted $\mathbf{\Theta}^M = \{\mathbf{W}^M, \mathbf{H}^M\}$.

This model has the advantage of being rather realistic, because it allows several sources to be active at the same time. This tends to better reflect the real world polyphonic situation. The alternative GSMM framework [Benaroya et al., 2006] is less suited to the task at hand, since it assumes that at each frame, there is only one component active, corresponding to a situation where the instruments of the background are playing one after the other. The instantaneous mixture assumption is more flexible and also leads to faster estimation schemes, ultimately equivalent to a certain kind of **Non-negative Matrix Factorisation (NMF)** problem as developed in Chapter 4.

Why does the model adopted for the accompaniment seem much simpler than for the leading voice when the accompaniment is assumed more complex? This is actually not so surprising: to a certain extent, a simple model also means a flexible model. On the contrary, the more complicated the model, the more constrained it is. The more we know or specify what the target is, the more constrained, hence complicated, the model is. The lead instrument is defined as being monophonic, harmonic and following a source/filter production process, its model is therefore rather sophisticated. The accompaniment is not as well identified as the desired lead voice, hence a more flexible model, or, in other words, a simpler model, is needed.

3.3.2.3 Frame level model for the mixture: summary

At last, the frame-wise model of signal leads to the following expression of the likelihood, for the mixture, conditionally upon the leading voice state $Z_n = (k, u)$:

$$\mathbf{x}_n|k,u \sim \mathcal{N}_c(\mathbf{0}, \mathrm{diag}(b_{kun}\mathbf{w}_k^{\Phi} \bullet \mathbf{w}_u^{F_0} + \mathbf{W}^M\mathbf{h}_n^M)) \tag{3.16}$$

In the above expression, we omitted to recall that it is the leading voice state Z_n which is at state (k, u). In the remainder of this document, when there is no ambiguity, this simpler expression will be used. Additionally, $p(.|Z_n = (k, u))$ will also be replaced by $p(.|k, u)$ when applicable. At last, the variance vector for the mixture, conditionally upon the source/filter state (k, u) is denoted $\mathbf{s}_n^{\mathrm{GSMM}|ku}$ and verifies:

$$\mathbf{s}_n^{\mathrm{GSMM}|ku} = b_{kun}\mathbf{w}_k^{\Phi} \bullet \mathbf{w}_u^{F_0} + \mathbf{W}^M\mathbf{h}_n^M \tag{3.17}$$

One additional state can be included using this framework, namely **the silence or "rest" state**, for the **"non-voiced"** frames for the lead instrument. Indeed, it might have been awkward to set to 0 the variance on rest frame on the sole lead instrument model. Although it comes naturally, considering the interpretation of the variances, it would have led to some undefined equation, with divisions by zeros. Instead, with the mixture model, the lead instrument silence strategy using a variance equal to 0 makes sense and leads to:

$$\mathbf{s}_n^{\mathrm{GSMM}|non-voiced} = \mathbf{0} + \mathbf{W}^M\mathbf{h}_n^M \tag{3.18}$$

This new state is easily included in the GSMM framework. To keep it simple, from here on, this silence state is implicitly included in the framework, even though it may not appear explicitly in all equations, mainly for the sake of readability.

The parameters involved in the (S)GSMM as well as the variances they define are summarized in Table 3.1.

Table 3.1: (S)GSMM: Parameters for the leading voice and the accompaniment. All the parameters are estimated, except if mentioned. If a parameter is used only in the GSMM or the SGSMM framework, and not in both, then this is also indicated.

	Description	Remarks
	Mixture $\mathbf{s}_n^{\text{GSMM}\|ku} = \mathbf{s}_n^{V,\text{GSMM}\|ku} + \mathbf{s}_n^M$	
	Leading Instrument $\mathbf{s}_n^{V,\text{GSMM}\|ku} = b_{kun}\mathbf{w}_k^\Phi \bullet \mathbf{w}_u^{F_0}$	
$\mathbf{W}^\Phi$	Matrix of spectral envelopes for the filter part	GSMM
$\mathbf{w}_k^\Phi$	Vector of filter spectral envelope k	GSMM
w_{fk}^Φ	Filter spectral envelope k, at frequency bin f	GSMM
$\mathbf{W}^{F_0}$	Dictionary of source comb spectra	Fixed
$\mathbf{w}_u^{F_0}$	Source comb spectrum u	Fixed
$w_{fu}^{F_0}$	Source comb spectrum u at frequency bin f	Fixed
$\mathbf{W}^\Gamma$	Dictionary of smooth elementary filter parts	SGSMM, Fixed
$\mathbf{w}_p^\Gamma$	Vector of smooth elementary filter p	SGSMM, Fixed
w_{fp}^Γ	Smooth elementary filter p at frequency bin f	SGSMM, Fixed
$\mathbf{B}$	Amplitude tensor	
b_{kun}	Amplitude for the couple $Z^\Phi = k, Z^{F_0} = u$, at frame n	
$\mathbf{H}^\Gamma$	Amplitude matrix for the decomposition of $\mathbf{W}^\Phi$ on $\mathbf{W}^\Gamma$	SGSMM
$\mathbf{h}_k^\Gamma$	Amplitude vector for filter k	SGSMM
h_{pk}^Γ	Amplitude for filter k on element p	SGSMM
	Accompaniment $\mathbf{s}_n^M = \mathbf{W}^M\mathbf{h}_n^M$	
$\mathbf{W}^M$	Matrix of spectral shape for the accompaniment	
$\mathbf{w}_r^M$	Spectral shape for element r of $\mathbf{W}^M$	
w_{fr}^M	Spectral shape r, at frequency bin f	
$\mathbf{H}^M$	Matrix of amplitudes for the accompaniment	
$\mathbf{h}_n^M$	Vector of amplitudes for the accompaniment at frame n	
h_{rn}^M	Amplitude associated with element r of $\mathbf{W}^M$ at frame n	

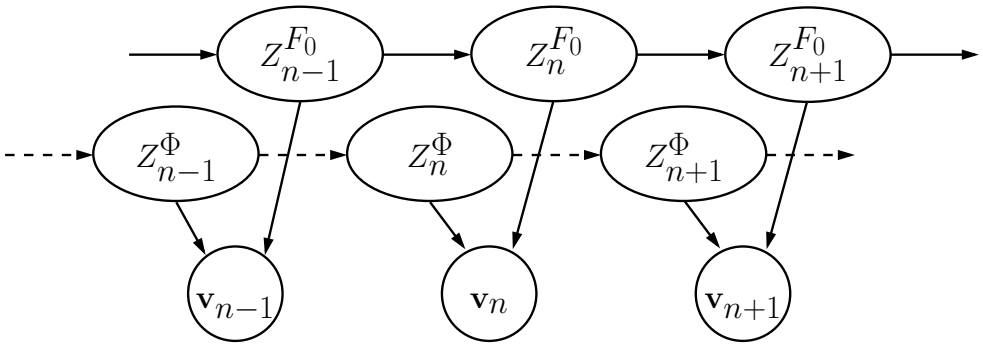


Figure 3.13: Graphical model of the generating HMM for the leading voice signal. The filter state sequence Z^{Φ} and the source state sequence Z^{F_0} are HMM sequences, with dependencies to the previous frame only.

3.3.3 Physical state layer: constraining the fundamental frequency evolution of the singing voice

Figure 3.6 shows the structure of the frame-wise model, but from a realistic generative point of view, there is a significant lack in the temporal dependencies, namely the temporal evolution of the states. Indeed, one should expect the state sequence $Z = \{Z_n, n \in [1, N]\}$ to exhibit some structure: $Z^{F_0}_n$ is physically related to the previous frame state $Z^{F_0}_{n-1}$, if we consider that the time-frequency representation corresponds to a "natural" (or real) instrument. The source part, and more precisely the pitch of the lead instrument cannot have a "random" evolution, especially when the time-frequency representation uses overlapping windows, meaning that the pitch varies slowly with respect to the frame sequence. Similarly, the filter state sequence Z^{Φ} corresponds to the evolution of the timbral changes, such as the vocal tract frequency response evolution. This sequence must also exhibit a certain degree of smoothness, since modifications of physical systems are not instantaneous.

In the proposed GSMM, a natural way to model these temporal evolutions is to consider a Markov model for each of the state sequences, leading to a hidden Markov model (HMM)for the leading voice, as depicted on Figure 3.13. Both sequences are assumed mutually independent: in the source/filter model, for a singer, the pitch and the timbre are respectively controlled by the vocal chords and the vocal tract, which are two distinct physical systems. The evolutions of Z^{Φ} and Z^{F_0} can therefore be considered independent one from the other.

For the source part, the HMM is defined with the following components:

- **Observation sequence**: the mixture STFT matrix $\mathbf{X}$.
- **The hidden state sequence** Z^{F_0}, where at each frame n, $Z^{F_0}_n = u$ represents a given pitch.
- **The *prior* probabilities** of $u \in [1, U]$, π_u.
- **The transition probabilities**, from pitch number u to v, $\forall (u, v) \in [1, U]^2$, characterized by a cost function favoring smooth pitch transitions rather than jumps, $Q(u, v) = p(Z^{F_0}_n = v | Z^{F_0}_{n-1} = u)$.

The parameters such as the *prior* probabilities and the transition probabilities may be learnt on a database of isolated singers. In the proposed works, these parameters were set in advance instead, based on some knowledge on the nature of the signal. Indeed, we are interested in any sort of leading voice, be it male or female singer, or any other wind instrument for instance. For this reason, there is no preference needed as concerns the fundamental frequency range, and uniform priors are satisfying in this case: $\pi_u \propto 1$.

We primarily target music excerpts belonging to the Western style. We can therefore expect that a certain stability and, more importantly, transitions depending on the differences of fundamental frequencies on a logarithmic frequency scale must be encoded in the transition probabilities. Similarly, there may not be much information as whether the behaviour of the pitch in high frequencies should be different from the one in low frequencies, and the transition probabilities should therefore only depend on the differences between pitches on a logarithmic frequency scale:

$$Q(u,v) = q(\delta), \text{ where } \delta = \log_2 \mathcal{F}(v) - \log_2 \mathcal{F}(u)$$

q is a function from $\mathbb{R}$ to $\mathbb{R}^+$. In order to favor continuous melody lines, q should exhibit a global maximum at 0. Without specific knowledge, this function should also be symmetric, such that $Q(u,v) = Q(v,u)$. For instance, in our work, q is defined as:

$$\begin{aligned} q(\delta) &\propto \exp(\alpha.\text{round}(|12\delta|)) \\ q(\log_2(f_2) - \log_2(f_1)) &\propto \exp(\alpha.\text{round}(|12\log_2(f_1) - 12\log_2(f_2)|)) \end{aligned} \tag{3.19}$$

where the parameter α controls the smoothness degree of the melody line: the higher α, the more constant (or "horizontal") the melody line. The term $12\log_2(f)$ maps the frequency f on the Western musical scale, such that $|12\log_2(f_1) - 12\log_2(f_2)|$ is the difference between f_1 and f_2 expressed in semitones.

Similarly, **the sequence Z^Φ of the filter part** can be modelled thanks to a Markov model, although to many aspects, it would be interesting to consider a model close to the duration model developed in Section 3.3.4. For the purpose of this thesis, only the HMM framework has been investigated, since the focus was not on analyzing the resulting sequence of filter states.

To **summarize**, we obtain a double Markov chain, with two hidden state sequences Z^Φ and Z^{F_0}. This is equivalent to a single HMM chain, with hidden state sequence $Z = (Z^\Phi, Z^{F_0})$:

- **Observation sequence**: the mixture STFT matrix $\mathbf{X}$.
- **The hidden state sequence** Z, where at each frame n, $Z_n = (k,u)$.
- **The *prior* probabilities** of $(k,u) \in [1,K] \times [1,U]$, π_{ku}.
- **The transition probabilities**, from (k,u) to (l,v), $\forall k,l,u,v \in [1,K]^2 \times [1,U]^2$, $Q((k,u),(l,v)) = Q^\Phi(k,l)Q^{F_0}(u,v)$.

For this HMM chain, the transition probabilities between the frequencies are defined as above, with the function in Equation (3.19). For the transitions of the filter part, to simulate a stable evolution, the following definition can be used:

$$Q^\Phi(k,l) \propto \begin{cases} 1, \text{ if } k = l \\ \epsilon^\Phi << 1, \text{ if } k \neq l \end{cases} \tag{3.20}$$

For the fundamental frequency transcription task, further introduced in Section 6.1.1, the Maximum *A Posteriori* (MAP) sequence $\widehat{Z}^{F_0}$ is the sequence that maximizes the posterior probability $p(Z^{F_0}|\mathbf{X})$. Thanks to Bayes' law, we can write:

$$\begin{aligned}\widehat{Z}^{F_0} &= \underset{z^{F_0}\in[1,U]^N}{\arg\max}\ p(\mathbf{X}|Z^{F_0}=z^{F_0})p(Z^{F_0}=z^{F_0}) \\ &= \underset{z^{F_0}\in[1,U]^N}{\arg\max} \sum_{z^{\Phi}\in[1,K]^N} p(\mathbf{X}|Z^{\Phi}=z^{\Phi},Z^{F_0}=z^{F_0})p(Z^{\Phi}=z^{\Phi},Z^{F_0}=z^{F_0})\end{aligned}$$

To simplify the tracking of the desired sequence, it is also possible to adopt a less optimal solution which consists in estimating the best path for both sequences $\widehat{Z}=(\widehat{Z}^{\Phi},\widehat{Z}^{F_0})$ such that:

$$\widehat{Z}^{\Phi},\widehat{Z}^{F_0} = \underset{z^{F_0}\in[1,U]^N,z^{\Phi}\in[1,K]^N}{\arg\max} p(\mathbf{X}|Z^{\Phi}=z^{\Phi},Z^{F_0}=z^{F_0})p(Z^{\Phi}=z^{\Phi},Z^{F_0}=z^{F_0}) \quad (3.21)$$

This sequence can be computed using for instance the efficient Viterbi algorithm [Rabiner, 1989], further discussed in Section 5.4.1.

The sequence of fundamental frequencies $\widehat{Z}^{F_0}$ gives one frequency per frame. It is worth noticing that this state sequence corresponds to a physically relevant parameter for the signal: it gives the actually played/sung fundamental frequency. In a source separation framework, this state can greatly help in better isolating the corresponding source. However, in a transcription, especially for musical score estimation, the physical fundamental frequency line does not directly allow to conclude on which musical note was intended. In order to infer this higher level state, another hidden state layer can be added.

3.3.4 "Musicological" state layer to model note level duration

The note level previously sketched on Figure 3.3 can be defined on top of the frame-wise fundamental frequency states Z^{F_0}. The note level structure is depicted on Figure 3.14: for each frame n, there is one note played by the leading voice, E_n. This state is connected to the whole past E_ν, $\nu<n$, because a frame-to-frame dependency is not enough to describe high level semantics such as musical notes, as we will discuss later. The E layer is connected to the fundamental frequency layer such that a note E_n at frame n controls the emitted frequency state $Z_n^{F_0}$. However, this dependency may need to be further extended such that $Z_n^{F_0}$ depends on $Z_{n-1}^{F_0}$, for the physical consistency, E_n for the link with the current note and also E_{n-1}, especially to control stronger and weaker variations of the f_0 line. Indeed, this sequence may achieve greater jumps when changing notes ($E_n\neq E_{n-1}$), while staying quite steady within one note ($E_n=E_{n-1}$). All these dependencies are further detailed in this section.

The layers for Z^{F_0} and E seem redundant, but they are actually complementary: E is the sequence of musical notes, quantified on the Western musical scale, while Z^{F_0} is the sequence of corresponding F0's. With such a double layer model, a musical transcription into a musical score is possible, thanks to E, while Z^{F_0} allows to describe more accurately the signal. These layers may also not have a deterministic relationship, especially in presence of vibrato singing, where the F0 line is not constant, while the note E stays the same. Dealing with such phenomena is one of the main targets of our model.

In [Vincent, 2004], explicit note duration models are proposed. These models are however limited to small variations of the spectral shapes, and therefore only small variations of the pitch. In order to take into account more variable signals, especially human voice

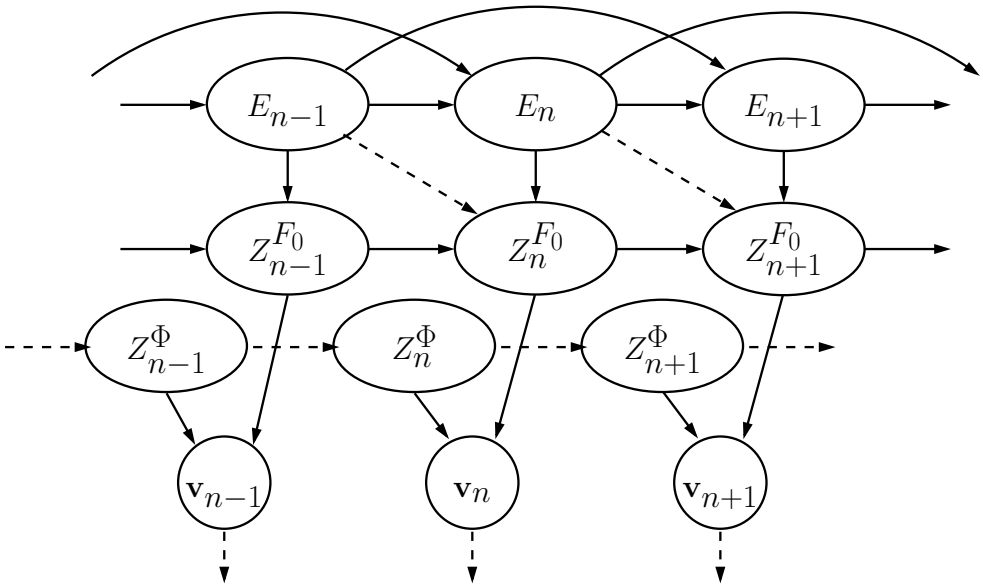


Figure 3.14: Generative model for the leading voice. The graph now includes all the hidden states of the proposed model. The observation $\mathbf{v}_n$ is generated by the source/filter states $Z^{F_0}_n$ and Z^{Φ}_n. These states depend on their value at the previous frame (HMM structure), and the sequence Z^{F_0} also depends on the note sequence E. At last, E_n depends on its whole past $E_{1:n-1}$

and vibrato, we adapted the segmental model such that a wider range of variations is allowed.

Let us first write the joint probability of the observations and all the state layers as depicted in Figure 3.14, with the independence assumptions made above:

$$p(\mathbf{X}, Z^{\Phi}, Z^{F_0}, E) = \underbrace{p(\mathbf{X}|Z^{\Phi}, Z^{F_0})}_{\text{“Frame-wise” model}} \times \overbrace{p(Z^{\Phi})}^{\text{Filter evolution}} \times \underbrace{p(Z^{F_0}|E)}_{\text{Physical source evolution}} \times \overbrace{p(E)}^{\text{Musicological model}} \tag{3.22}$$

where the different contributions write:

$$p(\mathbf{X}|Z^{\Phi}, Z^{F_0}) = \prod_n p(\mathbf{x}_n|Z^{\Phi}_n, Z^{F_0}_n) \tag{3.23}$$

$$p(Z^{\Phi}) = p(Z^{\Phi}_1) \prod_n p(Z^{\Phi}_n|Z^{\Phi}_{n-1}) \tag{3.24}$$

$$p(Z^{F_0}|E) = p(Z^{F_0}_1|E_1) \prod_n p(Z^{F_0}_n|Z^{F_0}_{n-1}, E_n, E_{n-1}) \tag{3.25}$$

$$p(E_{1:n}) = p(E_n|E_{1:n-1})p(E_{1:n-1}), \forall n \in [1, N] \tag{3.26}$$

The first two equations (3.23) and (3.24) have already been discussed respectively in Sections 3.3.2 and 3.3.3. The third equation (3.25) expresses the link between the physical layer, especially the fundamental frequency sequence, and the musicological layer, embodied by the note sequence E. This link will be discussed first in this section. At last, in the

fourth equation (3.26), the principle that will be used later in the actual estimation of the desired sequence is shown: we can compute the joint likelihood of the whole signal up to frame n with the joint likelihood of the signal up to frame $n-1$.

In Equation (3.25), **the conditional probability** $p(Z_n^{F_0}|Z_{n-1}^{F_0}, E_n, E_{n-1})$, for a given frame n, expresses how the evolution of the notes and the fundamental frequency of the previous frame constrain the fundamental frequency of the current frame. There are typically three behaviours of the sequence Z^{F_0}, depending on the sequence E. First, given the note E_n (in MIDI code) at frame n, the corresponding fundamental frequency $\mathcal{F}(Z_n^{F_0})$ should be "close" to the standard frequency f_0^{MIDI} such that, for a given MIDI code $E_n = n^{\text{MIDI}}$:

$$f_0^{\text{MIDI}} = \mathcal{F}^{\text{MIDI}}(n^{\text{MIDI}}) = 440 \times 2^{\frac{n^{\text{MIDI}}-69}{12}} \tag{3.27}$$

where the classical tuning 440Hz for the note A4, MIDI code 69, is assumed. Second, if the note state at frame $n-1$ and frame n are the same, $E_n = E_{n-1}$, then there should be a physical continuity between $Z_n^{F_0}$ and $Z_{n-1}^{F_0}$, such as the regularity imposed in the HMM structure of Section 3.3.3. At last, if $E_n \neq E_{n-1}$, this regularity is not needed. However, instead of considering all these cases to model the joint evolution of Z^{F_0} and E, the probability in Equation (3.25) can be approximated using only the first two cases, which means that $Z_n^{F_0}$ only depends on the previous value $Z_{n-1}^{F_0}$ and the current note state E_n, leading to:

$$p(Z_n^{F_0}|Z_{n-1}^{F_0}, E_n, E_{n-1}) \propto p(Z_n^{F_0}|Z_{n-1}^{F_0}) p(Z_n^{F_0}|E_n) \tag{3.28}$$

For the term $p(Z_n^{F_0}|Z_{n-1}^{F_0})$, the evolution of the physical layer of Section 3.3.3 can therefore be used once again here. It is indeed possible to consider the evolution of this layer rather independently from the rest: for singing voice signals, around note transitions, the fundamental frequencies do not always exhibit a clear jump. Instead, smooth transitions are visible, which motivates this approximation.

As for the conditional probability $p(Z_n^{F_0}|E_n)$, it is modelled using a "$\log_2$-Gaussian" density, centered around $\log_2 \mathcal{F}^{\text{MIDI}}(E_n)$ with a variance $(\sigma^{\text{MIDI}})^2$. This variance can be interpreted as a scale factor that allows more or less deviation from the standard frequency $\mathcal{F}^{\text{MIDI}}(E_n)$. The mapping on the logarithmic scale implies that σ^{MIDI} is homogeneous to one octave. The conditional probability writes:

$$p(Z_n^{F_0} = u|E_n = n^{\text{MIDI}}) \propto \exp\left(-\frac{(\log_2 \mathcal{F}(u) - \log_2 \mathcal{F}^{\text{MIDI}}(n^{\text{MIDI}}))^2}{2(\sigma^{\text{MIDI}})^2}\right) \tag{3.29}$$

Such a *prior* on the fundamental frequency of the state $Z_n^{F_0}$ constrains the fundamental frequency to be concentrated around the corresponding value for the note E_n. It is to be compared with the Gaussian distributions of the note-event model [Ryynänen and Klapuri, 2004]. Higher values for σ^{MIDI} allow more or less variability, notably for vibrato phenomena. We have used a value of $\sigma^{\text{MIDI}} \approx 0.12$ octave, which means that the F0s for a given note n^{MIDI} should be concentrated around about one and a half semitone from $\mathcal{F}^{\text{MIDI}}(n^{\text{MIDI}})$.

At last, in order to define **the likelihood of the note sequence** $p(E)$, the segmental duration model proposed in [Vincent, 2004] is used and adapted to the assumptions of our model. The lead instrument being monophonic, the sequence of notes E does not allow several notes during a single frame. At frame n, E_n takes values in the set of MIDI note numbers $[n_{\min}^{\text{MIDI}}, n_{\max}^{\text{MIDI}}]$, plus an additional value representing the silence, conventionally

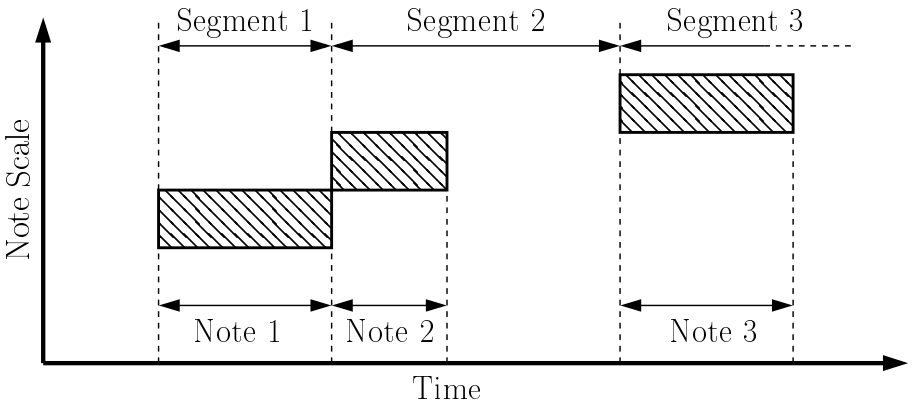


Figure 3.15: Definition of notes and segments for the segmental duration model [Vincent, 2004]. The notes are represented by hatched rectangles. The onsetting times of the notes are represented by dashed lines.

set to 0. The likelihood of the whole sequence of note states $E = \{E_1, \ldots, E_N\}$ is assumed to only depend on the durations of the musical notes that are encoded by E, and not on their value on the Western musical scale. The musicological role of the notes and their evolutions are therefore discarded from our model. It is however possible in future works to add *priors* similar to those of Ryynänen and Klapuri [2008b] for instance. But since such *priors* mainly make sense when coupled with some musical key estimation, which is outside the scope of our work, these possible constraints were ignored. The explicit duration model corresponds to the desire to model musical objects, with realistic behaviours.

The duration of the l^{th} note, with $l \in [1, L]$, is denoted d_l. Its starting frame is n_l. We define segments as the interval between each start of the notes (note onsets). A last value $n_{L+1} = N$ is arbitrarily set. The notes and segments are therefore constrained to fit the observation time range. The schema on Figure 3.15 represents what the difference between notes and segments is. Intuitively, the duration model on the notes helps to avoid arbitrarily short or long notes, while the duration model on the segment level constrains the notes to start at "suitable" distances, such that the melody line is not too fast (onsets that are too close to each other) or too sparse (onsets too distant from each other).

The probability for a note to have duration d is denoted $\mathcal{D}^{\text{note}}(d)$, while the probability for a segment to have duration d is $\mathcal{D}^{\text{seg}}(d)$. Both the segment and the note durations are assumed to be log-Gaussian distributed [Vincent, 2004]:

$$\mathcal{D}(d) = \begin{cases} \dfrac{\mathbf{N}(\log d; \mu^d, \sigma^d)}{\sum_{d_{min} < d'} \mathbf{N}(\log d'; \mu^d, \sigma^d)}, \text{ if } d_{min} < d \\ 0, \text{ if } d < d_{min} \end{cases} \tag{3.30}$$

where d_{min} is the minimal duration of a segment or a note.

The probability of a given sequence E, with the corresponding onsetting frames, therefore verifies:

$$p(E) = \mathcal{D}^{\text{seg}}(n_1) \prod_{l=1}^{L} \mathcal{D}^{\text{note}}(d_l) \mathcal{D}^{\text{seg}}(n_{l+1} - n_l) \tag{3.31}$$

In order to be able to compute the likelihood iteratively on the frame number, two types of probability functions need to be defined:

- The likelihood of a note and a segment lasting longer than a given value d:

$$\mathcal{Q}^{\text{note}}(d) = \sum_{d' \geq d} \mathcal{D}^{\text{note}}(d')$$
$$\mathcal{Q}^{\text{seg}}(d) = \sum_{d' \geq d} \mathcal{D}^{\text{seg}}(d')$$

- The continuation likelihood of a note and a segment knowing that it already lasted a given duration d:

$$\mathcal{T}^{\text{note}}(d) = \mathcal{Q}^{\text{note}}(d+1)/\mathcal{Q}^{\text{note}}(d)$$
$$\mathcal{T}^{\text{seg}}(d) = \mathcal{Q}^{\text{seg}}(d+1)/\mathcal{Q}^{\text{seg}}(d)$$

Using these definitions, it is possible to compute the conditional likelihood of E_n knowing the past sequence $E_{1:n-1}$:

$$p(E_n|E_{1:n-1}) = \begin{cases} \mathcal{T}^{\text{seg}}(n - n_{\text{last}}) & \text{, if } E_n = E_{n-1} = 0 \\ \mathcal{T}^{\text{seg}}(n - n_{\text{last}})\mathcal{T}^{\text{note}}(n - n_{\text{last}}) & \text{, if } E_n = E_{n-1} \neq 0 \\ \mathcal{T}^{\text{seg}}(n - n_{\text{last}})(1 - \mathcal{T}^{\text{note}}(n - n_{\text{last}})) & \text{, if } E_n = 0, E_{n-1} \neq 0 \\ (1 - \mathcal{T}^{\text{seg}}(n - n_{\text{last}}))\mathcal{T}^{\text{note}}(0) & \text{, if } E_n \neq 0, E_{n-1} = 0 \\ (1 - \mathcal{T}^{\text{seg}}(n - n_{\text{last}}))(1 - \mathcal{T}^{\text{note}}(n - n_{\text{last}})) & \text{, if } E_n \neq E_{n-1} \neq 0 \end{cases} \tag{3.32}$$

where n_{last} is the frame number where an attack last occured. Of course, after each of the last two assignments, the onsetting parameter is updated, such that $n_{\text{last}} \leftarrow n$. Each of the above lines corresponds to a specific situation when exploring the state space from frame $n-1$ to frame n: two consecutive silence frames, two consecutive identical note states, one silence following an active note, an active note after a silence, and at last an active note following another different note.

This formalism allows to compute the probability of partial paths, taken from all the possible paths up to frame $n-1$, $E_{1:n-1} \in [n_{\min}^{\text{MIDI}}, n_{\max}^{\text{MIDI}}]^{n-1}$, which is useful for the estimation of the desired sequence $\widehat{E}$. In a transcription application, $\widehat{E}$ can be estimated by MAP:

$$\begin{aligned} \widehat{E} &= \arg\max_E p(E|\mathbf{X}) \\ &= \arg\max_E \sum_{Z^\Phi, Z^{F_0}} p(E, Z^\Phi, Z^{F_0}|\mathbf{X}) \end{aligned} \tag{3.33}$$

However, as for the above HMM sequences, it is in practice easier to jointly determine the best sequences E, Z^Φ and Z^{F_0}:

$$\begin{aligned} \widehat{E}, \widehat{Z}^\Phi, \widehat{Z}^{F_0} &= \arg\max_{E, Z^\Phi, Z^{F_0}} p(E, Z^\Phi, Z^{F_0}|\mathbf{X}) \\ \widehat{E}, \widehat{Z}^\Phi, \widehat{Z}^{F_0} &= \arg\max_{E, Z^\Phi, Z^{F_0}} p(\mathbf{X}, Z^\Phi, Z^{F_0}, E) \end{aligned} \tag{3.34}$$

The joint likelihood for the observation and all the hidden state sequences is given by Equation (3.22). This less optimal criterion is necessary, since it allows to compute the joint likelihood for a limited number of states, using a pruning strategy which eliminates most

of the improbable states without having to compute their scores. Computing the joint likelihood for given sequences $\{E, Z^{\Phi}, Z^{F_0}\}$ implies estimating the corresponding parameters for the frame-wise model: each state sequence $\{E, Z^{\Phi}, Z^{F_0}\}$ requires its own set of parameters $\mathbf{\Theta}^{\text{GSMM}} = \{\mathbf{W}^{\Phi}, \mathbf{B}, \mathbf{W}^{M}, \mathbf{H}^{M}\}$ (or $\mathbf{\Theta}^{\text{SGSMM}} = \{\mathbf{H}^{\Gamma}, \mathbf{B}, \mathbf{W}^{M}, \mathbf{H}^{M}\}$ when using the smoothed version of $\mathbf{W}^{\Phi}$), which is unrealistic in practice, and some more approximations during the estimation are done, as explained in section 5.4.2.

At last, note how some features originally present in [Vincent, 2004] were altered in our framework. First the possibility of modelling reverberation and echo is dismissed: we assume that for a given frame n, either E_n is a silence state (no note played by the lead voice), either it is a note. This could of course be implemented in a future model, but may lead to the need to redefine the F0 layer and also to an even heavier computational load. Second, the descriptor level in our formulation is easier to interprete than the original one, since it is directly linked with physical quantities, namely the fundamental frequencies played by the lead instrument.

3.4 From the GSMM to the Instantaneous Mixture Model (IMM): links and differences

We have proposed in [Durrieu et al., 2008a] an alternative to the GSMM frame-wise model of the observation: the Instantaneous Mixture Model (IMM). The GSMM model for the singing voice may seem too complicated, and quite prone to mistakes and errors when the model does not fit the data anymore. We propose a modified model, derived from the GSMM, which, as a first aim, eases up the computational load.

The new model is first formally given, along with interpretations and discussions about its relevance in the present case. The temporal constraints from the GSMM model are then adapted to this framework, and other constraints are at last considered in order to bring the IMM closer to the original assumptions on the main melody, especially for the monophonic (mono-pitch) assumption.

3.4.1 IMM: formulation and interpretations

When analyzing the Maximum Likelihood or Maximum A Posteriori estimation problem for the above GSMM, one can identify one of the difficulties related to such a hidden state framework: computing the posterior probabilities $p(Z_n^{F_0}|\mathbf{x}_n)$ requires a particular care to avoid numerical problems, as explained in Appendix B.2.1. One way to simplify the problem, from the estimation point of view, is to drop the hidden state model for the frame-wise model: the model presented in this section therefore aims at replacing the GSMM proposed in Section 3.3.2.1 while still keeping the same interpretations for the source/filter model. The temporal constraints, physical and musicological, can be added after, as explained in Section 3.4.2, where the hidden states Z^{F_0} and Z^{Φ} are considered again, within a Bayesian framework.

The GSMM model assumption stated that the likelihood of the leading instrument was a weighted sum of conditional likelihoods. Instead, let us assume this time that the leading voice signal is a mixture of Gaussian components, like our accompaniment model, but with the same characteristics as the source/filter model designed for the GSMM, Section 3.3.2.1. We denote $\boldsymbol{\nu}_n^{ku}$ the Gaussian component whose covariance matrix is parameterized by the

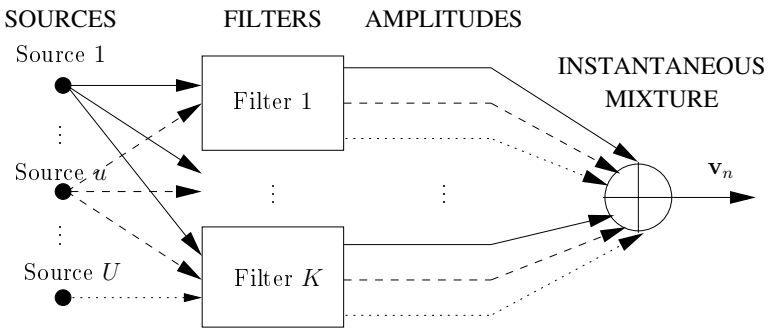


Figure 3.16: Schematic principle of the generative IMM for the main instrument part. At each frame, all the U sources, each filtered by the K filters, are multiplied by amplitudes and added together to produce the leading voice signal.

filter shape k and the source u:

$$\mathbf{v}_n = \sum_{ku} \boldsymbol{\nu}_n^{ku} \tag{3.35}$$

$$\boldsymbol{\nu}_n^{ku} \sim \mathcal{N}_c\left(\mathbf{0}, b_{kun}\mathrm{diag}(\mathbf{w}_k^{\Phi} \bullet \mathbf{w}_u^{F_0})\right) \tag{3.36}$$

This is equivalent to dropping the monopitch assumption of the leading instrument: the resulting model assumes that the leading voice signal is the sum of the individual Gaussian components, allowing several of them to be active for one given frame. The signal $\mathbf{v}_n$ can therefore be considered as the instantaneous mixture of several components, hence the name for this model: **the Instantaneous Mixture Model (IMM)**. The schematic generative process for the IMM, as depicted on Figure 3.16, is to be compared with the GSMM process, Figure 3.10. The "state selector" has been replaced by a simple addition operator. Assuming the independence between these different signals leads to a Gaussian distribution for $\mathbf{v}_n$, with on the diagonal of the covariance matrix the variance vector $\mathbf{s}_n^{V,\mathrm{IMM}}$:

$$\mathbf{s}_n^{V,\mathrm{IMM}} = \sum_{ku} b_{kun}\mathbf{w}_k^{\Phi} \bullet \mathbf{w}_u^{F_0} \tag{3.37}$$

In this framework, the scale parameters b_{kun} can be interpreted as activation coefficients: for an inactive pitch $\mathcal{F}(u)$ at frame n, then $b_{kun} = 0, \forall k$. The mono-pitch assumption of the GSMM could then be obtained with some sparsity constraint or penalization imposed on these coefficients, as will be discussed in Section 3.4.3.

The amplitude coefficients b_{kun} can be further split into their contributions to the filter part and to the source part of the leading voice, respectively h_{kn}^{Φ} and $h_{un}^{F_0}$, such that $b_{kun} \approx h_{kn}^{\Phi} h_{un}^{F_0}$. The benefit of such a decomposition is first that there are less coefficients to estimate, $(K+U)\times N$ instead of $K \times U \times N$, and second that the coefficients become much easier to interprete in terms of F0 energies. Equation (3.37) then writes:

$$\begin{aligned}
\mathbf{s}_n^{V,\mathrm{IMM}} &= \sum_{ku} h_{kn}^{\Phi}\mathbf{w}_k^{\Phi} \bullet h_{un}^{F_0}\mathbf{w}_u^{F_0} \\
\mathbf{s}_n^{V,\mathrm{IMM}} &= \left(\sum_{k} h_{kn}^{\Phi}\mathbf{w}_k^{\Phi}\right) \bullet \left(\sum_{u} h_{un}^{F_0}\mathbf{w}_u^{F_0}\right)
\end{aligned} \tag{3.38}$$

The global likelihood for the mixture, with this frame-wise model, is:

$$\mathbf{x}_n \sim \mathcal{N}_c\left(\mathbf{0}, \mathrm{diag}\left(\left(\left(\sum_k h_{kn}^{\Phi}\mathbf{w}_k^{\Phi}\right) \bullet \left(\sum_u h_{un}^{F_0}\mathbf{w}_u^{F_0}\right) + \left(\sum_r h_{rn}^{M}\mathbf{w}_r^{M}\right)\right)\right)\right) \tag{3.39}$$

$$\mathbf{x}_n \sim \mathcal{N}_c\left(\mathbf{0}, \mathrm{diag}\left(\left(\mathbf{W}^{\Phi}\mathbf{h}_n^{\Phi}\right) \bullet \left(\mathbf{W}^{F_0}\mathbf{h}_n^{F_0}\right) + \left(\mathbf{W}^{M}\mathbf{h}_n^{M}\right)\right)\right) \tag{3.40}$$

With the assumption that, in the frame-wise model, the FT vectors are independent from one frame to the other, and with the diagonal covariance matrix assumption, the likelihood of the whole observed STFT matrix $\mathbf{X}$ is equal to the product of the individual time-frequency bin likelihoods x_{fn}:

$$x_{fn} \sim \mathcal{N}_c\left(\mathbf{0}, \left(\sum_k h_{kn}^{\Phi}w_{fk}^{\Phi}\right)\left(\sum_u h_{un}^{F_0}w_{fu}^{F_0}\right) + \left(\sum_r h_{rn}^{M}w_{fr}^{M}\right)\right) \tag{3.41}$$

This particular observation likelihood for $\mathbf{X}$ can be denoted as:

$$\mathbf{X} \sim \mathcal{N}_c\left(\mathbf{0}, \left(\mathbf{W}^{\Phi}\mathbf{H}^{\Phi}\right) \bullet \left(\mathbf{W}^{F_0}\mathbf{H}^{F_0}\right) + \left(\mathbf{W}^{M}\mathbf{H}^{M}\right)\right) \tag{3.42}$$

where $\mathbf{H}^{\Phi}$ and $\mathbf{H}^{F_0}$ respectively are the amplitude (activation) matrices for the filter and the source parts of the lead instrument voice. In the expression (3.42), the matrix products may again recall the NMF approximation techniques [Lee and Seung, 2001], a link that is further developed and discussed in Chapter 4.

The smooth filter model can also be readily implemented by replacing the matrix $\mathbf{W}^{\Phi}$ by the expression in Equation (3.13). The model then becomes the Smooth filters-Instantaneous Mixture Model (SIMM) and the above equation verifies:

$$\mathbf{X} \sim \mathcal{N}_c\left(\mathbf{0}, \left(\mathbf{W}^{\Gamma}\mathbf{H}^{\Gamma}\mathbf{H}^{\Phi}\right) \bullet \left(\mathbf{W}^{F_0}\mathbf{H}^{F_0}\right) + \left(\mathbf{W}^{M}\mathbf{H}^{M}\right)\right) \tag{3.43}$$

The simplicity of this new model, *i.e.* avoiding the use of hidden states during the frame-wise parameter estimation, actually makes it more difficult to include the temporal constraint as proposed in Section 3.3.3 or the rest state introduced in Section 3.3.2.3. Approximations and proposals to compensate and estimate the desired fundamental sequence are discussed in the following section. The silence state remains an issue for the IMM, and only a heuristic solution was found to address it, as explained in Section 6.1.1.2. The parameters involved in the IMM and the SIMM are summarized in Table 3.2.

3.4.2 Adaptation of the temporal constraint for the evolution of the sequence Z^{F_0}

There are several ways of estimating the desired sequence. A framework close to the above GSMM can be derived within a Bayesian framework. Using the above IMM model, the desired sequence could then for instance be retrieved by MAP estimation. In comparison with the above Maximum Likelihood framework, in a Bayesian framework, even the parameters are considered as random variables, in addition to the other variables, observations and hidden state sequences.

The framework for the frame-wise statistical model SIMM[6] can be expressed as the probability of the observed signal, conditionally upon the parameter set $\mathbf{\Theta}^{\mathrm{SIMM}} =$

[6] IMM and SIMM are the same models, except for the smoothness of the filters, which does not influence the forth-coming developments. In such a case, we will refer to these models as the (S)IMM model. The results given in this section for the SIMM also hold for the IMM, replacing the matrix products $\mathbf{W}^{\Gamma}\mathbf{H}^{\Gamma}$ by $\mathbf{W}^{\Phi}$ when necessary.

Table 3.2: (S)IMM: Parameters for the leading voice and the accompaniment. All the parameters are estimated, except when indicated otherwise. If a parameter is exclusively used within the IMM or the SIMM, then it is also indicated.

Parameter	Description	Remarks
	Mixture $\mathbf{s}_n^{\text{IMM}} = \mathbf{s}_n^{V,\text{IMM}} + \mathbf{s}_n^m$	
	Leading Instrument $\mathbf{s}_n^{V,\text{IMM}} = \mathbf{W}^{\Phi}\mathbf{h}_n^{\Phi} \bullet \mathbf{W}^{F_0}\mathbf{h}_n^{F_0}$	
$\mathbf{W}^{\Phi}$	Matrix of spectral envelopes for the filter part	IMM
$\mathbf{w}_k^{\Phi}$	Vector of filter spectral envelope k	IMM
w_{fk}^{Φ}	Filter spectral envelope k, at frequency bin f	IMM
$\mathbf{W}^{F_0}$	Dictionary of source comb spectra	Fixed
$\mathbf{w}_u^{F_0}$	Source comb spectrum u	Fixed
$w_{fu}^{F_0}$	Source comb spectrum u at frequency bin f	Fixed
$\mathbf{W}^{\Gamma}$	Dictionary of smooth elementary filter parts	SIMM, Fixed
$\mathbf{w}_p^{\Gamma}$	Vector of smooth elementary filter p	SIMM, Fixed
w_{fp}^{Γ}	Smooth elementary filter p at frequency bin f	SIMM, Fixed
$\mathbf{H}^{\Phi}$	Amplitude matrix for the filter part of the lead instrument	
$\mathbf{h}_n^{\Phi}$	Amplitude vector for the filter part, at frame n	
h_{kn}^{Φ}	Amplitude for fiter k, at frame n	
$\mathbf{H}^{\Gamma}$	Amplitude matrix for the decomposition of $\mathbf{W}^{\Phi}$ on $\mathbf{W}^{\Gamma}$	SIMM
$\mathbf{h}_k^{\Gamma}$	Amplitude vector for filter k	SIMM
h_{pk}^{Γ}	Amplitude for filter k on element p	SIMM
$\mathbf{H}^{F_0}$	Amplitude matrix for the source part	
$\mathbf{h}_n^{F_0}$	Amplitude vector for the source part, at frame n	
$h_{un}^{F_0}$	Amplitude for source element u, at frame n	
	Accompaniment $\mathbf{s}_n^M = \mathbf{W}^M\mathbf{h}_n^M$	
$\mathbf{W}^M$	Matrix of spectral shape for the accompaniment	
$\mathbf{w}_r^M$	Spectral shape for element r of $\mathbf{W}^M$	
w_{fr}^M	Spectral shape r, at frequency bin f	
$\mathbf{H}^M$	Matrix of amplitudes for the accompaniment	
$\mathbf{h}_n^M$	Vector of amplitudes for the accompaniment at frame n	
h_{rn}^M	Amplitude associated with element r of $\mathbf{W}^M$ at frame n	

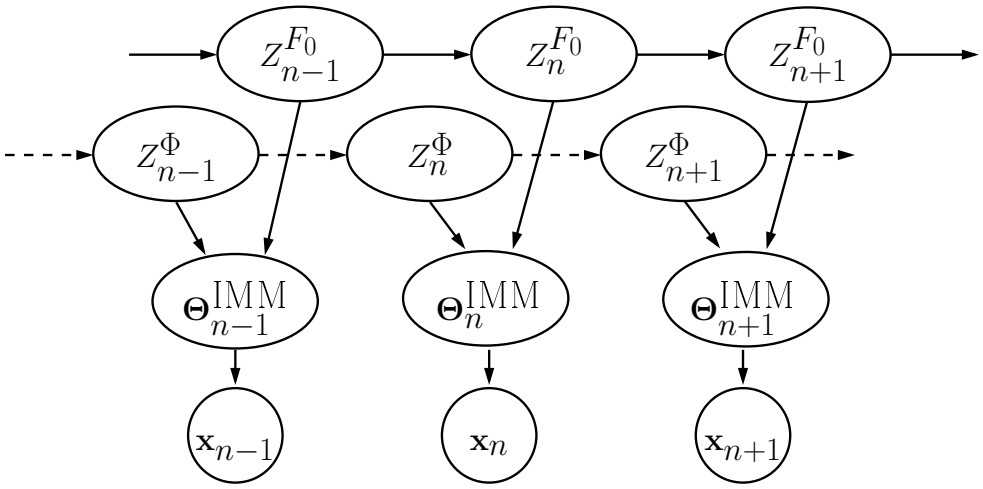


Figure 3.17: Graphical model for the (S)IMM within a Bayesian framework. The musicological layer E was omitted here, but can be added without modification from Section 3.3.4 to the layer Z^{F_0}.

$\{\mathbf{H}^{\Gamma}, \mathbf{H}^{\Phi}, \mathbf{H}^{F_0}, \mathbf{W}^{M}, \mathbf{H}^{M}\}$:

$$\mathbf{X}|\mathbf{\Theta}^{\text{SIMM}} \sim \mathcal{N}_c\left(\mathbf{0}, \left(\mathbf{W}^{\Gamma}\mathbf{H}^{\Gamma}\mathbf{H}^{\Phi}\right) \bullet \left(\mathbf{W}^{F_0}\mathbf{H}^{F_0}\right) + \left(\mathbf{W}^{M}\mathbf{H}^{M}\right)\right) \tag{3.44}$$

As shown on Figure 3.17, the difference with the previous GSMM framework is the introduction of the parameters as random variables between the observation layer ($\mathbf{x}_n$) and the sequence layers (Z^{F_0} and Z^{Φ}). The musicological layer E can be easily added in this graph, since the physical layer for the fundamental frequencies, Z^{F_0}, was not changed.

The joint likelihood (3.22) can then be written as follows:

$$p(\mathbf{X}, \mathbf{\Theta}^{(\text{S})\text{IMM}}, Z^{\Phi}, Z^{F_0}, E) = p(\mathbf{X}|\mathbf{\Theta}^{(\text{S})\text{IMM}})p(\mathbf{\Theta}^{(\text{S})\text{IMM}}|Z^{\Phi}, Z^{F_0})p(Z^{\Phi})p(Z^{F_0}|E)p(E) \tag{3.45}$$

where the different contributions write:

$$p(\mathbf{X}|\mathbf{\Theta}^{(\text{S})\text{IMM}}) = \prod_n p(\mathbf{x}_n|\mathbf{\Theta}^{(\text{S})\text{IMM}}_n) \tag{3.46}$$

$$p(\mathbf{\Theta}^{(\text{S})\text{IMM}}|Z^{\Phi}, Z^{F_0}) = \prod_n p(\mathbf{\Theta}^{(\text{S})\text{IMM}}_n|Z^{\Phi}_n, Z^{F_0}_n) \tag{3.47}$$

$$p(Z^{\Phi}) = p(Z^{\Phi}_1)\prod_n p(Z^{\Phi}_n|Z^{\Phi}_{n-1}) \tag{3.48}$$

$$p(Z^{F_0}|E) = p(Z^{F_0}_1|E_1)\prod_n p(Z^{F_0}_n|Z^{F_0}_{n-1}, E_n, E_{n-1}) \tag{3.49}$$

$$p(E_{1:n}) = p(E_n|E_{1:n-1})p(E_{1:n-1}), \forall n \in [1, N] \tag{3.50}$$

Equations (3.49) and (3.50) can be further developed in the same way as for the GSMM, respectively through Equations (3.28) and (3.31). Equation (3.46) is given by Equation (3.40), with $\mathbf{\Theta}^{(\text{S})\text{IMM}}_n = \{\mathbf{h}^{\Phi}_n, \mathbf{h}^{F_0}_n, \mathbf{h}^{M}_n, \mathbf{W}^{M}\}$ the parameter set limited to frame n.

The density of the parameters $\mathbf{\Theta}^{(\text{S})\text{IMM}}_n$ conditionally upon the states $Z^{\Phi}_n, Z^{F_0}_n$, namely $p(\mathbf{\Theta}^{(\text{S})\text{IMM}}_n|Z^{\Phi}_n, Z^{F_0}_n)$ in Equation (3.47), can be considered as the constraint on the parameters by the underlying states. Typically, what is expected from such constraints is

that the energy in $\mathbf{h}_n^{F_0}$ is concentrated around the coefficient $h_{un}^{F_0}$, when $Z_n^{F_0} = u$. It is interesting to note that, with this formalism, the GSMM actually is included in the above system of equations, with the following particular condition for $\mathbf{\Theta}_n^{(\mathrm{S})\mathrm{IMM}}$:

$$p(\mathbf{\Theta}_n^{(\mathrm{S})\mathrm{IMM}}|Z_n^{\Phi} = k, Z_n^{F_0} = u) \neq 0 \Rightarrow \forall (i,j) \neq (k,u), h_{in}^{\Phi} h_{jn}^{F_0} = 0 \tag{3.51}$$

In other words, the only amplitude coefficients that are allowed to be non-nul at frame n are the ones corresponding to the active state (k,u). In this case, one may either assume a certain *prior* distribution on the parameters or leave an uninformative *prior*, as is (implicitly) the case for the GSMM.

As for the use of this formalism for the IMM, the description of the appropriate *prior* is delayed to Section 5.2.3, where the motivation for the (S)IMM and the approximations that are necessary are also made explicit. In the next section, Section 3.4.3, an example of how to use this framework to constrain the parameters is described.

3.4.3 Constraints in SIMM to approximate the monophonic assumption

As for the GSMM, the statistical framework of the IMM model allows to define prior densities on the parameters. In the definition of this new model, the monophonic assumption was dropped, while it could be considered as an important characteristic of the leading instrument. Although the proposed scheme, *i.e.* the IMM with the above Bayesian framework, produces fairly good results in our experiments, it is worth introducing such a possibility, which may help further improving the results.

In order to illustrate this possibility, a prior density on the amplitude coefficients $\mathbf{H}^{F_0}$ is defined. It aims at penalizing the amplitude of a fundamental frequency if the amplitude of the corresponding lower octave is high. Let $u_8 = u - 12U_{st}$, for $12U_{st} < u < U$. For an amplitude coefficient $h_{un}^{F_0}$ at frame n, source u, if the lower octave coefficient, $h_{u8n}^{F_0}$, is relatively high, then $h_{un}^{F_0}$ should be constrained to be low. Inversely, if $h_{u8n}^{F_0}$ is relatively low, then the constraint should not apply, and should tend to some uniform distribution.

To obtain this effect, a Gamma prior (see Appendix A.2) for $h_{un}^{F_0}$ can be defined, with the scale parameter β_G proportional to the amplitude of the lower octave $h_{u8n}^{F_0}$, for $12U_{st} < u < U$:

$$h_{un}^{F_0} \sim \mathcal{G}\left(\alpha_G, h_{u8n}^{F_0}\right) \tag{3.52}$$

$$\log p(h_{un}^{F_0}|h_{u8n}^{F_0}) = \alpha_G \log h_{u8n}^{F_0} - \log\Gamma(\alpha_G) + (\alpha_G - 1)\log h_{un}^{F_0} - h_{u8n}^{F_0} h_{un}^{F_0} \tag{3.53}$$

The parameter α_G can be set to some value under 1, to fullfill the desired requirement. Indeed, as can be seen on Figure 3.18, for $\alpha_G = 0.9$, the distribution is almost "uniform" for low values of $h_{u8n}^{F_0}$. With higher values of $h_{u8n}^{F_0}$, low values of $h_{un}^{F_0}$ are more probable than high values. In other terms, high values for $h_{u8n}^{F_0}$ tend to penalize high values for $h_{un}^{F_0}$. For $\alpha_G > 1$, the distribution exhibits a mode, which is not desired here. On Figure 3.19, where $\alpha_G = 5$, for low values of $h_{u8n}^{F_0}$, the distribution tends to favor higher values, when the desired effect is to obtain a uniform distribution.

By doing so, the decomposition of the mixture signal onto all the possible states of the source/filter, Equation (3.41), can be better controlled: a harmonic comb from the observation signal has less chances to be decomposed onto itself (or the harmonic comb u in $\mathbf{W}^{F_0}$ with the closest fundamental frequency) plus the harmonic comb corresponding to its octave $(u + 12U_{st})$.

It is interesting to note that, taking the log-likelihood of such a prior shows that this choice corresponds to a "decorrelation" penalization (see for instance [Chen et al., 2006]).

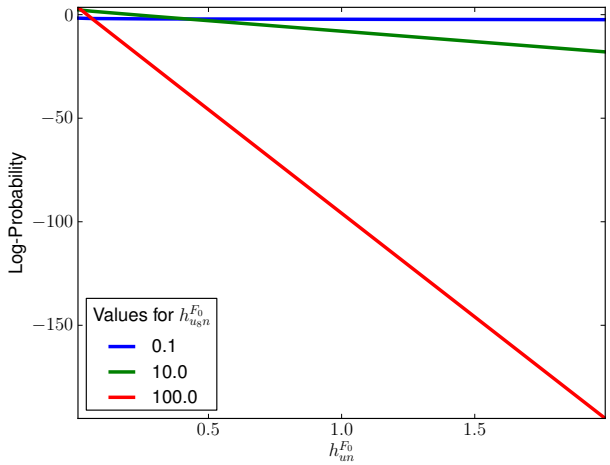


Figure 3.18: Gamma distributions for several values $h_{u8n}^{F_0}$, with $\alpha_G = 0.9$.

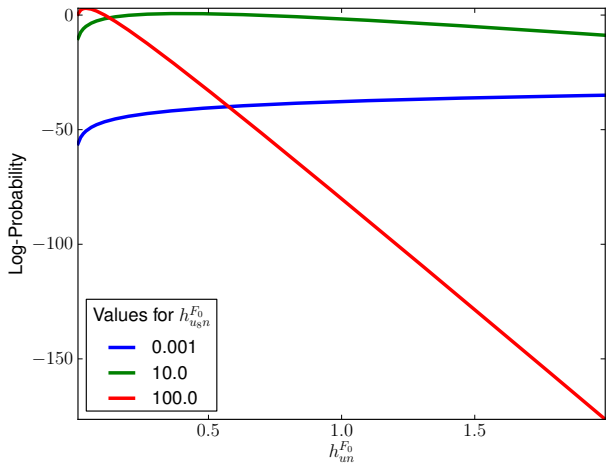


Figure 3.19: Gamma distributions for several values $h_{u8n}^{F_0}$, with $\alpha_G = 5$.

Indeed, in Equation (3.53), the term $h^{F_0}_{usn}h^{F_0}_{un}$ within a MAP estimation will be minimized such that both these amplitudes can not be high at the same time.

3.5 Signal Model Summary

In this chapter, two models have been derived: the source/filter Gaussian Scaled Mixture Model (GSMM) and the Instantaneous Mixture Model (IMM). Both models can also be further augmented with a structural smoothness constraint on the filters of the leading instrument.

We first recall the equations to which the signal obeys in the (S)GSMM model, and then give the equations corresponding to the (S)IMM model.

3.5.1 Source/Filter (S)GSMM

The joint likelihood of the observations $\mathbf{X}$, the filter and source state sequences Z^{Φ} and Z^{F_0} of the leading instrument, and its note state sequence E, for the (Smooth-filter) Gaussian Scaled Mixture Model ((S)GSMM), verifies:

$$p(\mathbf{X}, Z^{\Phi}, Z^{F_0}, E) = p(\mathbf{X}|Z^{\Phi}, Z^{F_0})p(Z^{\Phi})p(Z^{F_0}|E)p(E) \tag{3.54}$$

The different expressions of the right hand side of Equation (3.54) are given by several assumptions on the signal. First, the frames of $\mathbf{X}$ are independent conditionally upon the leading instrument states Z^{Φ} and Z^{F_0}. The sequence Z^{Φ} is a (hidden) Markov process, while Z^{F_0} is a Markov process, conditionally upon the note sequence E. At last, the evolution of E is given by an explicit model on the durations of the notes.

All the above assumptions can be translated into equations as follows:

$$p(\mathbf{X}|Z^{\Phi}, Z^{F_0}) = \prod_n p(\mathbf{x}_n|Z^{\Phi}_n, Z^{F_0}_n) \tag{3.55}$$

$$p(Z^{\Phi}) = p(Z^{\Phi}_1)\prod_n p(Z^{\Phi}_n|Z^{\Phi}_{n-1}) \tag{3.56}$$

$$p(Z^{F_0}|E) = p(Z^{F_0}_1|E_1)\prod_n p(Z^{F_0}_n|Z^{F_0}_{n-1}, E_n) \tag{3.57}$$

$$p(E) = \mathcal{D}^{\text{seg}}(n_1)\prod_{l=1}^{L}\mathcal{D}^{\text{note}}(d_l)\mathcal{D}^{\text{seg}}(n_{l+1} - n_l) \tag{3.58}$$

where n_l and d_l respectively are the onsetting frame number and duration (in number of frames) of the l^{th} note in the sequence E.

The conditional density in Equation (3.55) is a multivariate complex Gaussian distribution, centered, with a diagonal covariance matrix. The diagonal of that matrix is the sum of two elementary contributions, the first one being the leading instrument source/filter spectrum and the second one the instantaneous mixture of elementary components for the accompaniment:

$$p(\mathbf{x}_n|Z^{\Phi}_n = k, Z^{F_0}_n = u) = \mathbf{N}_c(\mathbf{x}_n; \mathbf{0}, \text{diag}(b_{kun}\mathbf{w}^{\Phi}_k \bullet \mathbf{w}^{F_0}_u + \mathbf{W}^M\mathbf{h}^M_n)) \tag{3.59}$$

The other probabilities that appear in the above equations, which mainly characterize the evolutions of the sequences, are defined using parametric distributions. The filter transition between the states mainly favors steady states in the sequence, the transition between the

fundamental frequencies penalizes great jumps and at last, the durations are assumed to follow a log-Gaussian distribution.

$$p(Z_1^{\Phi} = k) \propto 1 \tag{3.60}$$

$$p(Z_n^{\Phi} = k_2 | Z_{n-1}^{\Phi} = k_1) \propto \begin{cases} 1, \text{ if } k_2 = k_1 \\ \epsilon^{\Phi} << 1, \text{ if } k_2 \neq k_1 \end{cases} \tag{3.61}$$

$$\begin{aligned} p(Z_n^{F_0} = u_2 | Z_{n-1}^{F_0} = u_1, E_n = n^{\text{MIDI}}) \\ \propto \exp\left[-\alpha.\text{round}\left(|12\log_2 \mathcal{F}(u_2) - 12\log_2 \mathcal{F}(u_1)|\right)\right] \\ \times \exp\left(-\frac{[\log_2 \mathcal{F}(u_2) - \log_2 \mathcal{F}^{\text{MIDI}}(n^{\text{MIDI}})]^2}{2(\sigma^{\text{MIDI}})^2}\right) \end{aligned} \tag{3.62}$$

$$\mathcal{D}(d) = \begin{cases} \dfrac{\mathbf{N}(\log d; \mu^d, (\sigma^d)^2)}{\sum_{d_{min} < d'} \mathbf{N}(\log d'; \mu^d, (\sigma^d)^2)}, \text{ if } d_{min} < d \\ 0, \text{ if } d < d_{min} \end{cases} \tag{3.63}$$

The parameters of the log-Gaussian distribution for the durations were arbitrarily set to $\mu^{d,\text{seg}} = 0.5\text{s}$, $\mu^{d,\text{note}} = 0.3\text{s}$, $\sigma^{d,\text{seg}} = \sigma^{d,\text{note}} \approx 0.050\text{s}$, all these values are to be converted in number of frames according to the chosen frame rate. These values are the ones used in [Vincent, 2004].

In Chapter 5, the approximations and algorithms necessary to estimate the parameters and sequences are developed. More specifically, several systems, with different levels of approximations, are extracted from the above source/filter (S)GSMM model.

3.5.2 Source/Filter (S)IMM

Similarly to the (S)GSMM, the joint likelihood of the observation $\mathbf{X}$, the filter state sequence Z^{Φ}, the source sequence Z^{F_0} and the note sequence E, defining the (Smooth-filter) Instantaneous Mixture Model ((S)IMM) from Section 3.4, verifies:

$$p(\mathbf{X}, \mathbf{\Theta}^{(\text{S})\text{IMM}}, Z^{\Phi}, Z^{F_0}, E) = p(\mathbf{X}|\mathbf{\Theta}^{(\text{S})\text{IMM}}) p(\mathbf{\Theta}^{(\text{S})\text{IMM}}|Z^{\Phi}, Z^{F_0}) p(Z^{\Phi}) p(Z^{F_0}|E) p(E) \tag{3.64}$$

where the different probabilities and conditional probabilities write:

$$p(\mathbf{X}|\mathbf{\Theta}^{(\text{S})\text{IMM}}) = \prod_n p(\mathbf{x}_n|\mathbf{\Theta}_n^{(\text{S})\text{IMM}}) \tag{3.65}$$

$$p(\mathbf{\Theta}^{(\text{S})\text{IMM}}|Z^{\Phi}, Z^{F_0}) = \prod_n p(\mathbf{\Theta}_n^{(\text{S})\text{IMM}}|Z_n^{\Phi}, Z_n^{F_0}) \tag{3.66}$$

$$p(Z^{\Phi}) = p(Z_1^{\Phi}) \prod_n p(Z_n^{\Phi}|Z_{n-1}^{\Phi}) \tag{3.67}$$

$$p(Z^{F_0}|E) = p(Z_1^{F_0}|E_1) \prod_n p(Z_n^{F_0}|Z_{n-1}^{F_0}, E_n) \tag{3.68}$$

$$p(E) = \mathcal{D}^{\text{seg}}(n_1) \prod_{l=1}^{L} \mathcal{D}^{\text{note}}(d_l) \mathcal{D}^{\text{seg}}(n_{l+1} - n_l) \tag{3.69}$$

In comparison with the (S)GSMM equations, the main difference in the above set is the presence of the parameter set $\mathbf{\Theta}^{(\text{S})\text{IMM}}$ which appears explicitly as a random variable. This Bayesian framework is necessary to motivate the tracking of the melody (as embodied by

the source state sequence Z^{F_0}). $\mathbf{\Theta}^{(\mathrm{S})\mathrm{IMM}}$, which does not directly provide the F0 sequence, is another description layer, from which the desired sequence Z^{F_0} will be estimated. For the (S)IMM, the variance of the conditional likelihood of the observation equals the instantaneous mixture of all the possible source spectra and all possible filters. The accompaniment contribution stays the same as for the (S)GSMM:

$$p(\mathbf{x}_n|\mathbf{\Theta}_n^{(\mathrm{S})\mathrm{IMM}}) = \mathbf{N}_c(\mathbf{x}_n; \mathbf{0}, \mathrm{diag}(\mathbf{W}^{\Phi}\mathbf{h}_n^{\Phi} \bullet \mathbf{W}^{F_0}\mathbf{h}_n^{F_0} + \mathbf{W}^{M}\mathbf{h}_n^{M})) \tag{3.70}$$

The conditional *prior* distribution, $p(\mathbf{\Theta}_n^{(\mathrm{S})\mathrm{IMM}}|Z_n^{\Phi}, Z_n^{F_0})$ needs to be defined. This will be made clear in Section 5.2.3. The evolution equations are identical to the ones given before, since these higher level layers from the (S)GSMM are kept in the (S)IMM:

$$p(Z_1^{\Phi} = k) \propto 1 \tag{3.71}$$

$$p(Z_n^{\Phi} = k_2|Z_{n-1}^{\Phi} = k_1) \propto \begin{cases} 1, \text{ if } k_2 = k_1 \\ \epsilon^{\Phi} << 1, \text{ if } k_2 \neq k_1 \end{cases} \tag{3.72}$$

$$\begin{aligned} p(Z_n^{F_0} = u_2|Z_{n-1}^{F_0} = u_1, E_n = n^{\mathrm{MIDI}}) \\ \propto \exp\big[-\alpha.\mathrm{round}\big(|12\log_2\mathcal{F}(u_2) - 12\log_2\mathcal{F}(u_1)|\big)\big] \\ \times \exp\left(-\frac{[\log_2\mathcal{F}(u_2) - \log_2\mathcal{F}^{\mathrm{MIDI}}(n^{\mathrm{MIDI}})]^2}{2(\sigma^{\mathrm{MIDI}})^2}\right) \end{aligned} \tag{3.73}$$

$$\mathcal{D}(d) = \begin{cases} \dfrac{\mathbf{N}(\log d; \mu^d, (\sigma^d)^2)}{\sum_{d_{min}<d'}\mathbf{N}(\log d'; \mu^d, (\sigma^d)^2)}, \text{ if } d_{min} < d \\ 0, \text{ if } d < d_{min} \end{cases} \tag{3.74}$$

As for the (S)GSMM, several possible ways of estimating the parameters can be derived from the (S)IMM equations, and some of them are developed in Chapter 5.

Chapter 4

Probabilistic Non-negative Matrix Factorisation (NMF)

In this chapter, we emphasize the link between the Gaussian signal model and methods based on non-negative matrix factorization (NMF) of the power spectrogram. This link is a keystone in designing the algorithms of Chapter 5. One of the mostly used NMF estimation methods, the so-called "multiplicative gradient" method, was indeed adapted to fit the proposed parameterization of the power spectrogram. The general principle to design the updating rules can be readily used, as explained in Section 5.2.2 and in Appendix B.1.

Parts of the results from this section have been published in [Févotte et al., 2009a]. We first recall the principles of NMF, especially applied to audio signals, and then demonstrate the equivalence between the framework proposed in this thesis and NMF with Itakura-Saito (IS) divergence. At last the appropriateness of IS divergence for audio signal processing is discussed.

4.1 Non-negative Matrix Factorisation

Non-negative matrix factorisation (NMF) has recently been used in many fields as a simple yet efficient tool to reduce matrix dimension for positive valued matrices. It was made popular by Lee and Seung [2001] for image processing and clustering. For audio processing, NMF has also been widely used, notably by Smaragdis and Brown [2003] and Virtanen [2007], among others.

NMF methods for audio processing rely on a decomposition of a (non-negative) time-frequency representation of the signal. One of the main assumptions is that an "elementary" sound, say, a note played by an instrument, is characterized by a typical spectral shape, for instance a spectral harmonic comb. This note is activated when the note is actually played, and deactivated otherwise.

In practice, these two parameters, the spectral shape and the activation energy, are encoded in two matrices, respectively $\mathbf{W}$ of size $F \times R$ and $\mathbf{H}$ of size $R \times N$, where F is the number of frequency bins of the representation, R is the assumed number of different spectral shapes and N is the number of analysis frames. Let $\mathbf{S}$ be a non-negative valued time-frequency representation, such as an amplitude or a power spectrogram. The NMF of $\mathbf{S}$ consists in finding $\mathbf{W}$ and $\mathbf{H}$ that minimize a distortion measure D between $\mathbf{S}$ and

the matrix product **WH**:

$$\widehat{\mathbf{W}}, \widehat{\mathbf{H}} = \arg\min_{\mathbf{W},\mathbf{H}} D(\mathbf{S}||\mathbf{WH})$$

The distortion measure D is generally the sum over all the elements of both matrices of a scalar distortion d. Usual distortion measures include the Euclidian (EUC) distance, the Kullback-Leibler (KL) divergence ([Lee and Seung, 2001], [Virtanen, 2007]) or the Itakura-Saito (IS) divergence ([Févotte et al., 2009a]).

4.2 Statistical interpretation of Itakura-Saito-NMF (IS-NMF)

In this section, a first result on the equivalence between the Maximum Likelihood (ML) inference of the Gaussian statistical model and the minimization of the Itakura-Saito divergence between the squared magnitude of the Fourier transform and the variance of the Gaussian distribution is provided. The equivalence between the Gaussian composite model and the Itakura-Saito Non-negative Matrix Factorisation (IS-NMF) is then proved and discussed.

Let $y = \rho e^{\phi}$ be a complex random variable (for example, the Fourier transform of an audio signal). Let us assume that y follows a centered complex proper Gaussian distribution, with variance denoted s^y. The probability density function (PDF) of y, conditionally upon s^y is given, in the cylindrical coordinate system, by the joint likelihood of its magnitude and phase (see Appendix A.1.1 for details):

$$p(\rho, \phi|s^y) = \frac{\rho}{\pi s^y} \exp\left(-\frac{\rho^2}{s^y}\right)$$

In the model proposed in this work, the different variances s^y for the signals such as the leading voice or the accompaniment are parameterized and need to be estimated from the observation y.

Theorem 1 (Equivalence between ML inference and IS minimization) *Maximum Likelihood (ML) estimation of the variance s^y of the proper complex Gaussian signal y is equivalent to estimating the parameter s^y which minimizes the Itakura-Saito (IS) divergence between the squared magnitude of y and s^y.*

Before proving this theorem, the Itakura-Saito divergence needs to be defined:

Definition 8 (Itakura-Saito (IS) scalar divergence:) *The Itakura-Saito (IS) divergence, between two positive real numbers a and b is:*

$$d_{IS}(a||b) = -\log\frac{a}{b} + \frac{a}{b} - 1 \tag{4.1}$$

◇

Proof of Theorem 1 Let $y = \rho e^{\phi}$ be a complex random variable. y is assumed to follow a complex proper Gaussian distribution, with mean $\mu^y = 0$ and variance s^y. The estimated

$\widehat{s}^y$ that maximizes the maximum likelihood is such that:

$$\begin{aligned}\widehat{s}^y &= \arg\max_{s^y} \mathbf{N}_c(y; 0, s^y) \qquad (4.2)\\ &= \arg\max_{s^y} \log \mathbf{N}_c(y; 0, s^y)\\ &= \arg\max_{s^y} \log(\rho) - \log(\pi s^y) - \frac{\rho^2}{s^y}\\ &= \arg\max_{s^y} -\log(s^y) - \frac{\rho^2}{s^y}\\ &= \arg\min_{s^y} -\log(\frac{\rho^2}{s^y}) + \frac{\rho^2}{s^y} - 1\\ \widehat{s}^y &= \arg\min_{s^y} d_{\mathrm{IS}}(\rho^2||s^y) \qquad (4.3)\end{aligned}$$

The identity between Equations (4.2) and (4.3) concludes the proof. ∎

Theorem 1 also holds, under various conditions, for multivariate Gaussian processes. However, estimating the variance for each time-frequency bin of the STFT of the observation is trivial, since the minimization in Eq. (4.3) is obtained for $\widehat{s}^y = \rho^2$. In our work, the variance for the whole STFT is parameterized by a parameter set $\mathbf{\Theta}$. Controlling the structure of the variance through these parameters may allow to extract more information than the raw sinusoidal information provided by the STFT.

The observations are Fourier transform vectors, stacked into an STFT matrix. The IS divergence to compare two such matrices can be defined as in the following definition.

Definition 9 (IS divergence between 2 matrices:) *Let A and B two $I \times J$ positive valued matrices. The IS divergence between A and B is defined as:*

$$D_{IS}(A||B) = \sum_{i,j} d_{IS}(a_{ij}||b_{ij}) = \sum_{i,j} -\log\frac{a_{ij}}{b_{ij}} + \frac{a_{ij}}{b_{ij}} - 1 \qquad (4.4)$$

◇

Let $\mathbf{Y}$ be the STFT of signal y. For each frame n of $\mathbf{Y}$, the covariance matrix is diagonal, with a diagonal parameterized by a parameter set $\mathbf{\Theta}$ which is to be estimated: $\mathbf{\Sigma}_n^y = \mathrm{diag}(\mathbf{s}_n^y(\mathbf{\Theta}))$. The frames are assumed to be independent one from the other. $\mathbf{\Theta}$ follows a *prior* distribution $p(\mathbf{\Theta})$, which partially reflects the spectral and temporal structures for the resulting variances.

The following theorem establishes the equivalence between Maximum A Posteriori (MAP) estimation of the parameters in $\mathbf{\Theta}$ and the penalized minimization of the IS divergence between the power spectrogram $|\mathbf{Y}|^2 = \mathbf{S}$ and the parameterized variance $\mathbf{S}(\mathbf{\Theta})$, with penalization terms derived from the *prior* distribution $p(\mathbf{\Theta})$.

Theorem 2 (MAP parameter inference and penalized IS minimization) *Maximum A Posteriori (MAP) estimation of the parameters in $\mathbf{\Theta}$, assuming there exists a unique solution, is equivalent to estimating the set $\mathbf{\Theta}$ which minimizes the penalized Itakura-Saito (IS) divergence between the squared magnitude of $\mathbf{Y}$ and $\mathbf{S}(\mathbf{\Theta})$.*

The penalization terms are given by the parameter set prior distribution.

$$\widehat{\mathbf{\Theta}} = \arg\max_{\mathbf{\Theta}} p(\mathbf{\Theta}|\mathbf{Y}) = \arg\min_{\mathbf{\Theta}} \left[D_{IS}(\mathbf{S}||\mathbf{S}(\mathbf{\Theta})) - \log p(\mathbf{\Theta})\right]$$

Proof

Assuming there is one unique solution that maximizes the posterior distribution of the parameters in Θ, given the observations, this solution verifies:

$$\begin{aligned}\widehat{\boldsymbol{\Theta}} &= \arg\max_{\boldsymbol{\Theta}} p(\Theta|\mathbf{Y}) \\ &= \arg\max_{\boldsymbol{\Theta}} p(\mathbf{Y}|\boldsymbol{\Theta})p(\boldsymbol{\Theta}) \\ &= \arg\min_{\boldsymbol{\Theta}} -\log p(\mathbf{Y}|\boldsymbol{\Theta}) - \log p(\boldsymbol{\Theta}) \\ &= \arg\min_{\boldsymbol{\Theta}} \left[D_{IS}(\mathbf{S}||\mathbf{S}(\boldsymbol{\Theta})) - \log p(\boldsymbol{\Theta})\right] \quad \blacksquare\end{aligned}$$

This theorem is useful for two reasons: first it allows to use all the mathematical and optimisation techniques that have been derived for IS minimisation, especially the well furnished NMF literature, and second, it allows to easily define *prior* distributions for the parameters in $\boldsymbol{\Theta}$ and include them in the estimation process. The conditions on the uniqueness of $\boldsymbol{\Theta}$ could be dropped, but the equivalence would then rely on the fact that any element of the set of solutions for the MAP problem, $\{\boldsymbol{\Theta}|\arg\max_{\boldsymbol{\Theta}} p(\boldsymbol{\Theta}|\mathbf{Y})\}$, also belongs to the set of solutions for the IS minimisation problem.

These equivalence theorems provide us with a nice interpretability of the solutions we obtain, notably through the properties of the IS divergence, discussed in Section 4.3.

At last, let us state the theorem of equivalence between the Gaussian composite model and NMF using IS minimisation (IS-NMF) as derived by Févotte et al. [2009a]:

Theorem 3 (ML inference and NMF with IS minimisation) *Consider the generative model defined by, $\forall n = 1, \ldots, N$:*

$$\mathbf{x}_n = \sum_{r=1}^{R} \mathbf{c}_n^r \tag{4.5}$$

where $\mathbf{x}_n$ and $\mathbf{c}_n^r$ belong to $\mathbb{C}^{F\times 1}$, $\forall r \in [1, R]$ and

$$\mathbf{c}_n^r \sim \mathcal{N}_c(\mathbf{0}_F, h_{rn}\, diag(\mathbf{w}_r)) \tag{4.6}$$

The components $\mathbf{c}_n^1, \ldots, \mathbf{c}_n^R$ are assumed mutually independent and individually and independently distributed. Let $\mathbf{S}$ be the matrix with entries verifying $s_{fn} = |x_{fn}|^2, \forall (f, n) \in [1, F] \times [1, N]$.

The maximum likelihood estimation of $\mathbf{W}$ and $\mathbf{H}$ from $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_N]$ is equivalent to the NMF of $\mathbf{S}$ into $\mathbf{S} \approx \mathbf{WH}$, where the Itakura-Saito divergence is used.

Proof Thanks to the independence conditions of the components, the distribution of the observation is also a complex proper Gaussian, centered and with variance equal to the sum of all the variances:

$$\mathbf{x}_n \sim \mathcal{N}_c(\mathbf{0}_F, \sum_{r=1}^{R} h_{rn}\mathrm{diag}(\mathbf{w}_r)) \tag{4.7}$$

The ML estimation consists in determining the matrices $\mathbf{W}$ and $\mathbf{H}$ that maximize the criterion $C_{\mathrm{ML}}(\mathbf{W}, \mathbf{H})$ equal to the log-likelihood in Equation (4.7).

As stated by Theorems 1 and 2, this is equivalent to minimizing the IS divergence between $\mathbf{S}$ and the product $\mathbf{WH}$:

$$\arg\max_{\mathbf{W},\mathbf{H}} C_{\mathrm{ML}}(\mathbf{W},\mathbf{H}) = \arg\min_{\mathbf{W},\mathbf{H}} D_{\mathrm{IS}}(\mathbf{S}||\mathbf{WH}). \tag{4.8}$$

which ends the proof. ∎

This theorem is interesting because it maps a flexible statistical model on the widely used NMF tool. The link between our statistical framework and the NMF approximation permits to use all the mathematical background associated with NMF: in this work, we have focussed on adapting the multiplicative updates to our need. However, many other possibilities for parameter estimation in such an NMF framework exist, faster algorithms and implementations may also be considered if the computation time matters. The purpose of this study is to validate the model for both the transcription and the separation, and algorithm optimisation is therefore outside the scope of this work.

4.3 Properties of the Itakura-Saito (IS) divergence

Before discussing the results of the estimations using our Gaussian models (or equivalently with the IS divergence minimisation), it is worth browsing some properties of the IS divergence, which may help us to better understand the result we obtain for our audio specific applications. Two properties are especially important in our applications: the scale invariance of the IS divergence and the (non-)convexity of the resulting cost function.

The first property, **the scale invariance of the IS divergence**, is easy to verify. Let x and y be two positive real scalar values and λ a positive scale factor. Then the following proposition holds:

Proposition 2 (Scale invariance of IS divergence) *The value of the IS divergence between x and y is equal to the IS divergence between λx and λy, where $x, y, \lambda \in \mathbb{R}^{+*}$:*

$$d_{IS}(x||y) = d_{IS}(\lambda x||\lambda y) \tag{4.9}$$

Proof

$$\begin{aligned} d_{\mathrm{IS}}(x||y) &= -\log\frac{x}{y} + \frac{x}{y} - 1 \\ &= -\log\frac{\lambda x}{\lambda y} + \frac{\lambda x}{\lambda y} - 1 \\ d_{\mathrm{IS}}(x||y) &= d_{\mathrm{IS}}(\lambda x||\lambda y) \end{aligned}$$ ∎

This property of the IS divergence is particularly interesting for audio applications. Indeed, as many works on psychoacoustics and perception have derived ([Stevens, 1936] and [Warren, 1970]), the perception of sound intensity is proportional to the sound pressure level in decibel, hence proportional to the logarithm of the energy of the signal. This implies that for a system to catch signals that are ultimately important to a human ear, the components with relatively low energy need to be considered with care. When using the other popular distortion measures, namely the Euclidean (EUC) distance and the Kullback-Leibler (KL)

divergence, one can also easily show that [Févotte et al., 2009a]:

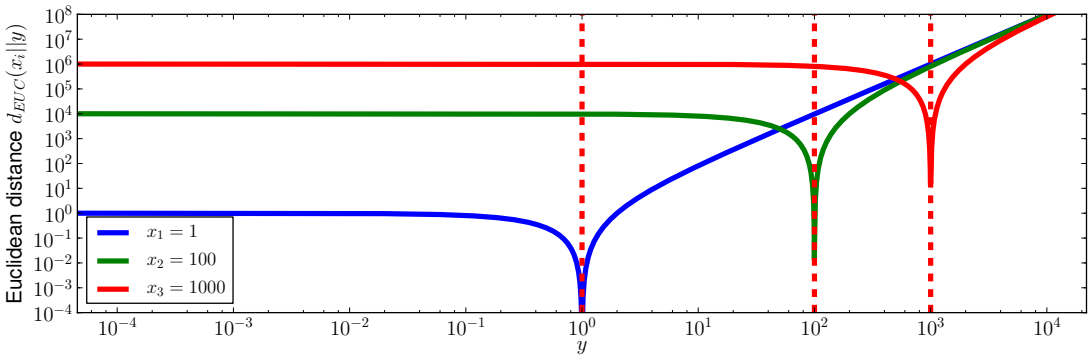
$$d_{\mathrm{EUC}}(\lambda x||\lambda y) = \frac{1}{2}(\lambda x - \lambda y)^2 = \lambda^2 d_{\mathrm{EUC}}(x||y) \tag{4.10}$$

$$d_{\mathrm{KL}}(\lambda x||\lambda y) = \lambda x \log \frac{\lambda x}{\lambda y} - \lambda x + \lambda y = \lambda d_{\mathrm{KL}}(x||y) \tag{4.11}$$

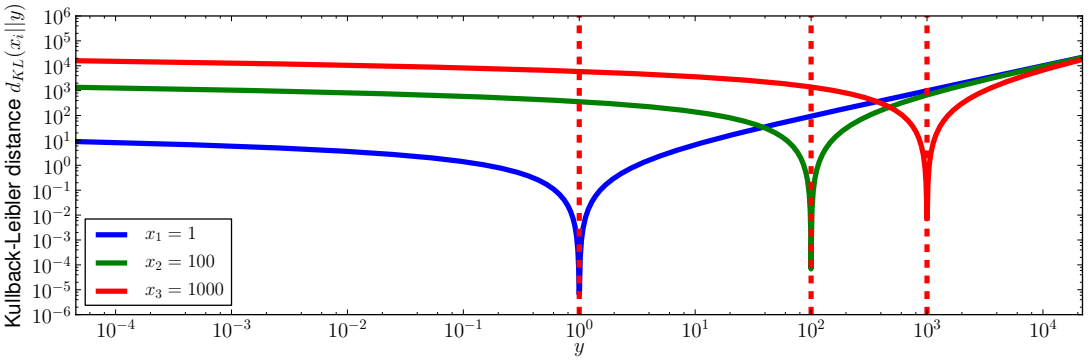
In the case of power STFT modelling as in Févotte et al. [2009a], for two frequency bins, the spectrum could exhibit a relatively low energy component x_1 with a high energy component $x_2 = \lambda x_1$, with $\lambda > 1$. Let us consider the interval $I_i^\epsilon = [y_i^{-\epsilon}, y_i^\epsilon]$ such that, for a given value ϵ of the cost function, $\forall y \in I_i^\epsilon$, $d(x_i||y) < \epsilon$. Depending on the chosen cost function d (EUC, KL or IS), we observe that the corresponding intervals sensibly vary. Indeed, as can be seen on Figure 4.1, for the Euclidean distance, for a given ϵ, on a logarithmic scale for y, or equivalently on a perceptive scale, the interval I_2^ϵ is much smaller than I_1^ϵ: the Euclidean distance allows more "perceptual" error for low energy components. This phenomenon is still present for the Kullback-Leibler divergence, but completely disappears for the Itakura-Saito divergence: the size on a logarithmic scale of the intervals I_i^ϵ does not depend on the value of x_i.

The scale invariance of the IS divergence therefore makes it an ideal divergence to work with for audio signal processing. It allows to give as much weight in estimating high energy components as estimating low energy ones. This can however also be a drawback, in some circumstances: the use of the IS divergence implies that any component of the processed signal has to be modelled in the covariance matrix of the Gaussian, even noise components, which would be implicitly ruled out in systems using the Euclidean or the Kullback-Leibler cost functions.

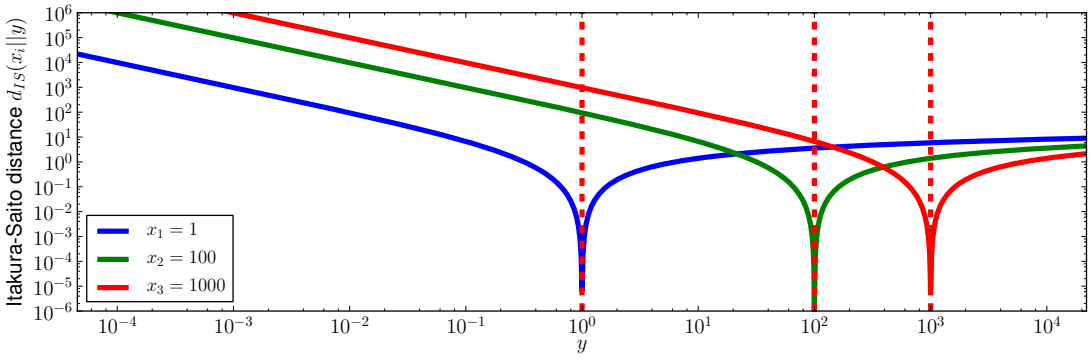
The second property is actually a burden from an estimation point of view: **the IS divergence, as used in our application, is not convex** over the whole parameter domain. There is in general no analytic solution for the optimisation problem, such that we need iterative (mostly gradient) algorithms to find the best parameter set. Most of the optimal gradient methods, such as the Newton gradient method (for instance in a very similar framework as in [Cardoso et al., 2008]), may therefore be bound to fail, since they often rely on a parabolic approximation of the cost function. The convexity assumption may be true if the initial solution given to these iterative algorithms is close enough to the actual solution, such that more elaborate initialisation strategies need to be involved when using the Newton algorithm. One should however note that practical solutions such as the multiplicative gradient method seem to lead to fairly good estimations.



(a) Euclidean (EUC) distance



(b) Kullback-Leibler (KL) divergence



(c) Itakura-Saito (IS) divergence

Figure 4.1: Comparison between the Euclidean, Kullback-leibler and Itakura-Saito divergences, with respect to scale changes.

Chapter 5

Parameter and sequence estimation

In this chapter, some practical algorithms and systems are proposed. From the estimation point of view, it may indeed not be possible to directly use the models of Chapter 3. In Section 5.1, the principle for the adopted approximations is first given, and the five proposed systems to estimate the sequences of fundamental frequencies F0, the sequence of melody notes and the separated signals are presented.

In Section 5.2, we then describe the algorithms used to estimate the parameters intervening in the IMM and SIMM model. In Section 5.3, the Expectation-Maximisation (EM) algorithms necessary to estimate the parameters of the GSMM and SGSMM models are derived. At last, in Section 5.4, the strategies and algorithms to track the desired F0 and note sequences are presented.

5.1 Transcription and separation as statistical estimation

The targetted applications, the melody extraction and the separation of the lead instrument, can be expressed as statistical estimation problems.

The general model as stated in Chapter 3 gives the joint likelihood of the observations and all the variables $p(\mathbf{X}, Z^{\Phi}, Z^{F_0}, E)$ (Maximum Likelihood framework - (S)GSMM) or $p(\mathbf{X}, \mathbf{\Theta}^{\text{IMM}}, Z^{\Phi}, Z^{F_0}, E)$ (Bayesian framework - (S)IMM).

However, jointly estimating all the parameters and sequences involved in these likelihoods may be computationally too complicated. Several approximations have been proposed [Durrieu et al., 2008a]: the GSMM framework can be replaced by the IMM, which notably avoids the need for an EM algorithm. The algorithm flow has also been decomposed into several steps which can be held independently, with good results. In this section, the theoretical motivations for the approximations made in [Durrieu et al., 2008a], [Durrieu et al., 2009a] and [Weil et al., 2009b] are first discussed. Then the five different proposed systems aiming at estimating, transcribing and separating the main melody are presented.

5.1.1 Estimation by Maximum Likelihood (ML) and Maximum A Posteriori (MAP) principle

Depending on the application, the above models may be used to estimate the separated signals $\widehat{\mathbf{V}}$ and $\widehat{\mathbf{M}}$ knowing the mixture $\mathbf{X} = \mathbf{V} + \mathbf{M}$ or to estimate the sequences Z and E in order to perform a transcription of the melody. Any of these estimations are made through the joint likelihoods $p(\mathbf{X}, Z^{\Phi}, Z^{F_0}, E)$ or $p(\mathbf{X}, \mathbf{\Theta}^{\text{IMM}}, Z^{\Phi}, Z^{F_0}, E)$.

However, as was already discussed in Section 3.3.3 and 3.3.4, finding the best path for sequences Z^{F_0} or E would strictly speaking require marginalising these likelihoods over the other variables (parameters and sequences). This may sometimes be possible, sometimes requiring techniques such as Monte Carlo methods [Cemgil and Kappen, 2003]. However, for our applications, we have chosen some strategies that approximate the optimal result.

the main principle is to estimate all the variables by **gradually increasing the complexity of the model**. In practice, this means that, in the graphical models on Figures 3.14 and 3.17, we start with the first layers, *i.e.* the parameter layer $\mathbf{\Theta}$ or the physical one Z, discarding all the other layers, especially the temporal dependencies. This basically allows to estimate the parameters in $\mathbf{\Theta}$ such that they fit the signal, without temporal or musical constraints. A second layer can then be added, and so forth, so as to track the different desired sequences.

The systems we propose in the following sections implement this principle at different scales. We however also provide a system (system F-III, see Section 5.1.2 for details) which aims at including more constraints directly from the first estimation steps. It seems indeed possible to design an algorithm that would directly take into account all the dependencies of the model, even for the musicological layer. Considering the results so far obtained by our preliminary tests with that system, the benefit of such an integrated system is not obvious, and will be discussed in Chapter 6.

5.1.2 Predominant fundamental frequency estimation

For this task, only the model up to the physical model is necessary, and the musicological layer is discarded. The problem can be addressed by estimating the sequence Z^{F_0} that maximizes the posterior likelihood $p(Z^{F_0}|\mathbf{X})$. However, as we discussed earlier, it is easier, and to some extent more meaningful to estimate the sequence Z^{F_0} along with the corresponding parameter set $\mathbf{\Theta}$ and the sequence Z^{Φ}:

$$\widehat{\mathbf{\Theta}}, \widehat{Z}^{F_0}, \widehat{Z}^{\Phi} = \underset{\mathbf{\Theta}, Z^{F_0}, Z^{\Phi}}{\arg\max}\, p(\mathbf{X}, Z^{F_0}, Z^{\Phi}, \mathbf{\Theta}) \tag{5.1}$$

The systems proposed and studied to address this estimation are given below. All of them, except system F-III, rely on the aforementioned approximation. First the frame level parameterization is considered, without the temporal aspect. Then the temporal aspect from the physical layer is included.

All three systems use the voiced dictionary $\mathbf{W}^{F_0}$, without the additional unvoiced element. This latter element is included in $\mathbf{W}^{F_0}$ only for the source separation system described below.

F-I **Predominant F0 estimation with the (S)GSMM:** first estimate the parameters $\mathbf{\Theta}^{\text{GSMM}} = \{\mathbf{B}, \mathbf{W}^{\Phi}, \mathbf{H}^{M}, \mathbf{W}^{M}\}$ or $\mathbf{\Theta}^{\text{SGSMM}} = \{\mathbf{B}, \mathbf{H}^{\Gamma}, \mathbf{H}^{M}, \mathbf{W}^{M}\}$, maximizing the frame level probability, corresponding to the (S)GSMM, i.e. without temporal constraint. Then use the Viterbi algorithm in order to retrieve the corresponding optimal path for the sequence (Z^{Φ}, Z^{F_0}), with the temporal structure proposed in Section 3.3.3. In other terms, in equations, the estimation process writes, chronologically and with I the number of iterations for the first estimation round[1]:

[1] As will be seen later, all the algorithms presented in this work are iterative optimisation algorithms.

1. Parameter estimation:

$$(\widehat{\mathbf{\Theta}}^{(\mathrm{S})\mathrm{GSMM}})^{(i)} = \underset{\mathbf{\Theta}^{(\mathrm{S})\mathrm{GSMM}}}{\arg\max}\, E\left[\log p(\mathbf{X}, Z^{F_0}, Z^{\Phi}; \mathbf{\Theta}^{(\mathrm{S})\mathrm{GSMM}})|\mathbf{X}; (\mathbf{\Theta}^{(\mathrm{S})\mathrm{GSMM}})^{(i-1)}\right] \quad \text{for } i \in [1, I] \tag{5.2}$$

2. Sequence tracking:

$$\widehat{Z}^{F_0}, \widehat{Z}^{\Phi} = \underset{Z^{F_0}, Z^{\Phi}}{\arg\max} \prod_n p(\mathbf{x}_n|Z_n^{F_0}, Z_n^{\Phi}; \widehat{\mathbf{\Theta}}^{(\mathrm{S})\mathrm{GSMM}}) \times p(Z_1^{F_0}, Z_1^{\Phi}) \prod_{n>1} p(Z_n^{F_0}, Z_n^{\Phi}|Z_{n-1}^{F_0}, Z_{n-1}^{\Phi}) \tag{5.3}$$

The estimation details for Equation (5.2) and the corresponding algorithm are given in Section 5.3. The Viterbi algorithm that allows to estimate the desired sequences Z^{F_0} and Z^{Φ}, from Equation (5.3), is discussed in Section 5.4.1.

F-II **Predominant F0 estimation with the (S)IMM:** as for system F-I, the parameters for the frame layer are first estimated without temporal constraint. Then the Viterbi algorithm is run, assuming a MAP framework, with correctly defined conditional probability for the parameter set $\mathbf{\Theta}^{(\mathrm{S})\mathrm{IMM}}$ conditionally upon the sequence (Z^{Φ}, Z^{F_0}). The estimation process then follows:

1. Parameter estimation:

$$\widehat{\mathbf{\Theta}}^{(\mathrm{S})\mathrm{IMM}} = \underset{\mathbf{\Theta}^{(\mathrm{S})\mathrm{IMM}}}{\arg\max}\, p(\mathbf{X}|\mathbf{\Theta}^{(\mathrm{S})\mathrm{IMM}}) \tag{5.4}$$

2. Sequence tracking:

$$\widehat{Z}^{F_0} = \underset{Z^{F_0}}{\arg\max} \prod_n p(\mathbf{x}_n|\widehat{\mathbf{\Theta}}_n^{(\mathrm{S})\mathrm{IMM}}) p(\widehat{\mathbf{\Theta}}_n^{(\mathrm{S})\mathrm{IMM}}|Z_n^{F_0}) \times p(Z_1^{F_0}) \prod_{n>1} p(Z_n^{F_0}|Z_{n-1}^{F_0}) \tag{5.5}$$

The estimation of the parameters without temporal constraint of Equation (5.4) is done through multiplicative updating rules inspired by NMF: the link with NMF techniques was given in Chapter 4, and the multiplicative updating rules are given in Section 5.2. Note that all the derivations are also given in Appendix B.1.

Once the parameters are estimated, the optimal path Z^{F_0} operating the trade-off between the energy of the leading instrument and the physical continuity of the fundamental frequency is estimated: the Viterbi algorithm described in Section 5.4.1 "decodes" the sequence as defined by Equation (5.5). For system F-II, we assume that the probability product $p(\mathbf{x}_n|\widehat{\mathbf{\Theta}}_n^{(\mathrm{S})\mathrm{IMM}}) p(\widehat{\mathbf{\Theta}}_n^{(\mathrm{S})\mathrm{IMM}}|Z_n^{F_0} = u)$ is proportional to $h_{un}^{F_0}$. Indeed, this way, the first term in Equation (5.5) corresponds to the energy of the source pitch u in the mixture, while the second term, $p(Z_1^{F_0}) \prod_{n>1} p(Z_n^{F_0}|Z_{n-1}^{F_0})$, corresponds to the sequence evolution constraint, defined in Section 3.3.3. The reason why such a scheme allows to take into account the energy and the continuity at the same time can easily be seen: when the evolution is not constrained, i.e. $p(Z_n^{F_0}|Z_{n-1}^{F_0}) \propto 1$,

then the best path is such that, for any frame n, $\widehat{Z}^{F_0} = \arg\max_u h_{un}^{F_0}$, namely the fundamental frequency maximizing the energy. On the contrary, without the energy information, the optimal path would be a constant state. The HMM chain allows to retrieve a path which is "between" these two extreme solutions. A more formal motivation for such a choice is given in Section 5.2.3.

F-III **Predominant F0 estimation with the HMM:** for this system, we directly consider the HMM framework including the physical layer for the source/filter sequences. This model is denoted the **Hidden Markov - Gaussian Scaled Mixture Model (HM-GSMM)**. The criterion is the same as Equation 5.2, and the sequence is estimated exactly with the same Equation 5.3. The estimation process is therefore the same as system F-I:

1. Parameter estimation:

$$(\widehat{\boldsymbol{\Theta}}^{(\mathrm{S})\mathrm{GSMM}})^{(i)} = \underset{\boldsymbol{\Theta}^{(\mathrm{S})\mathrm{GSMM}}}{\arg\max}\, E[\log p(\mathbf{X}, Z^{F_0}, Z^{\Phi}; \boldsymbol{\Theta}^{(\mathrm{S})\mathrm{GSMM}})|\mathbf{X}; (\boldsymbol{\Theta}^{(\mathrm{S})\mathrm{GSMM}})^{(i-1)}] \quad \text{for } i \in [1, I] \tag{5.6}$$

2. Sequence tracking:

$$\widehat{Z}^{F_0}, \widehat{Z}^{\Phi} = \underset{Z^{F_0}, Z^{\Phi}}{\arg\max} \prod_n p(\mathbf{x}_n|Z_n^{F_0}, Z_n^{\Phi}; \widehat{\boldsymbol{\Theta}}^{(\mathrm{S})\mathrm{GSMM}}) \times p(Z_1^{F_0}, Z_1^{\Phi}) \prod_{n>1} p(Z_n^{F_0}, Z_n^{\Phi}|Z_{n-1}^{F_0}, Z_{n-1}^{\Phi}) \tag{5.7}$$

The difference between F-I and F-III, although not visible from Equations (5.2) and (5.6) are discussed in Section 5.3.3. The main difference appears, technically, when computing the posterior probability. Indeed, with the (S)GSMM, one only considers the posterior probabilities $p(Z_n^{\Phi} = k, Z_n^{F_0} = u|\mathbf{x}_n)$, while in the HMM framework, the posterior probability cannot be reduced to the GSMM form, and one needs to compute the posterior probability, conditionally upon the whole sequence of observation $\mathbf{X}$: $p(Z_n^{\Phi} = k, Z_n^{F_0} = u|\mathbf{X})$.

For this estimation, the **STFT resolution limits** are partly overcome by the use of the spectral combs (as represented in Figures 3.7 and 3.8). Indeed, the *posterior* probability of a given fundamental frequency is determined thanks to the whole spectral range, and not only thanks to the corresponding STFT frequency bin. This technique however also has its limits: when the peaks of the spectral comb are not distinguishable anymore, the resulting estimation for Z^{F_0} becomes unreliable. This may happen for low fundamental frequencies, when the frequency of the first harmonic f_1 is within the resolution of the STFT from the fundamental frequency f_0. For a cosine window or a Hann window, at a sampling rate of 44100Hz, with a size of 2048 samples (around 46ms), the bandwidth of the main lobe of the Fourier transform is around 40Hz. 40Hz is therefore, in this case, the lowest value we should use for the possible fundamental frequencies.

5.1.3 Musical (notewise) transcription of the main melody

For this task, the model including the sequence of notes is necessary, and the estimation concerns all the following quantities:

$$\widehat{\mathbf{\Theta}}, \widehat{Z}^{F_0}, \widehat{Z}^{\Phi}, \widehat{E} = \underset{\mathbf{\Theta}, Z^{F_0}, Z^{\Phi}}{\arg\max}\, p(\mathbf{X}, \mathbf{\Theta}, Z^{F_0}, Z^{\Phi}, E) \tag{5.8}$$

The proposed system to address this problem is based on the IMM for the frame level model, and the estimation is again done one layer after the other. As for the above systems, F-I, F-II and F-III, the dictionary $\mathbf{W}^{F_0}$ only includes spectral combs, and not the additional unvoiced element.

MUS-I **Musical transcription of the main melody with IMM:** first, the parameters are estimated thanks to the IMM frame level model, without any temporal contraint. Then, the (sub-)optimal sequence of fundamental frequencies $\widetilde{Z}^{F_0}$ is computed with the physical layer temporal constraints. At last, this sequence is used as initial F0 candidates to explore the possible note sequences E, which narrows down the space of potential sequences by eliminating as many unlikely hypothesis as possible. The process then writes:

1. Parameter estimation:

$$\widehat{\mathbf{\Theta}}^{(\mathrm{S})\mathrm{IMM}} = \underset{\mathbf{\Theta}^{(\mathrm{S})\mathrm{IMM}}}{\arg\max}\, p(\mathbf{X}|\mathbf{\Theta}^{(\mathrm{S})\mathrm{IMM}}) \tag{5.9}$$

2. Sequence tracking and candidate F0 selection:

$$\widetilde{Z}^{F_0} = \underset{Z^{F_0}}{\arg\max} \prod_n p(\mathbf{x}_n|\widehat{\mathbf{\Theta}}_n^{(\mathrm{S})\mathrm{IMM}}) p(\widehat{\mathbf{\Theta}}_n^{(\mathrm{S})\mathrm{IMM}}|Z_n^{F_0}) \times p(Z_1^{F_0}) \prod_{n>1} p(Z_n^{F_0}|Z_{n-1}^{F_0}) \tag{5.10}$$

3. Note sequence tracking:

$$\widehat{E}, \widehat{Z}^{F_0} = \underset{E, Z^{F_0}}{\arg\max}\, p(\mathbf{X}|\widehat{\mathbf{\Theta}}^{(\mathrm{S})\mathrm{IMM}}) p(\widehat{\mathbf{\Theta}}^{(\mathrm{S})\mathrm{IMM}}|Z^{F_0}) p(Z^{F_0}|E) p(E) \tag{5.11}$$

The estimation Equation (5.9) is identical to the IMM estimation, and done with the multiplicative algorithm developed in Section 5.2. Again, as for the IMM of system F-II, Equation (5.10) is done by assuming that the probability product $p(\mathbf{x}_n|\widehat{\mathbf{\Theta}}_n^{(\mathrm{S})\mathrm{IMM}}) p(\widehat{\mathbf{\Theta}}_n^{(\mathrm{S})\mathrm{IMM}}|Z_n^{F_0} = u)$ is proportional to $h_{un}^{F_0}$. At last, the sequence decoding of Equation (5.11) is described in Section 5.4.2, with the beam search strategy already proposed by Vincent [2006] for a similar application. This system has been proposed in [Weil et al., 2009b].

5.1.4 Leading instrument / accompaniment separation

This task is slightly different since we are interested in estimating the separated signals $\widehat{\mathbf{V}}$ and $\widehat{\mathbf{M}}$. The least square estimator is the conditional expectation of $\mathbf{V}$ (or $\mathbf{M}$) given $\mathbf{X}$. As discussed later in Section 6.2, in practice, the signals' STFTs are obtained by Wiener masking, and then the desired time domain signals are computed by overlap-add procedure:

this is called an adpative Wiener filtering [Benaroya et al., 2006]. The expression for the Wiener estimator is:

$$\widehat{\mathbf{V}} = \frac{\mathbf{S}^V}{\mathbf{S}^V + \mathbf{S}^M} \bullet \mathbf{X}$$
$$\widehat{\mathbf{M}} = \frac{\mathbf{S}^M}{\mathbf{S}^V + \mathbf{S}^M} \bullet \mathbf{X}$$

where $\mathbf{S}^V$ and $\mathbf{S}^M$ respectively are the (estimated) variances for the leading instrument and for the accompaniment. The division operations are meant element by element, and $\bullet$ represent the Hadamard product. They depend on the parameter set $\mathbf{\Theta}$, and we need an estimation of this set in order to proceed to the actual separation. Once again, this task only requires frame-wise estimations, and the model which is used discards the musicological layer:

$$\hat{\mathbf{\Theta}}, \hat{Z}^{F_0}, \hat{Z}^{\Phi} = \underset{\mathbf{\Theta}, Z^{F_0}, Z^{\Phi}}{\arg\max}\, p(\mathbf{X}, Z^{F_0}, Z^{\Phi}, \mathbf{\Theta}) \tag{5.12}$$

Note that the needed estimators are formally the same for the source separation task (Eq. (5.12)) and for the predominant fundamental frequency estimation (Eq. (5.1)). This clearly shows the link between these two applications and the possibility of jointly estimating and separating. This unified framework is quite novel since it allows to estimate the fundamental frequencies on a rather fine scale, while being able to separate the analyzed sounds. In [Vincent, 2004], the authors consider a similar framework, but the states for the musical notes correspond to the Western music scale.

SEP-I **Leading instrument / accompaniment separation:** To achieve the source separation task, the three previously proposed frameworks for fundamental frequency estimation, F-I, F-II or F-III, can be equally used. The system F-II is the easiest to configure and adapt, and provides a faster algorithm than F-I or F-III. F-II is therefore a good choice and is the basis for the systems published in [Durrieu et al., 2008b], [Durrieu et al., 2009a] and [Durrieu et al., 2009b]. A first estimation round gives the parameter set $\widehat{\mathbf{\Theta}}^{(\mathrm{S})\mathrm{IMM}}$, without constraints. From this first estimate, and especially from the amplitudes for the source part $\mathbf{H}^{F_0}$, the optimal fundamental frequency sequence $\widehat{Z}^{F_0}$ is extracted. A second estimation round with a specific initial matrix $\widetilde{\mathbf{H}}^{F_0}$ then allows to obtain a parameter set that fits the estimated melody line, thus better fitting the mixture signal. The separation process flow is:

1. Parameter estimation, first round:

$$\bar{\mathbf{\Theta}}^{(\mathrm{S})\mathrm{IMM}} = \underset{\mathbf{\Theta}^{(\mathrm{S})\mathrm{IMM}}}{\arg\max}\, p(\mathbf{X}|\mathbf{\Theta}^{(\mathrm{S})\mathrm{IMM}}) \tag{5.13}$$

2. Sequence tracking (melody estimation):

$$\widehat{Z}^{F_0} = \underset{Z^{F_0}}{\arg\max} \prod_n p(\mathbf{x}_n|\bar{\mathbf{\Theta}}_n^{(\mathrm{S})\mathrm{IMM}}) p(\bar{\mathbf{\Theta}}_n^{(\mathrm{S})\mathrm{IMM}}|Z_n^{F_0}) \times p(Z_1^{F_0}) \prod_{n>1} p(Z_n^{F_0}|Z_{n-1}^{F_0}) \tag{5.14}$$

3. Parameter estimation, second round:

$$\widehat{\mathbf{\Theta}}^{(\mathrm{S})\mathrm{IMM}} = \underset{\mathbf{\Theta}^{(\mathrm{S})\mathrm{IMM}}}{\arg\max}\, p(\mathbf{X}|\mathbf{\Theta}^{(\mathrm{S})\mathrm{IMM}}) \text{ with initial amplitudes } \widetilde{\mathbf{H}}^{F_0} \tag{5.15}$$

4. Computing the separated signals with adaptive Wiener filters (given below for the SIMM):

$$\widehat{\mathbf{V}} = \frac{\mathbf{W}^{\Gamma}\widehat{\mathbf{H}}^{\Gamma}\widehat{\mathbf{H}}^{\Phi} \bullet \mathbf{W}^{F_0}\widehat{\mathbf{H}}^{F_0}}{\mathbf{W}^{\Gamma}\widehat{\mathbf{H}}^{\Gamma}\widehat{\mathbf{H}}^{\Phi} \bullet \mathbf{W}^{F_0}\widehat{\mathbf{H}}^{F_0} + \widehat{\mathbf{W}}^{M}\widehat{\mathbf{H}}^{M}} \bullet \mathbf{X}$$
$$\text{and } \widehat{\mathbf{M}} = \frac{\widehat{\mathbf{W}}^{M}\widehat{\mathbf{H}}^{M}}{\mathbf{W}^{\Gamma}\widehat{\mathbf{H}}^{\Gamma}\widehat{\mathbf{H}}^{\Phi} \bullet \mathbf{W}^{F_0}\widehat{\mathbf{H}}^{F_0} + \widehat{\mathbf{W}}^{M}\widehat{\mathbf{H}}^{M}} \bullet \mathbf{X} \tag{5.16}$$

The first two steps, Equation (5.13) and (5.14), are the same as for system F-II (Equations (5.4) and (5.5)). The last step Equation (5.15) is a re-estimation of the parameter set $\mathbf{\Theta}^{(\mathrm{S})\mathrm{IMM}}$, with a hard constraint on the amplitude coefficients for the source part of the lead instrument: $\widetilde{\mathbf{H}}^{F_0}$ is such that any coefficient which is not within one semitone from the estimated melody is set to 0:

$$\widetilde{h}^{F_0}_{un} = 0 \text{ if } |12\log_2 \mathcal{F}(u) - 12\log_2 \mathcal{F}(\widehat{Z}^{F_0}_n)| > \frac{1}{4} \tag{5.17}$$

which means that source elements u whose frequency $\mathcal{F}(u)$ is outside a scope of one semitone from the frequency $\mathcal{F}(\widehat{Z}^{F_0}_n)$ of the estimated melody are de-activated on frame n. Equations (5.16) give the formulas to compute the estimated separated lead instrument and accompaniment STFTs $\widehat{\mathbf{V}}$ and $\widehat{\mathbf{M}}$, which are trivially computed with the parameters of $\widehat{\mathbf{\Theta}}^{\mathrm{SIMM}} = \{\widehat{\mathbf{H}}^{\Gamma}, \widehat{\mathbf{H}}^{\Phi}, \widehat{\mathbf{H}}^{F_0}, \widehat{\mathbf{W}}^{M}, \widehat{\mathbf{H}}^{M}\}$ and the fixed spectral shape dictionaries $\mathbf{W}^{\Gamma}$ and $\mathbf{W}^{F_0}$. To obtain the time domain signals, the STFTs are inverted thanks to a classical Overlap and Add (OLA) procedure.

The source part dictionary $\mathbf{W}^{F_0}$ contains only the voiced elements during the first step. During the second step, one can add the unvoiced element, described in Section 3.3.2.1. However, in our system, it was preferred to insert it after the second round of estimation, and update the parameters during a third estimation round, as explained in Section 6.2.4.3.

5.1.5 Systems summary

In the next sections, the estimations corresponding to the different equations provided for each system are presented. Some estimation methods may be shared among several systems. The proposed systems are enumerated in Table 5.1, along with the model which is used ((S)GSMM or (S)IMM) for the parameter estimation, the sequence tracking method and the articles in which they were published.

Table 5.1: Proposed systems and their characteristics. The numbers beside each entry corresponds to the section of this document where the relevant algorithms are given.

System	Model for parameter estimation	Tracking method	Article
F-I	(S)GSMM 5.3.2	Viterbi 5.4.1	[Durrieu et al., 2009c, 2010]
F-II	(S)IMM 5.2.2	Viterbi 5.4.1	[Durrieu et al., 2008a, 2009c, 2010]
F-III	HM-(S)GSMM 5.3.3	Viterbi 5.4.1	*unpublished*
MUS-I	(S)IMM 5.2.2	Viterbi 5.4.1 Beam search 5.4.2	[Weil et al., 2009b]
SEP-I	(S)IMM 5.2.2	Viterbi 5.4.1	[Durrieu et al., 2009a,b]

5.2 IMM and SIMM: Multiplicative gradient algorithm

It is easier to describe the estimation process for the IMM before the one needed for the GSMM: indeed, the multiplicative gradient approach is common to these two model estimations. However, for the GSMM, this gradient approach occurs within a Generalized Expectation-maximization (GEM) algorithm, while for the IMM, the estimation algorithm directly is a gradient descent approach, without having to define an auxiliary function other than the posterior or joint likelihood, as is the case for the GEM algorithm (see Section 5.3 for details).

First, the derivations for the parameter estimation of the frame-wise models IMM and SIMM are presented: the MAP criterion to be maximized is derived in Section 5.2.1 and the corresponding multiplicative gradient updating rules are given in Section 5.2.2. In Section 5.2.3 the issue of estimating the desired fundamental frequency sequence is addressed: the use of the Viterbi algorithm on the amplitude coefficients associated with each pitch is described (the Viterbi algorithm in itself is described in Section 5.4.1). The monopitch assumption, inherent in the GSMM but discarded in the SIMM, is also considered with the inclusion of the prior distribution on the amplitudes introduced in Section 3.4.3.

In this section, $\boldsymbol{\Theta}$ refers to $\boldsymbol{\Theta}^{(\mathrm{S})\mathrm{IMM}}$.

5.2.1 Maximum A Posteriori (MAP) Criterion for the IMM/SIMM

For systems F-II, MUS-I and Sep-I, the parameters are estimated through Maximum Likelihood, as stated in Equations (5.4), (5.9), (5.13) and (5.15), with the re-estimation of the parameters needed in SEP-I, which formally is exactly the same process, except for the initialisation of the parameters. However, in this section we present a more general MAP formulation for the parameter estimation: with uninformative priors on the parameters, this is equivalent to ML estimation. With priors on $\boldsymbol{\Theta}$, as proposed in Section 3.4.3, the MAP criterion derived in this section can still be used, with the corresponding consequences in the updating rules explained in Section 5.2.3.

For the IMM and SIMM, since there is no hidden state, the MAP criterion is directly chosen as the logarithm of the posterior probability of the parameters, given the observa-

tion, or equivalently, the joint log-likelihood of the observation and the parameters:

$$C_{\text{IMM}}(\boldsymbol{\Theta}) = \log p(\mathbf{X}, \boldsymbol{\Theta})$$
$$C_{\text{IMM}}(\boldsymbol{\Theta}) = \sum_{f,n} \log \frac{|x_{fn}|}{\pi s_{fn}^{(\text{S})\text{IMM}}} - \frac{|x_{fn}|^2}{s_{fn}^{(\text{S})\text{IMM}}} + \log p(\boldsymbol{\Theta}) \tag{5.18}$$

The expression of the variance $s_{fn}^{(\text{S})\text{IMM}}$ in Equation (5.18) is given by Table 3.2 and depends on $\boldsymbol{\Theta}$, such that:

$$\mathbf{S}^{(\text{S})\text{IMM}} = \left(\mathbf{W}^{\Phi}\mathbf{H}^{\Phi}\right) \bullet \left(\mathbf{W}^{F_0}\mathbf{H}^{F_0}\right) + \mathbf{W}^{M}\mathbf{H}^{M} \tag{5.19}$$

where $\mathbf{W}^{\Phi} = \mathbf{W}^{\Gamma}\mathbf{H}^{\Gamma}$ for the SIMM. Estimating $\boldsymbol{\Theta}$ by MAP estimation boils down to finding the $\widehat{\boldsymbol{\Theta}}$ which maximizes the criterion Equation (5.18):

$$\widehat{\boldsymbol{\Theta}} = \arg\max_{\boldsymbol{\Theta}} C_{\text{IMM}}(\boldsymbol{\Theta}) \tag{5.20}$$

The criterion (5.18) suffers from several indeterminacies. Scale indeterminacy arises with the distribution of the energy between the dictionary matrices $\mathbf{W}^{\Gamma}$, $\mathbf{W}^{F_0}$ and $\mathbf{W}^{M}$ and their corresponding amplitude matrices $\mathbf{H}^{\Gamma}$ (and $\mathbf{H}^{\Phi}$), $\mathbf{H}^{F_0}$ and $\mathbf{H}^{M}$, as well as between $\mathbf{H}^{\Phi}$ and $\mathbf{H}^{F_0}$: we solve this problem by normalizing the columns of $\mathbf{W}^{\Gamma}$, $\mathbf{W}^{F_0}$, $\mathbf{W}^{M}$, $\mathbf{H}^{\Gamma}$ and $\mathbf{H}^{\Phi}$.

In the proposed systems, a multiplicative gradient approach has been investigated to achieve the estimation in Equation (5.20). In the following sections, the updating rules for the ML estimation (or MAP with uninformative priors) using the criterion in Equation (5.18) are given, with the algorithm that is the basis for articles such as [Durrieu et al., 2008a], [Durrieu et al., 2009b] or [Durrieu et al., 2010], with detailed calculations in Appendix B.1.

5.2.2 IMM/SIMM updating rules

The MAP estimate of $\boldsymbol{\Theta}$ can be classically derived by finding a parameter set $\widehat{\boldsymbol{\Theta}}$ such that all the partial derivatives of the criterion C_{IMM} with respect to all the elements of $\boldsymbol{\Theta}$, evaluated at $\widehat{\boldsymbol{\Theta}}$, are equal to 0 (one of the first order Karush, Kuhn and Tucker's conditions - KKT conditions [Kuhn and Tucker, 1951]). If the criterion is "regular" enough, then there is only one such set, otherwise, there may be several other "sub-optimal" solutions to the problem, due to potential local maxima in C_{IMM}.

In order to find one such parameter set, we apply a method now classical in NMF related algorithms: a multiplicative gradient approach. The idea of such an approach is identical with the principle of (additive) gradient ascent. The parameters in $\boldsymbol{\Theta}$ are updated such that the previous values are changed in the same direction as the gradient, evaluated at those previous values.

Let $\theta \in \boldsymbol{\Theta}$ be one of the parameters to be estimated. The partial derivative of the

criterion in Equation (5.18) with respect to this parameter is:

$$\frac{\partial C_{\text{IMM}}(\boldsymbol{\Theta}))}{\partial\theta} = \underbrace{-\left(\sum_{fn}\frac{\partial s_{fn}^{(\text{S})\text{IMM}}(\theta)}{\partial\theta}\frac{1}{s_{fn}^{(\text{S})\text{IMM}}(\theta)} + \left[\frac{\partial}{\partial\theta}\log p(\boldsymbol{\Theta})\right]^{-}\right)}_{\nabla^{-}} + \underbrace{\left(\sum_{fn}\frac{\partial s_{fn}^{(\text{S})\text{IMM}}(\theta)}{\partial\theta}\frac{|x_{fn}|^2}{s_{fn}^{(\text{S})\text{IMM}}(\theta)^2} + \left[\frac{\partial}{\partial\theta}\log p(\boldsymbol{\Theta})\right]^{+}\right)}_{\nabla^{+}} \tag{5.21}$$

where ∇^+ and ∇^- are positive terms, and with:

$$\frac{\partial}{\partial\theta}\log p(\boldsymbol{\Theta}) = \left[\frac{\partial}{\partial\theta}\log p(\boldsymbol{\Theta})\right]^{+} - \left[\frac{\partial}{\partial\theta}\log p(\boldsymbol{\Theta})\right]^{-}$$

where the terms in the right hand are all positive terms.

In Appendix B.1.1, the multiplicative gradient principle is further developed and motivated. Note that in the Appendix, the derivations are done with the Itakura-Saito divergence, such that the results must be inverted as concerns the terms ∇^+ and ∇^-. For the log-likelihood criterion, the updating rules should therefore be defined as:

$$\theta^{(i+1)} \leftarrow \theta^{(i)}\left(\frac{\nabla^{+}}{\nabla^{-}}\right)^{\omega} \tag{5.22}$$

where ∇^+ and ∇^- are computed with the values of the parameters in $\boldsymbol{\Theta}^{(i-1)}$, and $\omega \in]0,2[$ is a parameter allowing to control the speed of convergence of the algorithm, hence holding the same role as the step size for additive gradient approaches[2]. Following the above relation, the updating rules for all the parameters to be estimated, as explicitly mentioned in Table 3.2, can be easily derived. The details are given in Appendix B.1.2. The updating rules are summed up in Algorithm 5.1 for the IMM, and in Algorithm 5.2 for the SIMM, where for simplicity ω was set to 1.

In these algorithms, the order of update for the parameters was arbitrarily set: first $\mathbf{H}^{F_0}$, $\mathbf{H}^{\Phi}$, $\mathbf{H}^{M}$, $\mathbf{W}^{\Phi}$ and $\mathbf{W}^{M}$. Intuitively, this allows the parameters for the main instrument to adapt to the signal first, hence avoiding to leave some of the signal of interest in the accompaniment too early in the estimation. Except otherwise mentioned, the initial set of parameters for the estimation algorithm is randomly drawn.

Note also that in both algorithms as presented in this section, there is no use of priors for the parameter set, such that the criterion actually is a ML criterion, and the updating rules are obtained with the ∇ terms as in Equation (5.21), with $\frac{\partial}{\partial\theta}\log p(\boldsymbol{\Theta}) = 0$. In Section 5.2.3, the interpretation of the estimated parameters and their use to infer the fundamental frequency path Z^{F_0} are discussed, along with the updating rules corresponding to the prior on the amplitudes $\mathbf{H}^{F_0}$ as defined in Section 3.4.3, as an example of constraints on the parameters.

Figure 5.1 gives an example of the decomposition one obtains with the parameters of the SIMM. As expected, the matrix $\mathbf{H}^{F_0}$ does not exhibit as much sparsity as desired, and

[2]See [Badeau et al., 2009] for details on the values of ω allowing convergence of NMF multiplicative gradient approaches.

Algorithm 5.1 Updating rules for the IMM:
Estimating $\mathbf{\Theta}^{\mathrm{IMM}} = \{\mathbf{W}^{\Phi}, \mathbf{H}^{\Phi}, \mathbf{H}^{F_0}, \mathbf{W}^{M}, \mathbf{H}^{M}\}$

for $i \in [1, I]$ **do**

- Leading instrument source parameters:

$$\mathbf{H}^{F_0} \leftarrow \mathbf{H}^{F_0} \bullet \frac{(\mathbf{W}^{F_0})^T \mathbf{P}^{F_0}}{(\mathbf{W}^{F_0})^T \mathbf{Q}^{F_0}}$$

$$\text{where} \begin{cases} \mathbf{P}^{F_0} &= |\mathbf{X}|^2 \bullet (\mathbf{W}^{\Phi}\mathbf{H}^{\Phi})/(\mathbf{S}^{\mathrm{IMM}})^2 \\ \mathbf{Q}^{F_0} &= (\mathbf{W}^{\Phi}\mathbf{H}^{\Phi})/\mathbf{S}^{\mathrm{IMM}} \end{cases}$$

- Leading instrument filter parameters:

$$\mathbf{H}^{\Phi} \leftarrow \mathbf{H}^{\Phi} \bullet \frac{(\mathbf{W}^{\Phi})^T \mathbf{P}^{\Phi}}{(\mathbf{W}^{\Phi})^T \mathbf{Q}^{\Phi}}$$

$$\mathbf{W}^{\Phi} \leftarrow \mathbf{W}^{\Phi} \bullet \frac{\mathbf{P}^{\Phi} (\mathbf{H}^{\Phi})^T}{\mathbf{Q}^{\Phi} (\mathbf{H}_{\Phi})^T}$$

$$\text{where} \begin{cases} \mathbf{P}^{\Phi} &= |\mathbf{X}|^2 \bullet (\mathbf{W}^{F_0}\mathbf{H}^{F_0})/(\mathbf{S}^{\mathrm{IMM}})^2 \\ \mathbf{Q}^{\Phi} &= (\mathbf{W}^{F_0}\mathbf{H}^{F_0})/\mathbf{S}^{\mathrm{IMM}} \end{cases}$$

- Background music parameters:

$$\mathbf{H}^{M} \leftarrow \mathbf{H}^{M} \bullet \frac{(\mathbf{W}^{M})^T (|\mathbf{X}|^2/(\mathbf{S}^{\mathrm{IMM}})^2)}{(\mathbf{W}^{M})^T (1/\mathbf{S}^{\mathrm{IMM}})}$$

$$\mathbf{W}^{M} \leftarrow \mathbf{W}^{M} \bullet \frac{(|\mathbf{X}|^2/(\mathbf{S}^{\mathrm{IMM}})^2)(\mathbf{H}^{M})^T}{(1/\mathbf{S}^{\mathrm{IMM}})(\mathbf{H}^{M})^T}$$

end for

Algorithm 5.2 Updating rules for the SIMM:
Estimating $\mathbf{\Theta}^{\text{SIMM}} = \{\mathbf{H}^{\Gamma}, \mathbf{H}^{\Phi}, \mathbf{H}^{F_0}, \mathbf{W}^{M}, \mathbf{H}^{M}\}$

for $i \in [1, I]$ **do**

• Leading instrument source parameters:

$$\mathbf{H}^{F_0} \leftarrow \mathbf{H}^{F_0} \bullet \frac{(\mathbf{W}^{F_0})^T \mathbf{P}^{F_0}}{(\mathbf{W}^{F_0})^T \mathbf{Q}^{F_0}}$$

$$\text{where} \begin{cases} \mathbf{P}^{F_0} &= |\mathbf{X}|^2 \bullet (\mathbf{W}^{\Phi}\mathbf{H}^{\Phi})/(\mathbf{S}^{\text{SIMM}})^2 \\ \mathbf{Q}^{F_0} &= (\mathbf{W}^{\Phi}\mathbf{H}^{\Phi})/\mathbf{S}^{\text{SIMM}} \end{cases}$$

• Leading instrument filter parameters:

$$\mathbf{H}^{\Phi} \leftarrow \mathbf{H}^{\Phi} \bullet \frac{(\mathbf{W}^{\Phi})^T \mathbf{P}^{\Phi}}{(\mathbf{W}^{\Phi})^T \mathbf{Q}^{\Phi}}$$

$$\mathbf{H}^{\Gamma} \leftarrow \mathbf{H}^{\Gamma} \bullet \frac{(\mathbf{W}^{\Gamma})^T \mathbf{P}^{\Phi} (\mathbf{H}^{\Phi})^T}{(\mathbf{W}^{\Gamma})^T \mathbf{Q}^{\Phi} (\mathbf{H}^{\Phi})^T}$$

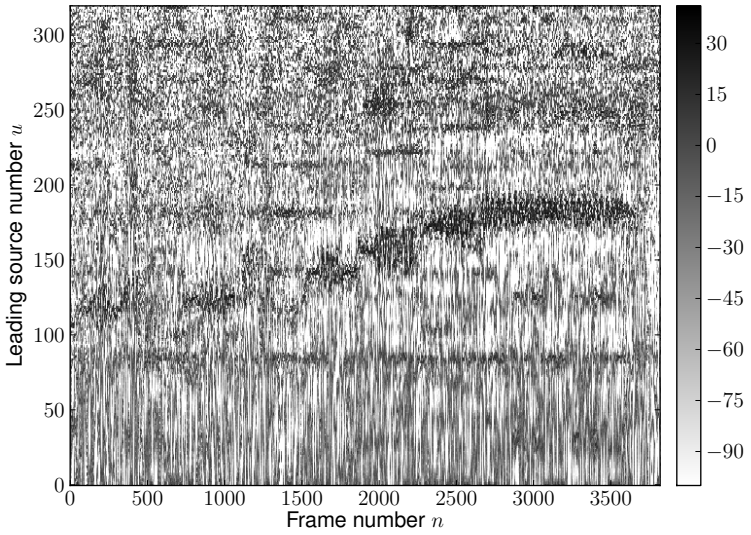
$$\text{where} \begin{cases} \mathbf{P}^{\Phi} &= |\mathbf{X}|^2 \bullet (\mathbf{W}^{F_0}\mathbf{H}^{F_0})/(\mathbf{S}^{\text{SIMM}})^2 \\ \mathbf{Q}^{\Phi} &= (\mathbf{W}^{F_0}\mathbf{H}^{F_0})/\mathbf{S}^{\text{SIMM}} \end{cases}$$

• Background music parameters:

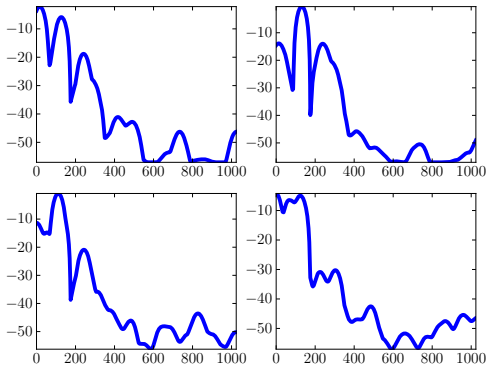
$$\mathbf{H}^{M} \leftarrow \mathbf{H}^{M} \bullet \frac{(\mathbf{W}^{M})^T (|\mathbf{X}|^2/(\mathbf{S}^{\text{SIMM}})^2)}{(\mathbf{W}^{M})^T (1/\mathbf{S}^{\text{SIMM}})}$$

$$\mathbf{W}^{M} \leftarrow \mathbf{W}^{M} \bullet \frac{(|\mathbf{X}|^2/(\mathbf{S}^{\text{SIMM}})^2)(\mathbf{H}^{M})^T}{(1/\mathbf{S}^{\text{SIMM}})(\mathbf{H}^{M})^T}$$

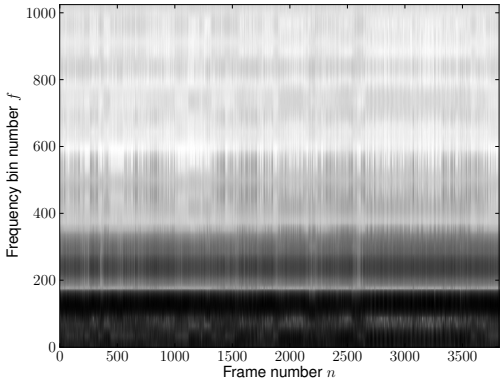
end for



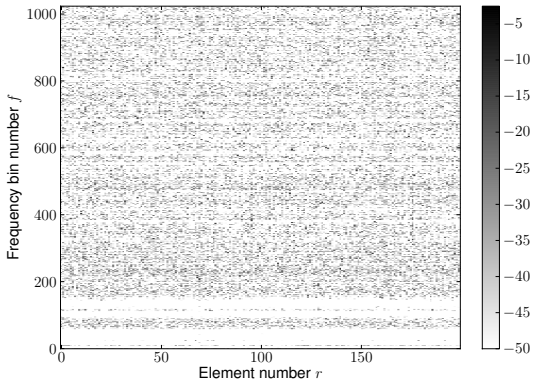
(a) $\mathbf{H}^{F_0}$



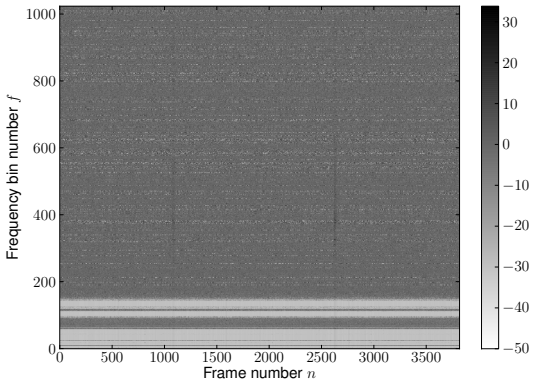
(b) $\mathbf{W}^{\Phi}$



(c) $\mathbf{W}^{\Phi}\mathbf{H}^{\Phi}$



(d) $\mathbf{W}^{M}$



(e) $\mathbf{W}^{M}\mathbf{H}^{M}$

Figure 5.1: Estimated SIMM parameters $\mathbf{H}^{F_0}$, $\mathbf{W}^{\Phi}$, $\mathbf{W}^{\Phi}\mathbf{H}^{\Phi}$, $\mathbf{W}^{M}$ and $\mathbf{W}^{M}\mathbf{H}^{M}$, for the ADC 2004 song "opera_male5.wav".

as a consequence, the matrices for the filters and the accompaniment are admittedly impossible to interprete. A post-processing corresponding to Equations (5.5), (5.10) and (5.14), which tracks the main melody through the estimated parameters by including the temporal dependencies as explained in Section 5.2.3, is necessary. Furthermore, the parameters for the accompaniment part, for the first round of estimation is clearly difficult to interprete, since most of the accompaniment signal was also transcribed in $\mathbf{H}^{F_0}$. This is the reason why it is advised to operate the second round of parameter estimation in Equation (5.15) proposed in system SEP-I: this allows to obtain a better spectral decomposition for the accompaniment, hence leading to Wiener filters that are more fitting to the signal, as is shown on Figure 5.2. Note however that the accompaniment NMF is still rather blurry, which is probably due to the general problem of totally unsupervised NMF. Imposing more structure in the NMF process of the accompaniment could help enhancing the accompaniment estimation part.

5.2.3 Approximations and constraints within the IMM/SIMM

Temporal constraint within the IMM/SIMM

In the IMM/SIMM framework, the IS divergence equivalence with the MAP problem, thanks to Theorem 2, provides an interesting interpretation to the parameter estimation, even without temporal constraints. It can be considered that the power spectrogram of the mixture is decomposed onto all the possible notes, and the amplitude coefficients $\mathbf{H}^{F_0}$ reflect most of the polyphonic content of the audio signal.

Even though the sequence of hidden states has been left aside, it is still possible, using the above interpretation of $\mathbf{H}^{F_0}$, to infer the desired sequence. Indeed, in the MAP framework defined by Equations (5.5), (5.10) or (5.14), the conditional observation probability times the prior density of the set, conditionally upon the state Z^{F_0}, $p(\mathbf{x}_n|\widehat{\mathbf{\Theta}}_n)p(\widehat{\mathbf{\Theta}}_n|Z_n^{F_0})$, still remains to be defined. For all three proposed systems using this scheme, F-II, MUS-I and SEP-I, $\widehat{\mathbf{\Theta}}$ is fixed after the first step, such that whatever the choice of the sequence Z^{F_0}, $p(\mathbf{x}_n|\widehat{\mathbf{\Theta}}_n)$ remains constant. This value merely reflects whether the parameters fit the observations or not.

As for $p(\widehat{\mathbf{\Theta}}_n|Z_n^{F_0})$, it is interesting to consider the desired sequence again, to define a suitable conditional prior. The sequence of fundamental frequencies Z^{F_0} corresponds to the main melody, predominant and continuous. The melody continuity is guaranteed (or at least enforced as much as possible) by the evolution equation:

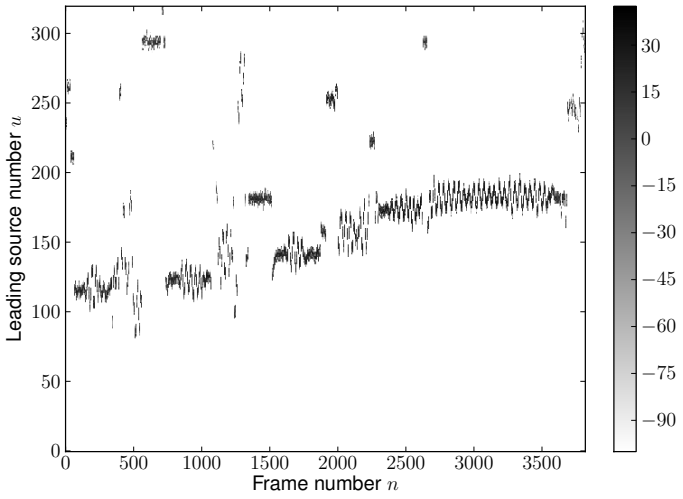
$$p(Z^{F_0}) = p(Z_1^{F_0}) \prod_{n>1} p(Z_n^{F_0}|Z_{n-1}^{F_0}) \tag{5.23}$$

The predominance can here be retranscribed through the interpretation of the amplitude parameters in $\mathbf{H}^{F_0}$. Indeed, since they reflect the polyphonic content, the higher a value $h_{un}^{F_0}$ for frame n and source u, the more likely the corresponding pitch $\mathcal{F}(u)$ was played at frame n. We would therefore expect that the "posterior" probability of a pitch $Z_n^{F_0} = u$ at frame n is proportional to $h_{un}^{F_0}$:

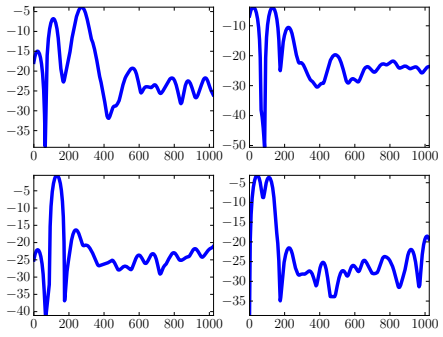
$$p(Z_n^{F_0} = u|\widehat{\mathbf{\Theta}}_n) \propto h_{un}^{F_0} \tag{5.24}$$

Then, thanks to Bayes' law, we can also expect that the prior density of the parameter set conditionally upon the state is proportional to the corresponding amplitude coefficient:

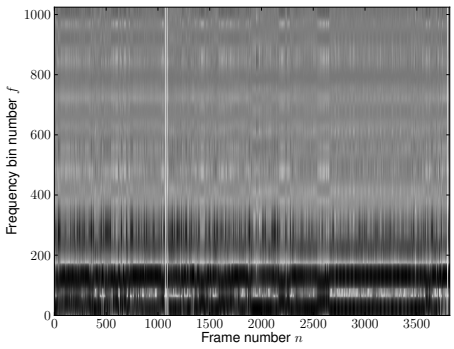
$$p(\widehat{\mathbf{\Theta}}_n|Z_n^{F_0} = u) \propto h_{un}^{F_0} \tag{5.25}$$



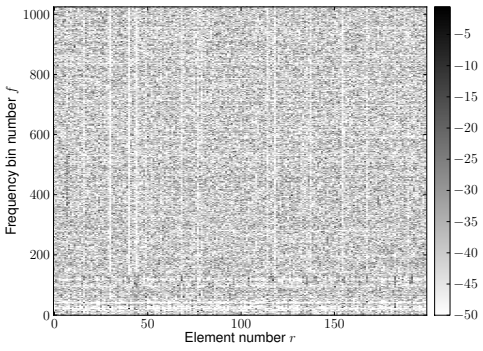
(a) $\mathbf{H}^{F_0}$



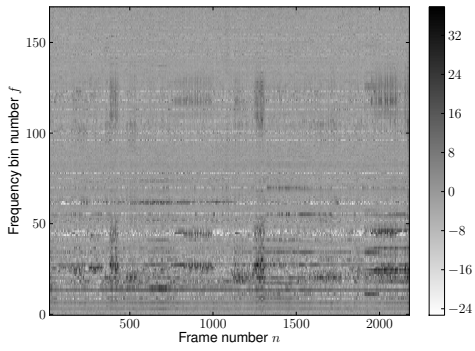
(b) $\mathbf{W}^{\Phi}$



(c) $\mathbf{W}^{\Phi}\mathbf{H}^{\Phi}$



(d) $\mathbf{W}^{M}$



(e) $\mathbf{W}^{M}\mathbf{H}^{M}$

Figure 5.2: Estimated SIMM parameters $\mathbf{H}^{F_0}$, $\mathbf{W}^{\Phi}$, $\mathbf{W}^{\Phi}\mathbf{H}^{\Phi}$, $\mathbf{W}^{M}$ and $\mathbf{W}^{M}\mathbf{H}^{M}$, for the ADC 2004 song "opera_male5.wav", second round (for system SEP-I).

Finally, we set the conditional probability in Equations (5.5), (5.10) and (5.14) such that:

$$p(\mathbf{x}_n|\widehat{\mathbf{\Theta}}_n)p(\widehat{\mathbf{\Theta}}_n|Z_n^{F_0}=u) \propto h_{un}^{F_0} \tag{5.26}$$

Using such a formula during the Viterbi tracking explained in Section 5.4.1, allows to take into account the energy predominance assumption, thanks to the explicit use of the amplitude matrix $\mathbf{H}^{F_0}$ during the decoding.

This IMM model is also, to a certain extent, reminiscent of works by Goto [2004], who also estimates weights which correspond to a similar harmonic decomposition. Note that, as Goto [2004], the proposed estimation is believed to yield a better musical/spectral representation of the mixture, hence potentially better performances than, for instance, auto-correlation based fundamental frequency estimations or the like. Indeed, computing salience functions such as the auto-correlation function (ACF) or other related functions as in [Klapuri, 2008] can usually be considered as a mere change of feature space: such systems transform the signal in order to obtain features (or representations) that are more discriminant for the pitch extraction task. Goto [2004] and our methods decompose the signal on the candidate pitches. Ideally, the amplitude matrix $\mathbf{H}^{F_0}$ represents only the existing pitches in the signal, hence avoiding sub-octave problems, which are likely to occur when using harmonic sums, or over-octave errors, with harmonic product based systems. However, the estimation is not completely safe from over-octave problems, since a harmonic comb can be decomposed onto a non-negative combination of himself and the harmonic comb of its upper octave. The following discussion aims at addressing this particular issue.

Prior on the parameter matrix $\mathbf{H}^{F_0}$

Following the Gamma prior defined in Section 3.4.3, and using the general formula to derive the updating rules in Equation (5.22), along with Equation (5.21), it is easy to derive the corresponding updating rule:

$$\mathbf{H}^{F_0} \leftarrow \mathbf{H}^{F_0} \bullet \frac{\mathbf{P}^{F_0}}{\mathbf{Q}^{F_0}} \tag{5.27}$$

where the matrices $\mathbf{P}^{F_0}$ and $\mathbf{Q}^{F_0}$ are defined by the following equations:

$$p_{un}^{F_0} = \sum_f w_{fu}^{F_0}|x_{fn}|^2 \frac{\left[\mathbf{W}^{\Phi}\mathbf{H}^{\Phi}\right]_{fn}}{(s_{fn}^{(\mathrm{S})\mathrm{IMM}})^2} \tag{5.28}$$

$$q_{un}^{F_0} = \sum_f w_{fu}^{F_0} \frac{\left[\mathbf{W}^{\Phi}\mathbf{H}^{\Phi}\right]_{fn}}{s_{fn}^{(\mathrm{S})\mathrm{IMM}}} + h_{u_8 n}^{F_0} - \frac{\alpha_G - 1}{h_{un}^{F_0}} \tag{5.29}$$

where $u_8 = u - 12U_{st}$, $u \in [12U_{st}U]$. However, as can be seen in Equation (5.29), this way of defining an update rule suffers from the fact that the criterion contributions, namely the conditional likelihood $p(\mathbf{X}|\mathbf{\Theta})$ on one hand and the prior distribution $p(\mathbf{\Theta})$ on the other hand, have different "dynamics", and are not homogeneous. Indeed the units of the derivatives of the former, $\sum_f w_{fu}^{F_0} \frac{\left[\mathbf{W}^{\Phi}\mathbf{H}^{\Phi}\right]_{fn}}{s_{fn}^{(\mathrm{S})\mathrm{IMM}}}$, are homogeneous to the inverse of the source amplitudes, while the derivative of the latter is equivalent to the amplitude $h_{u_8 n}^{F_0}$ or the inverse of the amplitude $\frac{\alpha_G - 1}{h_{un}^{F_0}}$. Note that this contradiction between these 2 terms can be partially solved by setting the scale parameter β_G of the Gamma distribution,

Equation (3.52), to some value proportional to $h^{F_0}_{u_8 n}$, but homogeneous to the inverse of an amplitude.

However, this solution does not address the resulting unbalanced equation in (5.29), and the scaling coefficient is not as obvious to set as it seems. Another simple way of circumventing the problem is to consider a weighted MAP criterion, instead of the MAP criterion in Equation (5.18), such that:

$$C'_{\mathrm{IMM}}(\mathbf{\Theta}) = \sum_{f,n} \log \frac{|x_{fn}|}{\pi s^{(\mathrm{S})\mathrm{IMM}}_{fn}} - \frac{|x_{fn}|^2}{s^{(\mathrm{S})\mathrm{IMM}}_{fn}} + \lambda_G \log p(\mathbf{\Theta}) \tag{5.30}$$

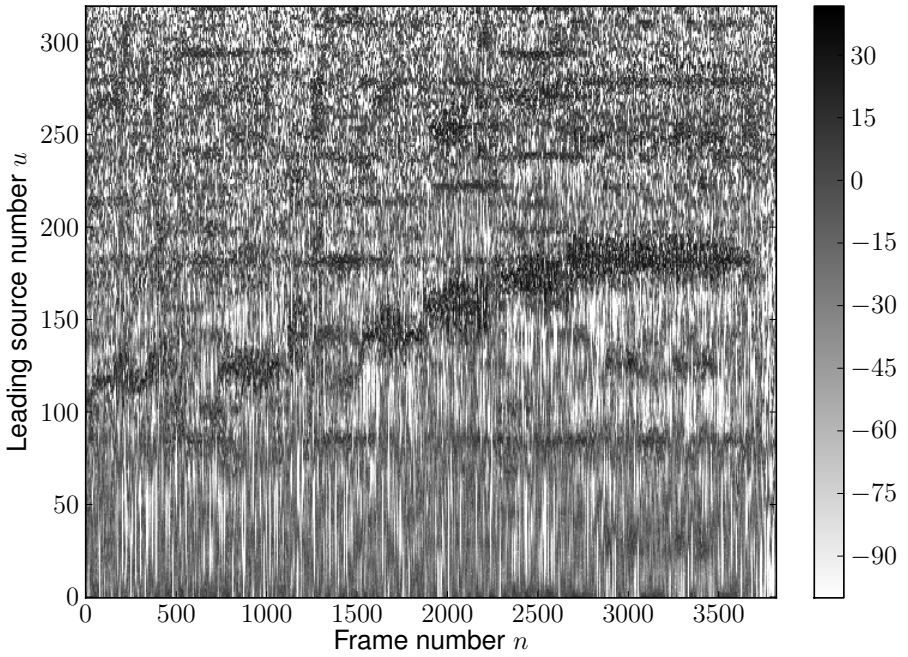
where λ_G allows to control the weight given to the prior distribution. This kind of weighted criterion can be justified as in [Vincent, 2004], for instance. This is also very similar to adding the constraints or penalizations in ML estimation using the Lagrangian multipliers, as often done for sparsity penalizations. Equation (5.29) thus becomes:

$$q^{F_0}_{un} = \sum_f w^{F_0}_{fu} \frac{\left[\mathbf{W}^{\Phi}\mathbf{H}^{\Phi}\right]_{fn}}{s^{(\mathrm{S})\mathrm{IMM}}_{fn}} + \lambda_G \left(h^{F_0}_{u_8 n} - \frac{\alpha_G - 1}{h^{F_0}_{un}} \right) \tag{5.31}$$

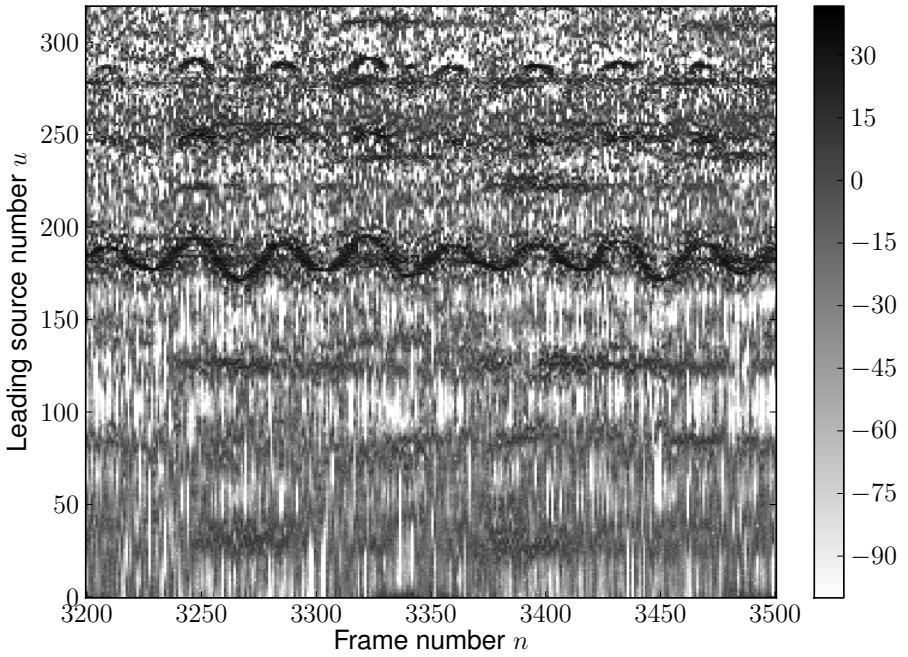
Unfortunately, as for many sparsity penalization strategies (e.g. [Virtanen, 2007, Smaragdis et al., 2008, Mohimani et al., 2008]), it is very difficult to evaluate the correct value for λ_G: if it is too small, the penalization has no effect, while if it is too high, then the constraint becomes predominant, and the result does not fit the signal anymore.

Figures 5.3, 5.4, and 5.5 present some resulting estimations of $\mathbf{H}^{F_0}$, with different values for λ_G, respectively 0, 0.0000556, and 0.01. All the figures use a grayscale colormap, from light to dark colors. As explained above, a small value of λ_G does not lead to great changes, compared with applying no constraint at all ($\lambda_G = 0$). High values tend to produce a matrix which is indeed sparse, with clear cancellation of (irrelevant or spurious) upper octaves. However, as could have been feared from this type of penalization, the value for a given coefficient at pitch u has an impact on the estimation of the lower octave u_8, although not explicitly constrained. Indeed, even with the generic algorithms 5.1 and 5.2, once a certain value is evaluated for a given pitch u, then all the other values are implicitly affected by it, especially if it particularly fits the signal. At last, for values of λ_G which are too "high" (in the example, $\lambda_G = 0.01$), the estimation is too sparse, and biased by the fact that the first coefficients, corresponding to the first octave, do not have any constraint on them. Because of these difficulties to configure the algorithm in a proper way, this feature is not implemented in the systems we have evaluated.

Figure 5.3: Estimation of the amplitude matrix $\mathbf{H}^{F_0}$, with $\lambda_G = 0$ - no constraint.

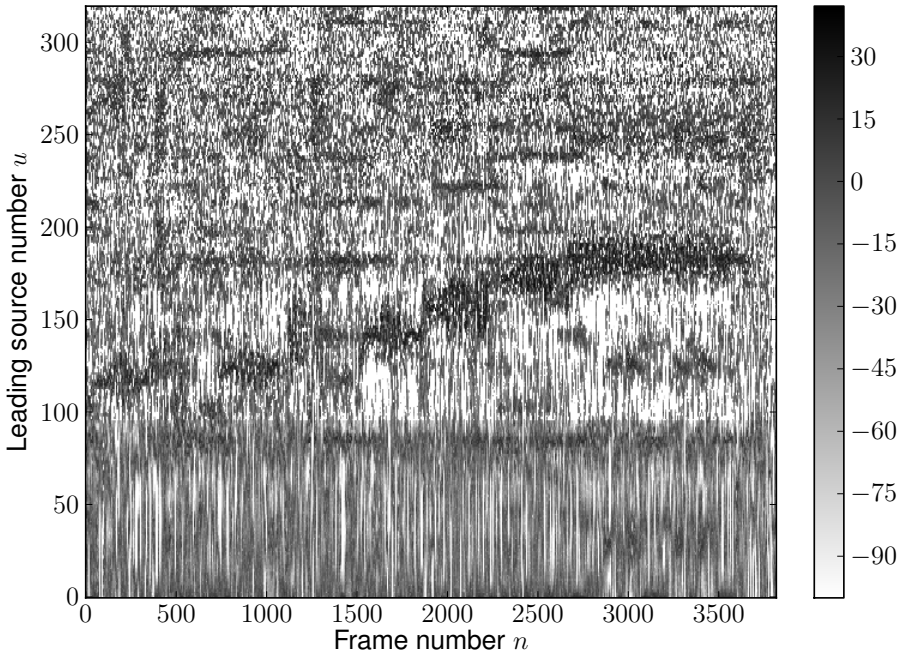


(a) Estimated parameter matrix $\mathbf{H}^{F_0}$, with $\lambda_G = 0$

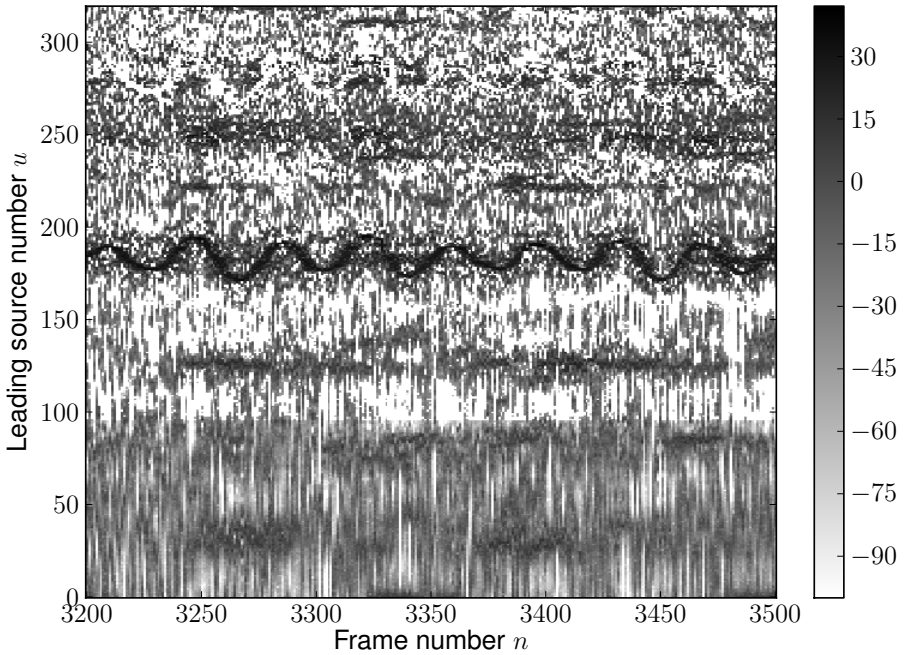


(b) Detail of the parameter matrix $\mathbf{H}^{F_0}$, with $\lambda_G = 0$

Figure 5.4: Estimation of the amplitude matrix $\mathbf{H}^{F_0}$, with $\lambda_G = 0.000556$.

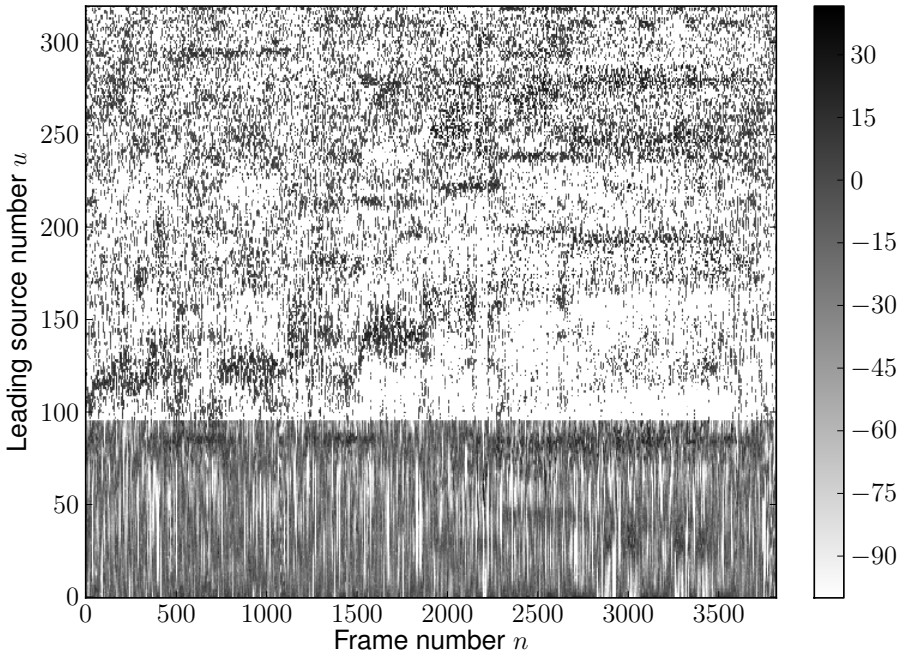


(a) Estimated parameter matrix $\mathbf{H}^{F_0}$, with $\lambda_G = 0.000556$

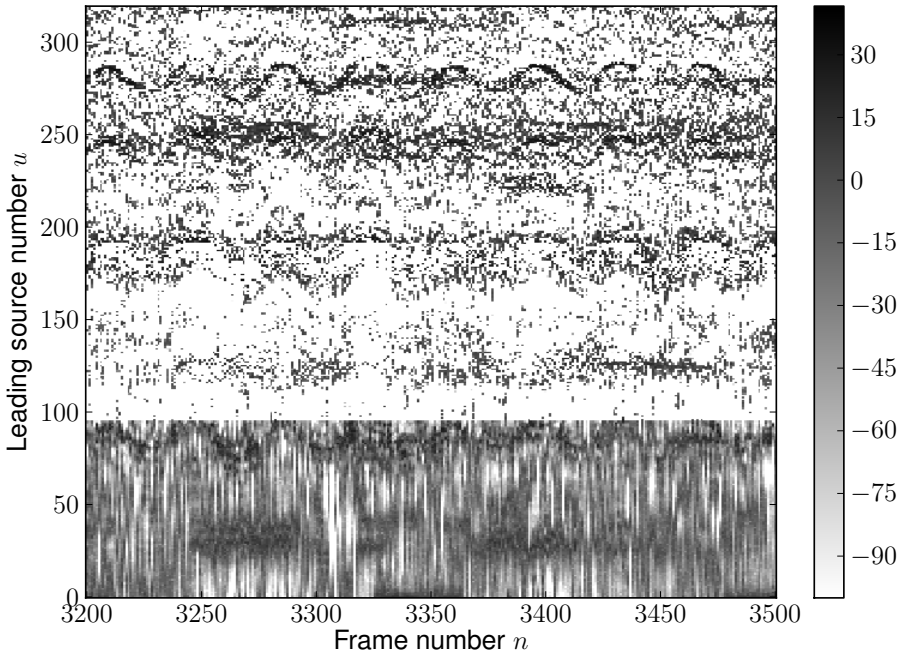


(b) Detail of the parameter matrix $\mathbf{H}^{F_0}$, with $\lambda_G = 0.000556$

Figure 5.5: Estimation of the amplitude matrix $\mathbf{H}^{F_0}$, with $\lambda_G = 0.01$.



(a) Estimated parameter matrix $\mathbf{H}^{F_0}$, with $\lambda_G = 0.01$



(b) Detail of the parameter matrix $\mathbf{H}^{F_0}$, with $\lambda_G = 0.01$

5.3 GSMM/SGSMM: Expectation-Maximisation (EM) algorithm

The GSMM framework, being a hidden state signal model, benefits from previous important works on the related parameter estimation, namely the Expectation-Maximisation (EM) algorithm [Dempster et al., 1977]. We first define the criterion for the maximum likelihood (ML) estimation of the parameters, focussing here on solving Equation (5.2) of system F-I. Then the corresponding updating rules and the iterative EM algorithm are given. At last, the inclusion of the temporal physical constraints, namely the HMM framework, is developed, using a forward-backward procedure [Rabiner, 1989] to compute the E step of the EM algorithm. This last case corresponds to system F-III, solving Equation (5.6). The results of this section have mainly been described in [Durrieu et al., 2010].

In this section, we consider the GSMM set of parameters $\mathbf{\Theta} = \mathbf{\Theta}^{(\text{S})\text{GSMM}}$.

5.3.1 Maximum Likelihood (ML) Criterion for the (S)GSMM

Equation (5.2), for system F-I, is an auxiliary function that is defined, in similar terms, in [Dempster et al., 1977]. As was discussed in Section 5.1, the parameters are first estimated through a ML principle, as defined in Equation (5.1). Dempster et al. [1977] showed that using the following criterion to update the parameter over several iterations allowed to obtain a non-decreasing observation log-likelihood, at iteration i of the algorithm:

$$C_{\text{GSMM}}(\mathbf{\Theta}, \mathbf{\Theta}^{(i-1)}) = E\left[\log p(\mathbf{X}, Z^{F_0}, Z^{\Phi}; \mathbf{\Theta}) | \mathbf{X}; \mathbf{\Theta}^{(i-1)}\right] \tag{5.32}$$

where, in the proposed framework, the complete data is given by the observation $\mathbf{X}$ plus the sequences Z^{F_0} and Z^{Φ}. Due to the iterative nature of this criterion, where two specific sets of parameters intervene: $\mathbf{\Theta}^{(i-1)}$ which is the "current" parameter set at iteration i, and $\mathbf{\Theta}$ which is an arbitrary parameter set. An alternative to this criterion is to take as complete data the set of variables $\{\mathbf{V}, Z^{F_0}, Z^{\Phi}, \mathbf{M}\}$, which is the choice of Ozerov et al. [2007]. In our systems, the first set of complete data was used, but further investigations on the latter complete data set may be necessary in order to compare both strategies.

The Expectation-Maximization (EM) algorithm is based on the maximization of this expectation of the joint log-likelihood for the observations and the hidden states, conditionally upon the observations. Let $i \in [1, I]$ the iteration number, $\mathbf{\Theta}^{(i)}$ the set of parameters updated at iteration i, $Z = \{Z_n = (Z_n^{\Phi}, Z_n^{F_0}); n \in [1, N]\}$ the sequence of active states for the whole observation sequence. The typical EM algorithm flow is as follows:

- E-step: Given the parameter set $\mathbf{\Theta}^{(i-1)}$, compute the criterion $C_{\text{GSMM}}(\mathbf{\Theta}, \mathbf{\Theta}^{(i-1)})$, $\forall \mathbf{\Theta}$,
- M-step: Find $\mathbf{\Theta}^{(i)}$ such that it maximizes the criterion:

$$C_{\text{GSMM}}(\mathbf{\Theta}^{(i)}, \mathbf{\Theta}^{(i-1)}) \geq C_{\text{GSMM}}(\mathbf{\Theta}, \mathbf{\Theta}^{(i-1)}), \forall \mathbf{\Theta}$$

Such a process, and especially the E-step, may seem quite heavy. In practice, the computations can be reduced to a limited number of quantities that depend only on $\mathbf{\Theta}^{(i-1)}$ and not on $\mathbf{\Theta}$ (*sufficient statistics*) which are easily identified when rewriting the criterion. The remainder of the section aims at this re-writing, while Section 5.3.2 aims at giving the EM principle associated with the "new" criterion.

A Lagrangian term can be added to the criterion, to express the condition over the prior probabilities: $\sum_{k,u} \pi_{ku} = 1$. Note that in this study, the prior probabilities were not re-estimated, and fixed to a unique value $\frac{1}{KU}$, such that no state is a priori more likely to occur. In order to make a complete analysis of the criterion, and in order to make further improvements on this model possible, this Lagrangian term needs to be added, and the criterion (5.32) becomes:

$$C_{\text{GSMM}}(\boldsymbol{\Theta}, \boldsymbol{\Theta}^{(i-1)}) = E\left[\log p(\mathbf{X}, Z; \boldsymbol{\Theta}) | \mathbf{X}; \boldsymbol{\Theta}^{(i-1)}\right] - \lambda \left(\sum_{k,u} \pi_{ku} - 1 \right) \tag{5.33}$$

Then, as shown by Dempster et al. [1977], to maximise the likelihood, $\boldsymbol{\Theta}$, at iteration i, can be set to:

$$\boldsymbol{\Theta}^{(i)} = \arg\max_{\boldsymbol{\Theta}} C_{\text{GSMM}}(\boldsymbol{\Theta}, \boldsymbol{\Theta}^{(i-1)}) \tag{5.34}$$

In the generalized form of the EM algorithm, the GEM algorithm, the updated parameter set can also be defined as any parameter set that majorizes the criterion, instead of maximizing it:

$$\boldsymbol{\Theta}^{(i)} | C_{\text{GSMM}}(\boldsymbol{\Theta}^{(i)}, \boldsymbol{\Theta}^{(i-1)}) \geq C_{\text{GSMM}}(\boldsymbol{\Theta}^{(i-1)}, \boldsymbol{\Theta}^{(i-1)}) \tag{5.35}$$

Strictly speaking, the proposed estimation algorithm for the GSMM is a GEM algorithm, since the partial derivatives obtained from the criterion do not allow closed form solutions. Gradient methods to find a parameter set verifying Equation (5.35), as shown in Section 5.3.2, are therefore necessary.

To obtain a criterion which can be easily derivated and used, it is interesting to write the log-likelihood as follows:

$$\log p(\mathbf{X}, Z) = \sum_{n} \log p(\mathbf{x}_n, Z_n) \tag{5.36}$$

$$= \sum_{n} \log p(\mathbf{x}_n | Z_n^{\Phi}, Z_n^{F_0}) + \log \pi_{Z_n^{\Phi} Z_n^{F_0}} \tag{5.37}$$

$$= \sum_{n,k,u} \left[\log p(\mathbf{x}_n | k, u) + \log \pi_{ku}\right] \mathbb{1}_{\left\{Z_n^{\Phi}=k, Z_n^{F_0}=u\right\}} \tag{5.38}$$

The first equation comes from the mutual independence of the observations over the frames, since in this estimation problem, we discarded the temporal constraints. The second equation is a classical result for conditional probabilities, and where Z_n was replaced by the corresponding active states Z_n^{Φ} and $Z_n^{F_0}$. At last, equation (5.38) is a "false sum" over the states. This equation allows us to find a convenient way of expressing the criterion (5.33):

$$C_{\text{GSMM}}(\boldsymbol{\Theta}, \boldsymbol{\Theta}^{(i-1)}) = \sum_{n,k,u} \left[\log p(\mathbf{x}_n | k, u; \boldsymbol{\Theta}) + \log \pi_{ku}\right] E\left[\mathbb{1}_{\{Z_n^{\Phi}=k, Z_n^{F_0}=u\}} | \mathbf{X}; \boldsymbol{\Theta}^{(i-1)}\right] - \lambda \left(\sum_{k,u} \pi_{ku} - 1 \right)$$

Furthermore, by definition of the expectation,

$$E\left[\mathbb{1}_{\{Z_n^{\Phi}=k, Z_n^{F_0}=u\}} | \mathbf{X}; \boldsymbol{\Theta}^{(i-1)}\right] = p(k, u | \mathbf{x}_n; \boldsymbol{\Theta}^{(i-1)})$$

where we used the fact that the couple state $(Z_n^{\Phi}, Z_n^{F_0})$ only depends on $\mathbf{x}_n$, and not on the whole sequence $\{\mathbf{x}_n, n \in [1, N]\}$. The E step of the EM algorithm actually consists in computing this quantity, thanks to Bayes' theorem:

$$p(k, u|\mathbf{x}_n; \mathbf{\Theta}^{(i-1)}) \propto p(\mathbf{x}_n|k, u; \mathbf{\Theta}^{(i-1)})\pi_{ku}^{(i-1)} \tag{5.39}$$

The conditional likelihood of the observations upon the states is given by equation (3.16), using the parameters in $\mathbf{\Theta}^{(i-1)}$. The expression of the criterion is at last given in equation (5.40), where $s_{fn}^{(\mathrm{S})\mathrm{GSMM}|ku}$ is calculated from the model parameters in $\mathbf{\Theta}$, with equation (3.17). The term "CST" is a constant independent from the parameter set $\mathbf{\Theta}$.

$$C_{\mathrm{GSMM}}(\mathbf{\Theta}, \mathbf{\Theta}^{(i-1)}) = \sum_{n,k,u} \left[\sum_f \left(\log \frac{|x_{fn}|}{\pi s_{fn}^{(\mathrm{S})\mathrm{GSMM}|ku}} - \frac{|x_{fn}|^2}{s_{fn}^{(\mathrm{S})\mathrm{GSMM}|ku}} \right) + \log \pi_{ku} \right] \times p(k, u|\mathbf{x}_n; \mathbf{\Theta}^{(i-1)}) - \lambda \left(\sum_{k,u} \pi_{ku} - 1 \right) + \mathrm{CST} \tag{5.40}$$

5.3.2 (S)GSMM updating rules and GEM algorithm

One can derive the updating rules the same way as in Section 5.2.2, since the criterion (5.40) contains a part which is similar to the criterion defined for the IMM, in Equation (5.18). However, for the GSMM, a relatively heavy step is required in order to compute the posterior probabilities $\gamma_n^{(i-1)}(k, u) = p(k, u|\mathbf{x}_n; \mathbf{\Theta}^{(i-1)})$.

The algorithms proposed in this section aim at increasing the value of the criterion given in Equation (5.40). They are Expectation-Maximization (EM) algorithms, because they consist in the alternation of two steps:

- The E-step: the posterior probabilities $\gamma_n^{(i-1)}(k, u)$ are computed, thanks to Bayes's law. The details for this computation and the resulting algorithm that avoids numerical issues are given in Appendix B.2.1, with Algorithm B.1.
- The M-step: the parameters are updated in a way that the criterion in Equation (5.40) is maximized. There is however no analytic solution to this maximization problem: the partial derivatives of the criterion do not lead to equations allowing to obtain an expression of the parameter θ using only the parameter set $\mathbf{\Theta}^{(i-1)}$, at iteration i. This is why, for the GSMM, the same multiplicative gradient approach as in Section 5.2.2 is needed at each iteration, during the M-step, leading to so-called generalized EM (GEM) algorithms. The details of the partial derivatives leading to the expressions of the multiplicative gradients as in Section 5.2.2 are given in Appendix B.2.

The algorithms for the GSMM and the Smoothed filter GSMM (SGSMM) are given in Algorithm 5.3. As for Algorithm 5.1, the initial set of parameters is randomly drawn.

5.3.3 Including constraints: Hidden Markov-GSMM (HM-GSMM) algorithm

Including the physical layer in the estimation corresponds to jointly estimating the parameters and the hidden state probabilities within a HMM framework. The criterion for the

Algorithm 5.3 EM algorithm for the (S)GSMM: Estimating $\boldsymbol{\Theta}$, equal to $\boldsymbol{\Theta}^{\text{GSMM}} = \{\mathbf{B}, \mathbf{W}^{\Phi}, \mathbf{H}^{M}, \mathbf{W}^{M}\}$ or $\boldsymbol{\Theta}^{\text{SGSMM}} = \{\mathbf{B}, \mathbf{H}^{\Gamma}, \mathbf{H}^{M}, \mathbf{W}^{M}\}$

for $i \in [1, I]$ **do**

- $\forall k, u, n,\ \ b_{kun} \leftarrow b_{kun} \dfrac{p^B_{kun}}{q^B_{kun}}$, where $\left\{ \begin{array}{rcl} p^B_{kun} & = & \displaystyle\sum_f \dfrac{w^{\Phi}_{fk} w^{F_0}_{fu} |x_{fn}|^2}{(s^{\text{(S)GSMM}|ku}_{fn})^2} \\ q^B_{kun} & = & \displaystyle\sum_f \dfrac{w^{\Phi}_{fk} w^{F_0}_{fu}}{s^{\text{(S)GSMM}|ku}_{fn}} \end{array} \right.$

E step: compute $\gamma_n^{(i-1)}(k,u) = p(k,u|\mathbf{x}_n; \boldsymbol{\Theta}^{(i-1)})$ with Algorithm B.1.
M step: update the parameters:

- (GSMM) $\forall f, k,\ \ w^{\Phi}_{fk} \leftarrow w^{\Phi}_{fk} \dfrac{p^{\Phi}_{fk}}{q^{\Phi}_{fk}}$, where $\left\{ \begin{array}{rcl} p^{\Phi}_{fk} & = & \displaystyle\sum_{u,n} \gamma_n^{(i-1)}(k,u) \times \dfrac{b_{kun} w^{F_0}_{fu} |x_{fn}|^2}{(s^{\text{(S)GSMM}|ku}_{fn})^2} \\ q^{\Phi}_{fk} & = & \displaystyle\sum_{u,n} \gamma_n^{(i-1)}(k,u) \dfrac{b_{kun} w^{F_0}_{fu}}{s^{\text{(S)GSMM}|ku}_{fn}} \end{array} \right.$
- (SGSMM) $\forall p, k,\ \ h^{\Gamma}_{pk} \leftarrow h^{\Gamma}_{pk} \dfrac{p^{\Gamma}_{pk}}{q^{\Gamma}_{pk}}$, where $\left\{ \begin{array}{rcl} p^{\Gamma}_{pk} & = & \displaystyle\sum_{f,u,n} \gamma_n^{(i-1)}(k,u) \times \dfrac{b_{kun} w^{\Gamma}_{fp} w^{F_0}_{fu} |x_{fn}|^2}{(s^{\text{SGSMM}|ku}_{fn})^2} \\ q^{\Gamma}_{pk} & = & \displaystyle\sum_{f,u,n} \gamma_n^{(i-1)}(k,u) \dfrac{b_{kun} w^{\Gamma}_{fp} w^{F_0}_{fu}}{s^{\text{SGSMM}|ku}_{fn}} \end{array} \right.$
- $\forall r, n,\ \ h^{M}_{rn} \leftarrow h^{M}_{rn} \dfrac{p^{H}_{rn}}{q^{H}_{rn}}$, where $\left\{ \begin{array}{rcl} p^{H}_{rn} & = & \displaystyle\sum_{k,u,f} \gamma_n^{(i-1)}(k,u) \dfrac{w^{M}_{fr} |x_{fn}|^2}{(s^{\text{(S)GSMM}|ku}_{fn})^2} \\ q^{H}_{rn} & = & \displaystyle\sum_{k,u,f} \gamma_n^{(i-1)}(k,u) \dfrac{w^{M}_{fr}}{s^{\text{(S)GSMM}|ku}_{fn}} \end{array} \right.$
- $\forall f, r,\ \ w^{M}_{fr} \leftarrow w^{M}_{fr} \dfrac{p^{W}_{fr}}{q^{W}_{fr}}$, where $\left\{ \begin{array}{rcl} p^{W}_{fr} & = & \displaystyle\sum_{k,u,n} \gamma_n^{(i-1)}(k,u) \dfrac{h^{M}_{rn} |x_{fn}|^2}{(s^{\text{(S)GSMM}|ku}_{fn})^2} \\ q^{W}_{fr} & = & \displaystyle\sum_{k,u,n} \gamma_n^{(i-1)}(k,u) \dfrac{h^{M}_{rn}}{s^{\text{(S)GSMM}|k,u}_{fn}} \end{array} \right.$

end for

HM-GSMM model has the same definition as for the GSMM criterion, in Equation (5.33). However, the joint likelihood term inside the expectation operator does not bear the same information. Indeed, this time, the probability of the sequence $Z = (Z^{\Phi}, Z^{F_0})$, $p(Z)$, cannot be factorized the same way as in (5.37). With the HMM assumption only, discarding the musicological constraint of the E layer, this evolution equation writes:

$$p(Z) = p(Z_1) \prod_{n=2}^{N} p(Z_n|Z_{n-1}) \tag{5.41}$$

where $p(Z_n|Z_{n-1})$ is the transition probability discussed in Section 3.3.3. After some derivations similar to those of Section 5.3.1, we obtain the following criterion:

$$C_{\text{HMM}}(\boldsymbol{\Theta}, \boldsymbol{\Theta}^{(i-1)}) = \sum_{n,k,u} \left[\sum_{f} \left(\log \frac{|x_{fn}|}{\pi s_{fn}^{\text{(S)GSMM}|ku}} - \frac{|x_{fn}|^2}{s_{fn}^{\text{(S)GSMM}|ku}} \right) + \log \pi_{ku} \right] \times p(Z_n^{\Phi} = k, Z_n^{F_0} = u|\mathbf{X}; \boldsymbol{\Theta}^{(i-1)}) - \lambda \left(\sum_{k,u} \pi_{ku} - 1 \right) + \text{CST} \tag{5.42}$$

As in Equation (5.33), the term "CST" is independent from any element of $\boldsymbol{\Theta}$. Note however that for the HM-GSMM, this term also contains the components concerning the transition probabilities. Since in the proposed work, these probabilities are not estimated, these terms do not need to be explicitly shown. The only noticeable and important difference between Equation (5.33) and Equation (5.42) are the posterior probabilities $p(Z_n^{\Phi} = k, Z_n^{F_0} = u|\mathbf{X}; \boldsymbol{\Theta}^{(i-1)})$. Indeed, in Equation (5.33), the conditioning is done only upon the value of the observation at the current frame n: $p(Z_n^{\Phi} = k, Z_n^{F_0} = u|\mathbf{x}_n; \boldsymbol{\Theta}^{(i-1)})$, while in the HM-GSMM framework, this simplification is not possible anymore. A forward/backward procedure, detailed in Appendix B.2.8 is needed to compute the HM-GSMM posterior probabilities.

Apart from this forward/backward procedure to compute these quantities, one can use exactly the same updating rules as in Algorithm 5.3 for the GSMM, since the criteria are equal for these aspects. The *posterior* probabilities, $\gamma_n^{(i-1)}(k, u)$, are for this case defined as:

$$\gamma_n^{(i-1)}(k, u) = p(k, u|\mathbf{X}; \boldsymbol{\Theta}^{(i-1)}) \tag{5.43}$$

These probabilities are computed with Algorithm B.2.

5.4 Temporal evolution of the states and sequence estimation

5.4.1 Viterbi algorithm to address the HMM of the physical layer for Z^{Φ} and Z^{F_0}

The decoding of the optimal path Z intervenes in all the proposed systems in Equations (5.3), (5.5), (5.7), (5.10) and (5.14). Although the different underlying models can be different, the decoding principle is the same: in this section, a general framework is presented, as well as the Viterbi algorithm that allows to efficiently decode the optimal path. This framework can be easily identified with any of the models corresponding to the different systems. These links are explicitly given at the end of the section.

For all the systems, the desired path maximizes the posterior probability of the sequence, given the observation. Thanks to Bayes' Law, this is also equivalent to maximizing the joint likelihood of the observation $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ and the (hidden) sequence $Z = \{Z_1, \ldots, Z_N\}$:

$$\widehat{Z} = \arg\max_{Z} p(\mathbf{X}, Z) \tag{5.44}$$

The sequence Z is modelled as a Markov process, as shown on Figure 3.13, such that

$$p(Z) = p(Z_1) \prod_{n>1} p(Z_n|Z_{n-1}) \tag{5.45}$$

where the transition probability at frame n from the previous state Z_{n-1} to the current state Z_n depends on the system and is given later. We denote by $q(i,j)$ the transition probability from state $Z_{n-1} = i$ to state $Z_n = j$, which does not depend on the frame number n.

Furthermore, conditionally upon the states, the observation vectors in $\mathbf{X}$ are independent, as shown in Figure 3.13. Therefore, the joint likelihood verifies:

$$\begin{aligned} p(\mathbf{X}, Z) &= p(\mathbf{X}|Z)p(Z) \\ &= \prod_{n} p(\mathbf{x}_n|Z_n) \times p(Z_1) \prod_{n>1} p(Z_n|Z_{n-1}) \\ &= p(Z_1)p(\mathbf{x}_1|Z_1) \prod_{n>1} p(\mathbf{x}_n|Z_n)p(Z_n|Z_{n-1}) \end{aligned} \tag{5.46}$$

At this stage, the conditional probabilities $p(\mathbf{x}_n|Z_n)$ are given by the estimated parameters $\mathbf{\Theta}$ (be it for the IMM or the GSMM), such that the unknown quantities in Equation (5.46) only consist of the desired sequence Z.

Equation (5.44) suggests that the joint likelihood should be calculated for each possible path before the optimal path can be selected as the one maximizing these quantities. In the case of $Z^{F_0} \in [1, U]^N$, there are U^N possible paths, namely, for typical values, around 1000^{3000}. Such a number of trials is prohibitive, and can be avoided thanks to the efficient **Viterbi algorithm** [Viterbi, 1967]. For the remainder of this section, Z_n is assumed to take values in $[1, U]$.

During the Viterbi algorithm, only the necessary quantities and pieces of information are kept. The main principle comes from the computation of an intermediary quantity δ_{un}, for all $u \in [1, U]$ and $n \in [1, N]$:

$$\delta_{un} = \max_{Z_{1:n-1}} p(\mathbf{x}_{1:n}, Z_{1:n-1}, Z_n = u) \tag{5.47}$$

This quantity is linked to the desired joint likelihood, since we have:

$$\max_{u} \delta_{uN} = \max_{Z_N = u} \max_{Z_{1:N-1}} p(\mathbf{x}_{1:N}, Z_{1:N-1}, Z_N = u) = \max_{Z} p(\mathbf{X}, Z) \tag{5.48}$$

It is easy to find a relation between $\delta_{u(n+1)}$ and δ_{vn}, $\forall v \in [1, U]$, for $1 \geq n$:

$$\begin{aligned}
\delta_{u(n+1)} &= \max_{Z_{1:n}} p(\mathbf{x}_{1:n+1}, Z_{1:n}, Z_{n+1} = u) \\
&= \max_{Z_n = v} \left(\underbrace{\max_{Z_{1:n-1}} p(\mathbf{x}_{1:n}, Z_{1:n-1}, Z_n = v)}_{\delta_{vn}} \right) p(\mathbf{x}_{n+1}|Z_{n+1} = u) p(Z_{n+1} = u|Z_n = v) \\
\delta_{u(n+1)} &= \left(\max_v \delta_{vn} q(v, u) \right) p(\mathbf{x}_{n+1}|Z_{n+1} = u) \qquad (5.49)
\end{aligned}$$

Thanks to this last relation, the decoding can be iteratively held with one loop over all the frames. At each frame, the Viterbi algorithm fills in the column vectors of two $U \times N$ matrices: the matrix of maximum joint likelihoods $\mathbf{\Delta}$ and the matrix of the antecedents $\mathbf{\Psi}$, which records all the necessary labels that correspond to the maxima obtained in $\mathbf{\Delta}$. The maximum of the last column of $\mathbf{\Delta}$ gives the maximum of the joint likelihood. Using the antecedents in $\mathbf{\Psi}$, the best path for $\widehat{Z}$ can be tracked back from that maximum value. The detailed computation rules are given in Algorithm 5.4.

At last, in order to find the optimal path Z or Z^{F_0} for the different proposed systems, the above generic framework can easily be re-adapted. The following remarks may also be of importance for the corresponding systems:

F-I The desired sequence is $Z = (Z^{F_0}, Z^{\Phi})$.

F-II To decode the sequence Z^{F_0}, the conditional probability is assumed to be proportional to the amplitude coefficients in $\mathbf{H}^{F_0}$, as explained in Section 5.2.3.

F-III Although the temporal constraint is also included during the estimation of the parameters, the decoding of the optimal sequence Z is still held with the Viterbi algorithm 5.4.

MUS-I The sequence estimated with the physical layer is actually used as an initialization for the beam search strategy aiming at decoding the sequence of notes.

SEP-I The Viterbi tracking is done between two rounds of parameter estimation. Such an iterative estimation is not optimal in some ways, but including the physical constraints as for system F-III has its limits. Preliminary tests done with system F-III do not lead to improvements in solo separation, on the contrary, it seems less robust than the proposed system SEP-I, using F-II as pre-processor for the melody tracking. Note that any other melody tracker could replace F-II, if required.

5.4.2 Beam search pruning strategy for the musical note layer E

At last, for system MUS-I, the note sequence E from the musicological layer can be decoded through Equation (5.11), which is constituted of 3 main parts: the frame-wise signal model from Section 3.4.1, the physical layer for the F0 sequence, Section 3.3.3, linked with the musicological layer of Section 3.3.4:

$$\underbrace{p(\mathbf{X}|\widehat{\mathbf{\Theta}}^{(\mathrm{S})\mathrm{IMM}}) \times p(\widehat{\mathbf{\Theta}}^{(\mathrm{S})\mathrm{IMM}}|Z^{F_0})}_{\text{Frame-wise level}} \quad \times \quad \underbrace{p(Z^{F_0}|E)}_{\text{Physical constraint}} \quad \times \quad \underbrace{p(E)}_{\text{Musicological constraint}} \qquad (5.56)$$

Algorithm 5.4 Viterbi algorithm

Initialization:
for $u \in [1, U]$ **do**

$$\delta_{u1} = p(Z_1 = u)p(\mathbf{x}_1|Z_1 = u) \tag{5.50}$$
$$\psi_{u1} = 0 \tag{5.51}$$

end for
Iteration:
for n from 2 to N **do**
  for $u \in [1, U]$ **do**

$$\delta_{un} = \left(\max_v \delta_{v(n-1)} q(v, u)\right) p(\mathbf{x}_n|Z_n = u) \tag{5.52}$$
$$\psi_{un} = \arg\max_v \delta_{v(n-1)} q(v, u) \tag{5.53}$$

  end for
end for
Termination:

$$\widehat{Z}_N = \arg\max_u \delta_{uN} \tag{5.54}$$

Backtracking:
for n from N to 2 **do**

$$\widehat{Z}_{n-1} = \psi_{\widehat{Z}_n n} \tag{5.55}$$

end for

Following the algorithm explained in [Vincent, 2004] in order to find the best note-level path results in an almost exhaustive search on the parameter space. Furthermore, the memory space needed for the beam search would be huge, since it would then be necessary to store every parameter matrix $\mathbf{\Theta}^{(\mathrm{S})\mathrm{IMM}}$ for each potential path.

Instead, following the principle given in Section 5.1.1, system MUS-I involves rounds of estimations that aim, in the end, at approximating the estimation of the optimal desired solution $\widehat{E}, \widehat{Z}^{F_0}$. First, without any temporal constraint, a first round of estimation provides a parameter set $\widehat{\mathbf{\Theta}} = \{\widehat{\mathbf{H}}^{\Gamma}, \widehat{\mathbf{H}}^{\Phi}, \widehat{\mathbf{H}}^{F_0}, \widehat{\mathbf{W}}^{M}, \widehat{\mathbf{H}}^{M}\}$. As discussed in Section 5.2.3, the source amplitude matrix $\widehat{\mathbf{H}}^{F_0}$ can be considered as a pitch salience function which reflects the polyphonic content of the song. The estimation of the sequence $\widehat{Z}^{F_0}$ from Equation (5.10) then provides a first estimate of the desired melody line, in terms of fundamental frequencies, by including the physical layer of Section 3.3.3. At last, the estimation of the note sequence is held through Equation (5.11), where the estimated parameters $\widehat{\mathbf{\Theta}}$ and the estimated sequence $\widehat{Z}^{F_0}$ are used to alleviate the computational load that would have been necessary otherwise with the beam search strategy proposed in [Vincent, 2004].

The principle of the beam search algorithm comes from the possibility to write the joint likelihood up to a frame n thanks to the joint likelihood up to frame $n-1$:

$$\begin{aligned} p(\mathbf{x}_{1:n}, \widehat{\mathbf{\Theta}}_{1:n}, Z^{F_0}_{1:n}, E_{1:n}) &= p(\mathbf{x}_n, \widehat{\mathbf{\Theta}}_n, Z^{F_0}_n, E_n | \mathbf{x}_{1:n-1}, \widehat{\mathbf{\Theta}}_{1:n-1}, Z^{F_0}_{1:n-1}, E_{1:n-1}) \\ &\qquad \times p(\mathbf{x}_{1:n-1}, \widehat{\mathbf{\Theta}}_{1:n-1}, Z^{F_0}_{1:n-1}, E_{1:n-1}) \\ &= \Big(p(\mathbf{x}_n|\widehat{\mathbf{\Theta}}_n) p(\widehat{\mathbf{\Theta}}_n|Z^{F_0}_n) p(Z^{F_0}_n|Z^{F_0}_{n-1}, E_n) p(E_n|E_{1:n-1}) \Big) \\ &\qquad \times p(\mathbf{x}_{1:n-1}, \widehat{\mathbf{\Theta}}_{1:n-1}, Z^{F_0}_{1:n-1}, E_{1:n-1}) \end{aligned} \tag{5.57}$$

In Equation (5.57), the different components have been discussed and defined in the previous sections such that:

$$p(\mathbf{x}_n|\widehat{\mathbf{\Theta}}_n) p(\widehat{\mathbf{\Theta}}_n|Z^{F_0}_n = u) \propto h^{F_0}_{un} \tag{5.58}$$

$$p(Z^{F_0}_n|Z^{F_0}_{n-1}, E_n) \propto p(Z^{F_0}_n|Z^{F_0}_{n-1}) p(Z^{F_0}_n|E_n) \tag{5.59}$$

where the relation in Equation (5.58) is discussed in Section 5.2.3, and the different probabilities in Equation (5.59) are defined in Section 3.3.4. The probability $p(E_n|E_{1:n-1})$ is given by Equation (3.32). Thanks to relation (5.57), the joint likelihood of the whole song can be computed iteratively over the frames. To determine which path E is optimal, this joint likelihood must be computed for all possible paths. However, at each frame, simple rules can be defined so as to prune the partial paths that are not likely to lead to feasible paths. For instance, the joint likelihood of a given partial path $E_{1:n}$ up to frame n may be 0, due to impossible transitions or very low corresponding probabilities/energies in $\mathbf{H}^{F_0}$.

To further prune the paths, only a few possible note candidates are considered, at each frame. These candidates are extracted from the estimated F0 sequence $\widehat{Z}^{F_0}_n$, at frame n:

$$E^C_n(\delta^{\mathrm{MIDI}}) = 12 \log_2 \left(\frac{\mathcal{F}(\widehat{Z}^{F_0}_n)}{440} \right) + 69 + \delta^{\mathrm{MIDI}} \tag{5.60}$$

where $\delta^{\mathrm{MIDI}} \in [-\Delta^{\mathrm{MIDI}}, \Delta^{\mathrm{MIDI}}]$, Δ^{MIDI} being the authorized deviation from the "standard" fundamental frequency of the note n^{MIDI} to the actual corresponding fundamental frequency $\mathcal{F}(\widehat{Z}^{F_0}_n)$ in the signal, expressed in semitones. This value allows for more or less tolerance to effects such as vibrato, imprecise attacks and other expressive effects.

We first initialize partial paths at frame 1 with the provided note candidates. Then each path is extended to the next frame through one of the following operations: continuation of silence, namely $E_2 = E_1 = 0$, continuation of a note, $E_2 = E_1 \neq 0$, deletion, $E_1 \neq 0$, $E_2 = 0$, onset from silence, $E_1 = 0$ and $E_2 \neq 0$ and replacement of note, $E_1 \neq E_2$, $E_1 \neq 0$ and $E_2 \neq 0$. The values for E_2 are the candidates $E_2^C(\delta^{\text{MIDI}})$ as defined in Equation (5.60). Another issue is to jointly estimate the sequence of fundamental frequencies with the source state sequence $\widehat{Z}^{F_0}$ corresponding to the note sequence $\widehat{E}$. For a given path $E_{1:n}$, at frame n, the source state is chosen such that:

$$\begin{aligned}\widehat{Z}_n^{F_0} &= \arg\max_u p(\mathbf{x}_n|\mathbf{\Theta}_n)p(\mathbf{\Theta}_n|Z_n^{F_0}=u)p(Z_n^{F_0}=u|\widehat{Z}_{n-1}^{F_0})p(Z_n^{F_0}=u|E_n) \\ &= \arg\max_u h_{un}^{F_0} p(Z_n^{F_0}=u|\widehat{Z}_{n-1}^{F_0})p(Z_n^{F_0}=u|E_n) \end{aligned} \tag{5.61}$$

where $p(Z_n^{F_0} = u|\widehat{Z}_{n-1}^{F_0}) = q(\widehat{Z}_{n-1}^{F_0}, u)$ and $p(Z_n^{F_0} = u|E_n)$ are respectively defined in Equations (3.19) and (3.29). Equation (5.61) does not of course give the optimal sequence in the sense of the global criterion or even in the sense of the HMM structure, as with the Viterbi algorithm. However, it provides an approximate yet relevant solution to the issue of finding the correct F0 to the current note.

Once all the partial paths have been extended with all the possibilities mentioned above, other pruning strategies may apply. First, only a few paths, N^{paths} paths, are kept as candidates to be extended on the next frame. The partial paths are therefore classified by decreasing value of their joint likelihood, and only the first N^{paths} paths are kept. If two partial path share the same current note, with same onsetting frame, then the path with the lowest probability can be dropped, since in any case, the other one will always be preferred to that one. This iterative scheme is repeated until the last frame of the song. The process is summed up in Algorithm 5.5. For each path, several quantities need to be stored: the fundamental frequency states $\widehat{Z}_{1:n}^{F_0}$, the corresponding note sequence $\widehat{Z}_{1:n}^{F_0}$, the note onsets, their durations and the last note onset.

Note at last that it would be possible to apply the Viterbi algorithm [Rabiner, 1989] in this case, namely by considering the duration as a random variable. However, this would also result in a substantial increase of memory use, since it implies considering as hidden states all the combinations of notes at all the possible durations.

Algorithm 5.5 Estimation of note sequence E for system MUS-I

Initialize paths
for all the candidates E_1^C **do**
  Find the optimal $\widehat{Z}_1^{F_0}$.
  Compute $p(\mathbf{x}_1, \widehat{\mathbf{\Theta}}_1, \widehat{Z}_1^{F_0}, \widehat{E}_1)$.
end for
Organize and prune the paths.
Extending the paths
for $n \in [2, N]$ **do**
  for all paths that were kept **do**
    Extend the path with one of the operations: continuation, deletion, replacement. The replacement and new notes have to be extracted from the candidate notes E_n^C.
    for each newly extended path **do**
      if $E_n \neq 0$ **then**
        Find the optimal $\widehat{Z}_n^{F_0}$.
      end if
      Compute $p(\mathbf{x}_{1:n}, \widehat{\mathbf{\Theta}}_{1:n}, \widehat{Z}_{1:n}^{F_0}, \widehat{E}_{1:n})$, using $p(\mathbf{x}_{1:n-1}, \widehat{\mathbf{\Theta}}_{1:n-1}, \widehat{Z}_{1:n-1}^{F_0}, \widehat{E}_{1:n-1})$.
    end for
  end for
  Organize and prune the paths.
end for
Termination
$\widehat{E}$ is chosen as the path E that maximizes $p(\mathbf{X}, \widehat{\mathbf{\Theta}}, \widehat{Z}^{F_0}, E)$.

Chapter 6

Applications

In this chapter, we first address the melody F0 estimation and musical transcription applications, and the proposed methods are evaluated. The results presented here were previously introduced in [Durrieu et al., 2008a] and [Durrieu et al., 2010]. The extension using the explicit duration model has been presented in [Weil et al., 2009b].

We then describe the leading instrument separation results that were presented in [Durrieu et al., 2009a] for the mono-channel case and in [Durrieu et al., 2009b] for the stereo extension.

6.1 F0 estimation and musical transcription of the main melody

We are interested in two applications: the estimation the sequence corresponding to the fundamental frequencies Z^{F_0} is a first step towards music database indexation purposes, which provides a description rather close to the signal. The objectives of the melody transcription are mainly to produce the sequence of notes E played by the leading instrument.

The target applications essentially are Query-by-Humming (QbH) applications, pure transcription, for "advanced" users, or song indexing in order to retrieve desired content-based information from a database.

The following sections are based on [Durrieu et al., 2008a], [Durrieu et al., 2010], [Durrieu et al., 2009c] and [Weil et al., 2009b], with results reported from the Music Information Retrieval Evaluation eXchange (MIREX) evaluation campaigns in 2008 and 2009, for the "Audio Melody Extraction" task.

6.1.1 Frame-wise F0 estimation of the melody

After the task definition has been recalled, we describe the proposed systems. The performance measures and the experiments are then discussed. These results were obtained through two international evaluation campaign on this topic MIREX 2008 (which was run again at our request) and MIREX 2009. In the process of submitting our contributions, some "tuning" experiments were made, and are also described below. The results of our systems are then discussed, and at last some hints about the performance of system F-III are given.

6.1.1.1 Task definition

The task of frame-wise "Audio Melody Extraction", as expected for the MIREX campaigns, has been defined in Section 2.2, in Definition 5. The systems are more specifically expected to comply with the following format:

- **Input:** a digital audio file, WAV format, with a sampling rate of 44100Hz.
- **Output:** a file with the format described in Definition 5, providing the desired melody fundamental frequency sequence.
- **Definition of the main melody:** the melody is not formally defined within the MIREX evaluation. Definition 4 along with the interpretations and discussions given in Section 2.1 are here assumed to hold.

6.1.1.2 Proposed methods

Two particular methods have been extensively investigated in this study, notably participating to MIREX 2008 and 2009: the first one is **system F-I**, which uses the GSMM, followed by the Viterbi algorithm.

The second method is **system F-II**, which uses the IMM, also followed by the Viterbi smoothing algorithm, this time applied to the amplitudes and not on the posterior probabilities, called system F-II. For system F-II, as it uses the IMM, there is an issue with modelling silences of the lead instrument. Indeed, this would amount to adding a spectrum with zeroes at all frequency bins to the matrix $\mathbf{W}^{F_0}$. The corresponding amplitude in $\mathbf{H}^{F_0}$ would not be defined, as the global contribution to the mixture would anyway be null. This amplitude can therefore not be interpreted as measuring the "strength" of silence, and another strategy needs to be investigated. The adopted approach is therefore different from the one used for the GSMM (systems F-I and F-III). A first attempt was to create an "amplitude" coefficient corresponding to the silence, as a function of all the amplitudes for the other states. However, such a method highly depends on the signal itself, and was hard to tune for general purposes. Another solution could be to detect the frames that are voiced or unvoiced using some classification method as a pre-processing [Ozerov et al., 2007], jointly [Hsu et al., 2009] or as a post-processing [Fujihara et al., 2008]. However, such a scheme would need some supervision, while the aim of the proposed approaches is rather to provide an unsupervised method. Instead, a more heuristic method was designed, based on the energy of the frames after separation, hence ruling out the frames where the energy of the estimated fundamental frequency is too low: after the Viterbi smoothing, the energy of the estimated leading voice for each frame is first computed, based on the parameters corresponding to the estimated main melody path. The frames are then classified into "leading voice" and non-"leading voice" segments with a threshold on their energies. The threshold is empirically chosen such that the remaining frames represent more than 99.95% of the total leading instrument energy. Fundamental frequencies of frames for which the energy is under the threshold are set to 0 after smoothing.

At last, the **third proposed system, F-III**, which uses the HMM structure directly during the estimation of the parameters was only roughly tested on some examples. The results for F-III are only sketched in Section 6.1.1.11, while the results obtained by F-I and F-II, which were both evaluated at the MIREX 2008 and 2009 campaigns, are reported from Section 6.1.1.4 to Section 6.1.1.10.

6.1.1.3 Performance measures

The proposed algorithms were evaluated with other systems at the MIREX 2008 and 2009 Audio Melody Extraction tasks. The metrics that were used are the same as for the MIREX 2005 edition of the task, described in Poliner et al. [2007]. These metrics are frame-wise (as opposed to note-wise) measures: in this setting, the onsets and offsets of the different notes are not considered, but only the fundamental frequency for a given frame. An estimated pitch that falls within a quarter tone from the ground-truth on a given frame and a frame correctly identified as unvoiced are true positives (TP). The main metrics are then:

- **Raw Pitch Accuracy (Acc.):** the accuracy only on the voiced frames:

$$\text{Raw Pitch Acc.} = \frac{\#\{\text{Voiced TP}\}}{\#\{\text{Voiced Frames}\}}$$

- **Overall Accuracy:** accuracy over all the frames, taking into account the silence (unvoiced) frames:

$$\text{Overall Acc.} = \frac{\#\{\text{TP}\}}{\#\{\text{Frames}\}}$$

The last measure, the overall accuracy, also takes into account the ability of the systems to determine whether the lead instrument plays or not, while the raw pitch accuracy informs about the ability of the systems to detect the right pitch when there is a non-null one.

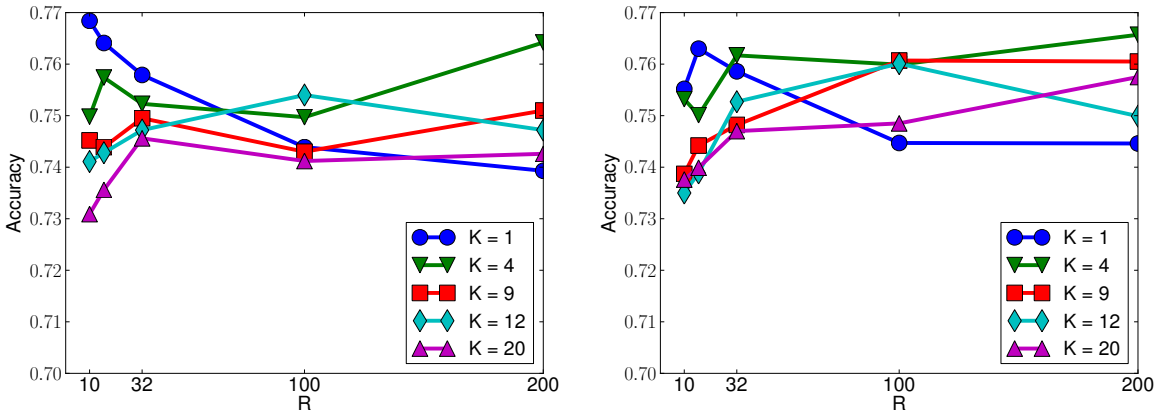
To evaluate whether the extraction of the melody was successful, another approach consists in evaluating the output with respect to a "back-end" application, such as a QbH or cover version detection system. This way, one can know whether the estimation of the melody was done in a way that it improves the result of the targetted application. One can consider our work on lead instrument separation partly as a validation application for our melody extraction system. The separation results are reported in Section 6.2.

6.1.1.4 Datasets for evaluation

The ISMIR04 database is composed of 20 songs and the MIREX05 dataset of 25 songs, both databases are described in Poliner et al. [2007]. For MIREX 2008, a new dataset (MIREX08) was also proposed, with 8 vocal Indian classical music excerpts[1]. At last, a new test set was provided for MIREX 2009, generated the same way as the MIR-1K dataset, which is a set of short excerpts of karaoke songs. The provided ground-truth for all the datasets is the framewise melody line of the predominant instrument, i.e. one fundamental frequency per frame. The hopsize between two frames is 10ms. The original songs are sampled at 44100Hz. Before processing, they are down-sampled to 11025Hz in our studies, mainly for computation time considerations. The duration of the excerpts may vary from 10 seconds up to about 30 seconds.

Also note that preliminary results for the IMM were published in Durrieu et al. [2008a]. The development database for MIREX 2008 was the ISMIR2004 set (20 files) with 13 files from the MIREX05 dataset. For MIREX 2009, MIR-1K could also be used as a development set. All these databases are further detailed in Appendix D.1.

[1] This subset is similar to the examples from http://www.ee.iitb.ac.in/daplab/MelodyExtraction/.



(c) IMM: hopsize of 5.8ms.

(d) SIMM: hopsize of 5.8ms.

Figure 6.1: F-II ((S)IMM): tuning of the model parameters K and R.

6.1.1.5 Practical choices for the model parameters

In our model, some parameters such as the number of spectral shapes for the filter or for the accompaniment, among others, need to be set beforehand. Different parameter combinations were tested with the IMM algorithm in order to choose a combination that leads to fairly good results in most cases. Some of the obtained average accuracies, on the development set for MIREX 2008 (ISMIR04 + parts of MIREX05 datasets), are displayed on Figure 6.1.

First, several values of the number of filters K and the number of accompaniment components R were tested. The obtained accuracies roughly range from 73% to 77%. Lower values of K and higher values for R tend to give better results. It is interesting to note that even for $K = 1$, i.e. with only one filter, the spectral combs of the leading voice source part are well adapted to the signal. In the IMM, as shown on Figure 6.1(c), the filter part is not constrained to be smooth. This may explain why even a single estimated filter for the whole signal was sometimes enough to provide good results. For melody transcription, it is not harmful to use such unconstrained filters. However, for applications where these filters are directly used for their semantic meaning, such as lyrics recognition, smoothing the filters may become necessary. In the SIMM framework, on Figure 6.1(d), where the filters are smoothed, the results for $K = 1$ are still the best ones for low values of R, but from $R = 32$ on, results with $K = 4$ are the best. This seems to show that constraining the filters and allowing enough components for the accompaniment may limit the above problem.

For our further experiments, we chose $K = 4$ and $R = 32$. These values ideally correspond to 4 filters, representing 4 different vowels, and to 32 components for the accompaniment, i.e. 32 different spectral shapes, one for each note or percussive sound. This choice also leads to good results while allowing good generalization capabilities.

We also tested a simpler model for the source spectral combs, replacing the amplitudes of the glottal model for each harmonic (see Appendix C) by $c_h = 1$. Theoretically, using such combs should be identical to the glottal model. However, according to our results, it is still better to use the glottal model. This model is indeed closer to actual natural sounds,

Table 6.1: Parameter values for U_{st}, for each system F-I, F-II and F-III.

U_{st}	MIREX08	MIREX09
F-I	4	4
F-II	4	8
F-III	4	4

with exponentially decreasing spectral envelopes. With spectral combs whose envelopes are uniform, the filter spectral shapes have more to compensate to fit the signals. The chosen iterative algorithms, especially the EM algorithm, are however very sensitive to the initialization. Since the filters are randomly initialized, the general initial set of values is probably closer to the desired solution with the glottal source model, hence leading to better results.

At last, since our GSMM implementation is much slower than our IMM implementation, we have assumed that the chosen parameter tuning was correct for both algorithms. Some other parameters also had to be manually set, such as the "strength" of the smoothing constraint on $\mathbf{H}^{F_0}$ in the Viterbi tracking, α, in Equation (3.19) and the number of elements per semitone in $\mathbf{W}^{F_0}$, U_{st}. The values used for U_{st} are given in Table 6.1. α was set to 10.0. This value could also be learnt from some database. When changing the frame rate (the STFT "hop size"), one should be careful to consequently change α, since bigger F0 jumps, with lower values of α, should be allowed with longer hop sizes.

6.1.1.6 Convergence

In spite of the lack of formal convergence proof for the proposed iterative methods, according to our simulations and tests, the chosen criteria $C_{\text{GSMM}}(\boldsymbol{\Theta}, \boldsymbol{\Theta}')$ and $C_{\text{IMM}}(\boldsymbol{\Theta})$ and, equivalently, the log-likelihood of the observation $\log p_{\boldsymbol{\Theta}}(\mathbf{X})$ increase over the iterations, as can be seen on the evolution of the observation log-likelihood for an excerpt of the MIREX development database on Fig. 6.2, for each model. The model parameters are therefore well estimated, or at least converge to a local maximum. However, concerning the melody estimation results, we noticed that running the algorithms with many more iterations paradoxically resulted in worse melody estimations. This may be due to a tuning problem of the fixed source spectra for the main voice $\mathbf{W}^{F_0}$. If a note in the main voice is detuned compared to the given dictionary, it will very likely be estimated as belonging to the accompaniment, especially if there are enough iterations for the accompaniment dictionary to fit such a signal.

6.1.1.7 Comparison between the proposed models (S)GSMM and (S)IMM

The (S)GSMM and (S)IMM algorithms (Algorithms 5.3, 5.1 and 5.2), respectively for system F-I and F-II, lead to parameters that really are different. Theoretically, the main disadvantage of the (S)IMM is the fact that several notes are allowed at the same time, even if they are constrained to share the same timbral envelope. In practice this timbre "constraint" is quite loose and the estimated amplitudes in $\mathbf{H}^{F_0}$ reflect most of the polyphonic content of the music, including some of the accompaniment, which leads to the need for the melody tracker introduced in section 5.4.1.

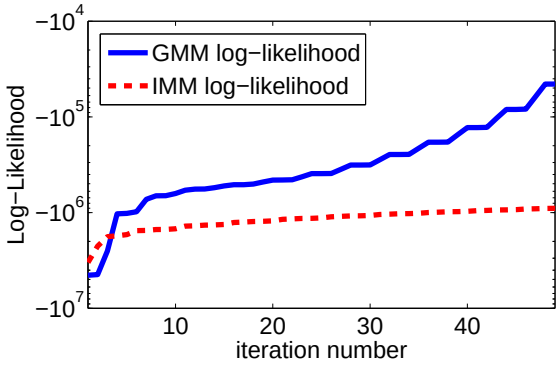


Figure 6.2: Evolution of the log-likelihood of the observations for the GSMM and IMM algorithms (Algorithm 5.3 and Algorithm 5.1).

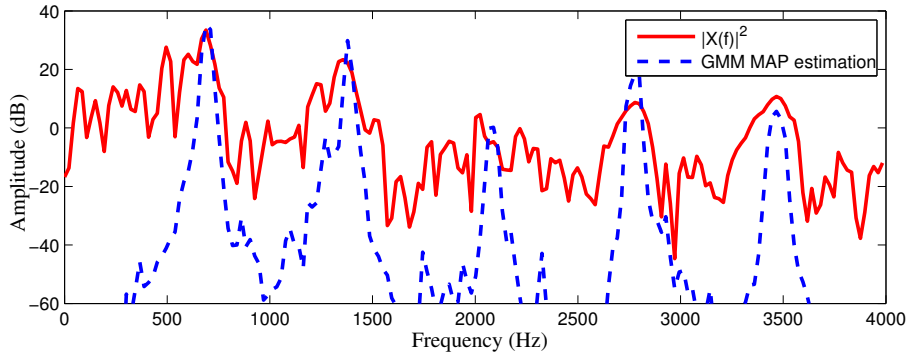
However, it turns out, in certain circumstances, to be an advantage over the (S)GSMM. Figure 6.3 shows some results obtained with our models: the estimated (approximated) spectrum for the main instrument is displayed over the original spectrum for each model. This frame is part of the file "opera_fem4.wav" from the ISMIR 2004 main melody extraction database, at $t = 9.665$s. On the original spectrum, one can see the main "note", at around $f_0 = 690$Hz, among several other accompaniment notes. This frame actually corresponds to a "chirp", transition from a minimum to a maximum F0 value, by the singer, during a vibrato: the higher the frequency, the wider the lobes of the main "harmonic comb". The estimations of the main note for F-I and F-II are both correct according to the ground-truth, and the peaks of the resulting combs fit to the ones of the original one. However, these figures show that the GSMM parameters do not fit the real data as closely as the IMM ones do. This tends to prove that the IMM is able to better fit vocal parts, especially on frequency transition frames (vibrato): on these segments, the GSMM assumption of having one constant fundamental frequency within the frame does not hold.

The IMM could also be used for a polyphonic instrument, but its design as shown on the diagram figure 3.16 does not allow different sources to have different timbres (filters): for a given filter k, at frame n, all the source excitations share the same amplitude h^{Φ}_{kn}. A more sensible model for polyphonic music analysis would be to directly replace the state selector in the GSMM diagram figure 3.10 by an instantaneous mixture. However, such a model leads to many more parameters to be estimated, hence to numerical problems and indeterminacies.

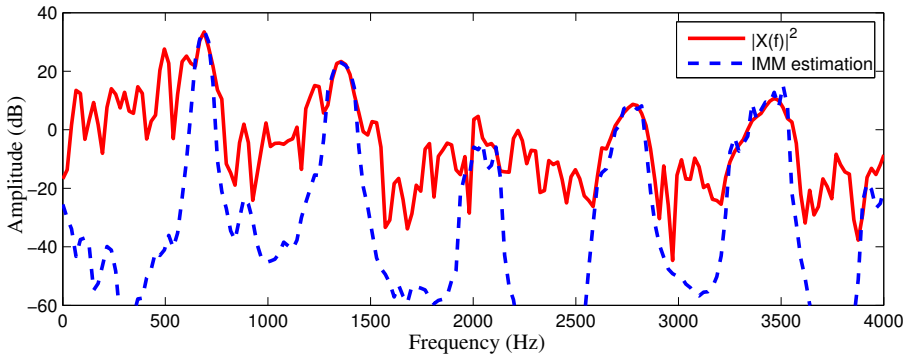
6.1.1.8 MIREX 2008: Main Melody Estimation Results

Table 6.2 provides the main results for the MIREX 2008 evaluation. The results for each of the different databases (ISMIR04, MIREX05 and MIREX08) are individually given. The "Total" column gives the average of these results, weighted by the number of files in the corresponding database.

The best result for each column is shown in bold font. We also provide the results of two other systems that were presented MIREX 2006. The proposed GSMM based system F-I is denoted "drd1" and the IMM F-II "drd2", for compliance with the notations of the evaluation campaign. The other systems "clly", "pc", "rk", "vr" are respectively described in [Cao and Li, 2008], [Cancela, 2008], [Ryynänen and Klapuri, 2006] and [Rao and Rao,



(a) F-I (GSMM) estimation result



(b) F-II (IMM) estimation result

Figure 6.3: "opera_fem4.wav": spectrum of a frame with a frequency chirp around $f_0 = 690$Hz of the main melody, and the corresponding estimated spectra by system F-I (GSMM) and F-II (IMM) algorithms (derived in Section 5.2 and 5.3).

	ADC2004		MIREX05	
System	Raw Pitch Acc.	Overall Acc.	Raw P. Acc.	Ov. Acc.
clly1	75.3%	50.2%	68.9%	48.9%
clly2	75.3%	68.0%	68.9%	61.4%
drd1 (F-I, GSMM)	65.9%	59.6%	57.4%	52.2%
drd2 (F-II, IMM)	**85.7%**	81.5%	**72.4%**	66.0%
pc	85.1%	**85.1%**	71.0%	**69.8%**
rk	82.4%	78.8%	69.7%	64.9%
vr	77.1%	70.1%	71.2%	63.5%
Average	78.1%	70.5%	68.5%	61.0%
Dressler	82.9%	82.5%	**77.7%**	**73.2%**
Poliner	73.2%	71.9%	66.2%	63.0%
	MIREX08		**Average on the 3 sets**	
System	Raw Pitch Acc.	Overall Acc.	Raw P. Acc.	Ov. Acc.
clly1	54.7%	51.4%	69.2%	49.8%
clly2	54.7%	49.7%	69.2%	62.1%
drd1 (F-I, GSMM)	85.8%	**76.0%**	64.9%	58.6%
drd2 (F-II, IMM)	81.8%	75.0%	**78.9%**	73.2%
pc	83.9%	73.3%	78.3%	**76.1%**
rk	83.5%	75.3%	77.3%	71.1%
vr	**88.2%**	66.7%	75.3%	67.1%
Average	76.1%	66.8%	73.3%	65.4%

Table 6.2: Results of the proposed algorithms compared to the other systems submitted to MIREX 2008 Audio Melody Extraction task. We also added the results by 2 participants from the MIREX 2006 edition of the task.

2008].

On average over the 3 databases, system F-II (drd2, IMM) obtained the best accuracy on the voiced frames, and the second overall accuracy. On the 2004 and 2005 sets, it also performed first for the voiced frames, second for the overall accuracy. On the 2008 dataset, it obtained over 80% on the voiced frames and 75% of overall accuracy. These results show that **F-II is robust to the variations of the database**.

System F-I, in average, did not perform so well, especially on the 2004 and 2005 datasets. On the other hand, on the 2008 set, it obtained the best overall accuracy. It seems to perform quite well in certain favorable cases: for the MIREX08 dataset, the polyphony is rather weak. The main voice - a singer - is prominent over a background music consisting of a soft harmonic pedal played by a traditional string instrument plus some Indian percussions. The 2005 database seems to be closer to the average Western world commercial music production, and is therefore quite diverse, with "stronger" polyphonies. In the GSMM framework, any melody line played in a song can lead to a local maximum of the criterion C_{GSMM}. If the initialization of the EM algorithm is too far from the desired solution, the parameters might converge towards one of those maxima, and

miss the main voice. It happens for instance when the main instrument is not a singer, or if other instruments have a relatively strong energy in the song. Note that this also affects the results with F-II (IMM), but up to a lesser scale than with F-I (GSMM). Alternatively, another reason may also come from a problem in the algorithm design. For some songs, the posterior probabilities degenerated in the EM algorithm, hence leading to incoherent results. This possibility, along with the huge computation time needed for F-I reflects how complicated it is in practice to manipulate and stabilize this sort of estimation algorithms.

Globally, on the provided development set (the 20 songs from ISMIR04 and 13 songs from the MIREX05 set), the percentage of voiced frames is about 85% for ISMIR04 and 63% for MIREX05. Successfully transcribing the main melody, with respect to the chosen evaluation criteria, therefore requires a good segmentation scheme into voiced/unvoiced frames for the main voice. Additionally, the system has to identify the main instrument and discriminate between its occurrences and other instruments that may also appear as "predominant" when the desired main voice is silent. This latter case happens more often with lower voiced frame percentages. Indeed, all the participating systems experienced a relative drop in performance on the MIREX05 set, which proves the need for better schemes to detect voiced frames. The approach of the system in [Dressler, 2005], which participated to the MIREX 2005 and 2006 audio melody extraction tasks, seems to overcome this problem and appears quite robust even in comparison with the results for MIREX 2008, as demonstrated by its participation to MIREX 2009 [Dressler, 2009].

At last, for both F-I and F-II, on some poorly transcribed songs, the Viterbi process misled the sequence to fit an erroneous "path", e.g. following a sequence one octave higher than the desired sequence. When the parameters of the models are poorly estimated or correspond to another instrument on one frame, the Viterbi algorithm propagates the errors to the neighbouring frames. The transcribed melody may therefore be, on some segments, the one played by an instrument other than the desired main instrument.

6.1.1.9 MIREX 2009: comparison with MIREX 2008 on development sets

We have used all three available databases in order to develop our algorithms. These sets are the ISMIR04 set (20 files of about 30 seconds each), the MIREX05 set (13 files, 20 seconds) and MIR-1K (1000 files, 10 seconds each).

During our tests, we obtained the "Raw Pitch/Total Accuracy" results given in Table 6.3. The fundamental frequency range was set for both algorithms to $[80, 800]$, with 4 pitches per semi-tone for system F-I (SGSMM) and 8 for F-II (SIMM). The reported F-II results correspond to a system with $K = 2$ and $R = 100$, after 50 iterations. For F-I, $K = 2$, $R = 20$, after 15 iterations. The results for F-I (GSMM) and F-II (IMM) obtained at MIREX 2008 are also reported. Note that the results for MIREX 2008, on the MIREX05 subset, were computed on the full set, and not only on the development set, as in the lines for F-I (SGSMM) and F-II (SIMM).

The results of Table 6.3 show that both systems have quite similar results, which is what we woud have expected, since system F-II (SIMM) is an approximation of the primary model provided by F-I (SGSMM), as concerns the monophonic assumption. The results on ADC04 and MIREX05 are of the same order as the performances for 2008 [University of Illinois Urbana Champaign USA, 2008]. The results for F-I (SGSMM) are much higher than those of F-I (GSMM) (drd1 at MIREX 2008 evaluation campaign), which may also be explained because of a bug rather than because of model differences. The results for F-II (SIMM) are slightly lower than the previously obtained ones for F-II (IMM): the

Algorithm	**ADC04**	**MIREX05**	**MIR-1K**
F-I (GSMM) (MIREX'08)	66/60	57/52	
F-II (IMM) (MIREX'08)	86/82	72/66	
F-I (SGSMM) (MIREX'09)	84/78	79/67	55/51
F-II (SIMM) (MIREX'09)	82/78	79/68	58/55

Table 6.3: Results of the tested algorithms, given for each development dataset, reported as "Raw pitch/Total Accuracy" (in percentage).

added filter smoothness does not generally improve melody estimation, at least with the chosen set of parameters. By constraining more the spectral shapes for the leading voice, compared with the MIREX 2008 submissions, we allow less flexibility for the parameters to adapt to the analyzed signal. This can result in more difficulties in detecting the correct fundamental frequencies.

6.1.1.10 MIREX 2009: results on test set

The datasets that were used for the Audio Melody Extraction evaluation campaign at MIREX 2009 are the ADC04, the MIREX05 test set (25 files), the MIREX08 set (8 files), and excerpts from the MIR-1K, denoted as the MIREX09 dataset. For MIREX 2009, our submitted algorithms were denoted as "drd1" for system F-I (SGSMM) and "drd2" for F-II (SIMM).

In Table 6.4, the results for each database, for the AME task at MIREX 2009 are given. The results obtained by our submissions are slightly under those of our best submission for 2008, F-II (IMM, "drd2"). We have already discussed a potential reason for such a decrease. Another reason could also come from the iterative nature of both our 2008 and 2009 submissions, which leads to algorithms that are quite sensitive to the initialisation. It is therefore hard to compare these submissions based on a single run.

Compared with the other systems, our 2009 submissions seem to perform fairly well, with good results on almost all the datasets, except for the -5dB MIR-1K set, on which most submitted systems also break down. Our models in general seem less adapted to the MIR-1K dataset. Compared with the first proposed dataset, ADC04, the MIR-1K dataset seems to better fit a more specific "**singing** melody extraction" task, rather than the general "**audio** melody extraction" which was originally stated. Indeed, in MIR-1K, the main instrument is a human singer, but some other instruments sometimes play the melody along with the singer, potentially at the upper octave, and not always at a lower energy. One may therefore define, for these examples, two melodies, instead of one. This ambiguity is solved if the task is redefined as tracking the singer. A potential way of improving the results on such a set would be to explicitly include a vocal/non-vocal classification step, for instance as a pre-processing as in Ozerov et al. [2007].

At last, our submissions seem to have better results on the vocal subsets, especially on the MIREX 2008 subset, for which F-I (SGSMM) obtained top results, and on the vocal pieces of the ADC04 and MIREX05 subsets. Non-vocal pieces across the datasets include several synthetized MIDI files. For such music pieces, the NMF based accompaniment model is usually able to fit to the whole signal spectrogram. The estimated leading instrument may in the worst case be identified with any of the instruments of the mixture.

	ADC2004		**MIREX05**	
System	Raw Pitch Acc.	Overall Acc.	Raw P. Acc.	Ov. Acc.
cl1	85.1%	76.6%	70.1%	61.0%
cl2	85.1%	75.4%	70.1%	62.6%
drd1 (F-I, SGSMM)	81.4%	75.7%	72.7%	65.8%
drd2 (F-II, SIMM)	81.2%	74.7%	70.4%	65.2%
hjc1	63.9%	47.4%	59.1%	41.1%
hjc2	50.5%	44.6%	44.0%	38.7%
jjy	83.3%	75.1%	69.5%	59.9%
kd	**87.1%**	**86.3%**	**76.4%**	**74.8%**
mw	82.3%	70.8%	75.0%	58.2%
pc	82.9%	82.5%	68.0%	66.5%
rr	76.9%	70.7%	69.0%	60.7%
toos	61.0%	52.5%	67.5%	51.6%
	MIREX08		**MIREX09 (+5dB)**	
System	Raw P. Acc.	Ov. Acc.	Raw P. Acc.	Ov. Acc.
cl1	50.8%	45.3%	70.3%	51.7%
cl2	50.8%	46.8%	70.3%	57.2%
drd1 (F-I, SGSMM)	**88.0%**	**81.2%**	81.0%	72.8%
drd2 (F-II, SIMM)	86.6%	80.0%	77.3%	72.8%
hjc1	67.6%	48.1%	84.9%	74.8%
hjc2	60.8%	46.5%	78.4%	75.0%
jjy	68.3%	61.2%	84.4%	51.7%
kd	87.8%	80.7%	**89.2%**	**78.4%**
mw	86.0%	73.5%	77.0%	50.0%
pc	81.8%	73.6%	63.7%	61.5%
rr	86.2%	79.0%	77.9%	76.7%
toos	79.8%	68.5%	84.8%	55.7%
	MIREX09 (0dB)		**MIREX09 (-5dB)**	
System	Raw P. Acc.	Ov. Acc.	Raw P. Acc.	Ov. Acc.
cl1	59.1%	44.0%	45.4%	34.5%
cl2	59.1%	49.2%	45.4%	39.9%
drd1 (F-I, SGSMM)	69.9%	60.1%	53.8%	45.5%
drd2 (F-II, SIMM)	66.5%	59.5%	50.5%	44.8%
hjc1	72.7%	53.2%	48.7%	38.5%
hjc2	51.7%	51.7%	21.5%	37.5%
jjy	75.9%	49.7%	58.5%	42.2%
kd	80.5%	**68.2%**	62.5%	**51.7%**
mw	67.3%	43.6%	53.1%	43.3%
pc	50.9%	51.5%	37.4%	41.6%
rr	68.6%	60.8%	54.7%	43.4%
toos	**82.3%**	53.6%	**74.9%**	48.6%

Table 6.4: Results of the proposed algorithms compared to the other systems submitted to MIREX 2009 Audio Melody Extraction task.

6.1.1.11 Preliminary results for system F-III

The system F-III was the longest to develop, and required some additional algorithmic tweakings before it could be used on the audio data. Indeed, the use of the HMM directly during the parameter estimation raised some numerical issues, especially during the forward-backward procedure Rabiner [1989]. These problems are addressed in Appendix B.2.8.

On Figure 6.4, the estimated posterior probabilities $p(Z_n^{F_0} = u|\mathbf{X})$ $(= p(Z_n^{F_0} = u|\mathbf{x}_n)$ for the GSMM of system F-I) as well as the corresponding optimal sequence $\widehat{Z}^{F_0}$ are shown for systems F-I and F-III.

There isn't any formal evaluation for system F-III yet. However, on some tests done on ADC 2004, the results of F-III are not as good as for system F-I. From these results, it seems that one problem comes from the design of the transitions between the pitches, and especially the introduction of the silence state. Indeed, the silence state was often preferred, probably when the source model for the leading instrument, *i.e.* one constant pitch during each frame, did not fit anymore to the data. This problem was already highlighted for system F-I, as pictured on Figure 6.3. The chirp problem seems even stronger in system F-III, because the posterior probability computation takes into account the neighbouring frames, hence spreading the model inaccuracies over several frames, while this was more limited in system F-I, in which the posterior probabilities were "locally" computed.

6.1.2 Notewise transcription of the melody

As part of a collaboration between the author and Jan Weil, from the Technische Universität Berlin, in Germany, within the European project K-Space, the article [Weil et al., 2009b] was published, with results reported in this section.

6.1.2.1 Task definition

The goal is to return the list of notes played by the leading instrument, i.e. their boundaries (onset and offset) as well as their label on the Western musical scale.

6.1.2.2 Performance measures

The note-wise evaluation of the main melody extraction has not yet been discussed on MIREX. The evaluation framework for the polyphonic note transcription has been more actively discussed and the metrics defined in MIREX for the polpyhonic case seem to fit the main melody estimation problem.

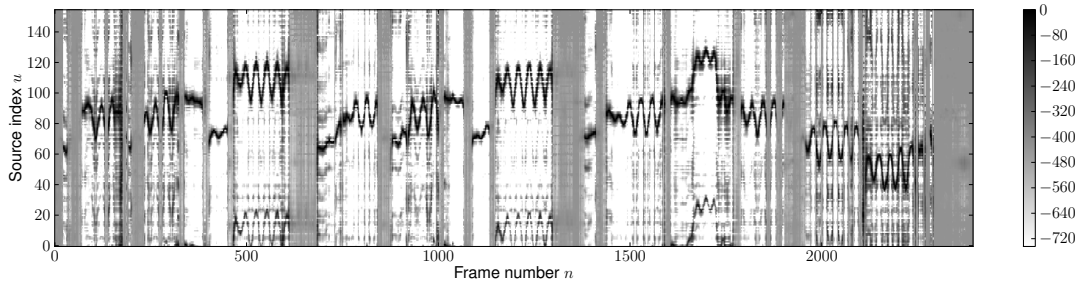
The results obtained by system MUS-I were therefore evaluated with the following recall R, precision P and F-measure F criteria:

$$R = \frac{\#\{\text{correctly transcribed notes}\}}{\#\{\text{reference notes}\}} \tag{6.1}$$

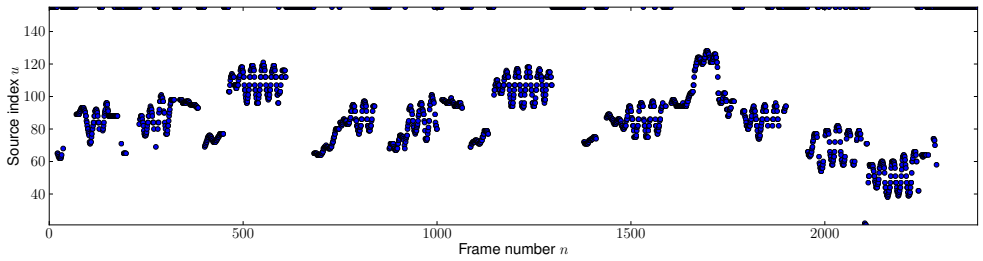
$$P = \frac{\#\{\text{correctly transcribed notes}\}}{\#\{\text{transcribed notes}\}} \tag{6.2}$$

$$F = 2\frac{RP}{(R+P)} \tag{6.3}$$

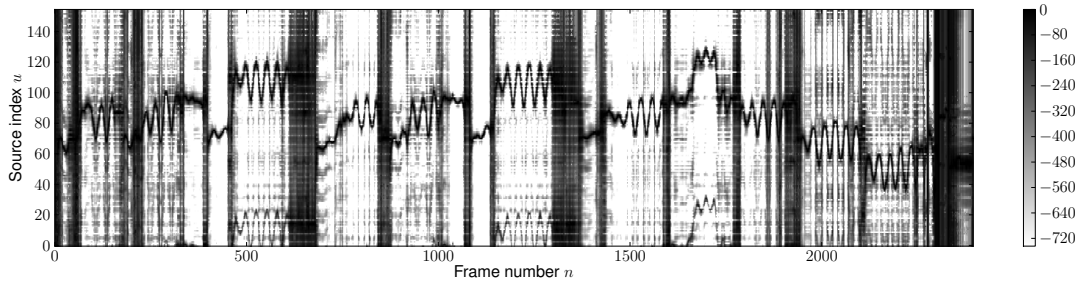
A transcribed note is considered correct if there is a note in the reference with the same MIDI note number of which the onsetting time is close to the one of the transcribed note.



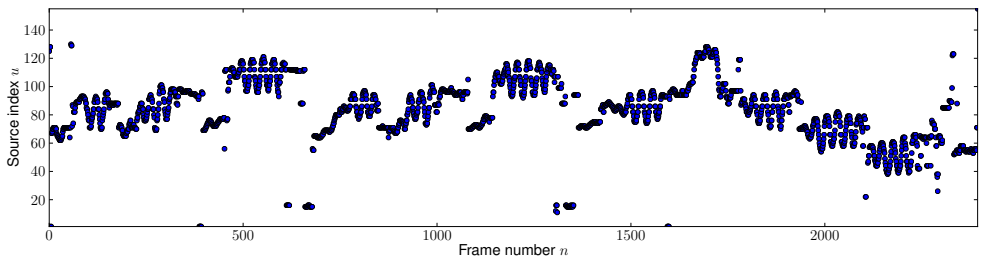
(a) F-I (GSMM): matrix of values $p(Z_n^{F_0} = u|\mathbf{x}_n)$



(b) F-I (GSMM): resulting optimal sequence $\widehat{Z}^{F_0}$



(c) F-III (HM-GSMM): matrix of values $p(Z_n^{F_0} = u|\mathbf{X})$



(d) F-III (HM-GSMM): resulting optimal sequence $\widehat{Z}^{F_0}$

Figure 6.4: Estimated (HM-)GSMM $p(Z_n^{F_0} = u|\mathbf{X})$, along with the corresponding "best path" $\widehat{Z}_n^{F_0}$. $\mathbf{W}^{\Phi}$, $\mathbf{W}^{\Phi}\mathbf{H}^{\Phi}$, $\mathbf{W}^{M}$ and $\mathbf{W}^{M}\mathbf{H}^{M}$, for the ADC 2004 song "opera_male5.wav".

The absolute difference between these onsetting times should be less than 25 ms. We compute the precision, recall, and f-measure, and we also computed the score obtained using the perceptually motivated measures in Daniel et al. [2008].

6.1.2.3 Results on a synthetic database (ISMIR 2009)

The system MUS-I, having been designed only recently, still has to prove its viability. However, due to several technical issues, especially because of the almost exhaustive search that is needed to solve it, only preliminary results on a synthetic database have been published [Weil et al., 2009b]. Using a time signature and temporally quantizing the melody to an estimated grid of beats leads to very promising results on a synthetic database, with musical scores visually acceptable and readable by musicians.

In order to assess the different modules presented in [Weil et al., 2009b], among which the proposed system MUS-I, a database for which the chords, the beat, and the melody line are annotated was needed. Assembling such a database by manually annotating audio recordings is highly time-consuming. The Band-In-A-Box[2] (BIAB) format then seemed a convenient way of generating the annotation in a semi-automatic way. BIAB is a software which generates musical accompaniment given a sequence of chords, a tempo, and a style; it also supports melody tracks. Thus, BIAB files contain all the information which is relevant for the lead sheet generation task: the key, the chord sequence, the tempo, the time-signature, and the melody notes.

We chose a subset of the Pop&Rock database gathered by members of the Yahoo BIAB user group. As the proposed main melody tracking module is explicitly modelled to match the human voice, the BIAB files were rendered using an oboe for the melody track; indeed, the acoustic characteristics of an oboe are very close to those of the human speech production sufficiently well. Due to the heavy computational load of MUS-I, we selected only 11 songs from this database, generated the MIDI files thanks to BIAB and annotated the files thanks to the corresponding melody track in the MIDI file. Details on these files may be found on-line[3].

On this database, the note tracking provided by MUS-I obtained an average recall, precision and f-measure of, respectively, 63 %, 68 % and 63 %. The average perceptive F-measure is 69 %. Fig. 6.5 also shows the box and whiskers for the 11 songs. The outlier corresponds to a song for which the melody was too fast and too variable for the melody tracker to follow. The results are promising; however, the BIAB subset was rather small and experiments on a bigger and more realistic database should be held before drawing conclusions. The results on the Quaero database may also convey some hints as about what should be done to improve this musical transcription.

6.1.2.4 Results for the Quaero evaluation campaign

In september 2009, the internal Quaero evaluation campaign aimed at measuring the performance of the different technologies developed within the project. The final purpose is to keep track of these technologies and be able to assess their improvement over time, during the project. The system MUS-I was therefore evaluated, both from the frame-wise and note-wise melody estimation.

[2] `http://www.band-in-a-box.com/`

[3] http://www.nue.tu-berlin.de/research/leadsheets/

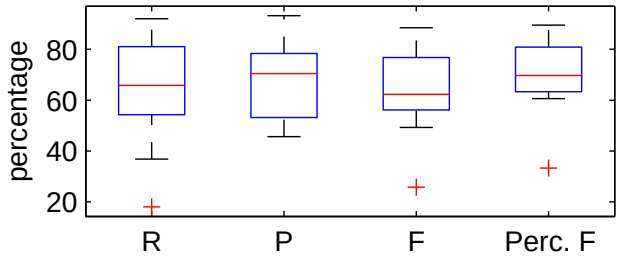


Figure 6.5: Box and whiskers plot of the results for melody estimation: Recall (R), Precision (P), F-measure (F) and perceptive F-measure (Perc. F).

The results obtained on the database described in Section D.2 are admittedly not as good as expected. **For the frame-wise melody estimation**, on the development dataset, MUS-I obtains a total average precision of 36 %, a recall of 58 % and F-measure of 44 %. As there were two specific subsets, a synthetic data one, generated from the MIDI files, and a real audio one, the original audio signals, the results were also computed separately, and MUS-I obtains $P = 30$ %, $R = 46$ % and $F = 36$ % on the synthetic dataset, while reaching $P = 42$ %, $R = 72$ % and $F = 53$ % on real audio excerpts. This apparently confusing result, where the system is able to perform better on real audio than on synthetic data, can be understood for the proposed system. Indeed, the synthetic generation of the data from the MIDI files could not be completely parameterized such that the resulting audio files seems to have some problems with the overall balance. Sometimes, the lead instrument actually does not lead the mixture, and is pretty weak in comparison to the other instruments. On the contrary, for the real audio, the professional sound engineering work made the lead instrument stand out of the mixture. Since MUS-I is mainly focussing on the energy pre-dominance of the melody, the observed better performances are not intuitively misleading. Compared with the MIREX evaluation, where the "raw pitch accuracy" can be almost identified with the recall R, reported here, the results are not very different. A significant difference between the Quaero dataset and the MIREX sets is that for the latter, it is mostly constituted of small excerpts of audio signals, in which most of the duration of the excerpt, the desired melody is present. The Quaero set, based on the RWC, is composed of complete songs. The melody is therefore not always present, during the introductions, the guitar breaks and the fade-out endings, for instance. The melody detection is therefore much harder, and reflects the melody definition ambiguity. A good recall on these songs means that most of the duration of the song, the melody was well transcribed. However, the relatively poor precision corresponds to the inability of the system to discriminate between the lead instrument, a singer voice in RWC-Pop, and the "accompaniment", such as a solo guitar.

These results were confirmed with the evaluation on the test set, with $P = 37$ %, $R = 58$ % and $F = 45$ %. Although we do not have the detailed results, it is very likely that the same phenomena about the differences between the synthetic and the real subsets may be observed.

As for the note-wise detection, on the Quaero database, it seems that MUS-I is not able to estimate the right notes. Indeed, with very poor results, $P = 8$ %, $R = 13$ %, $F = 9$ % and a MOR of 65 %, the system seems to have serious issues, be it on the synthetic or the real subset. A first element of answer about what happened could come again from the chosen database. The database is indeed constituted of pop songs, with

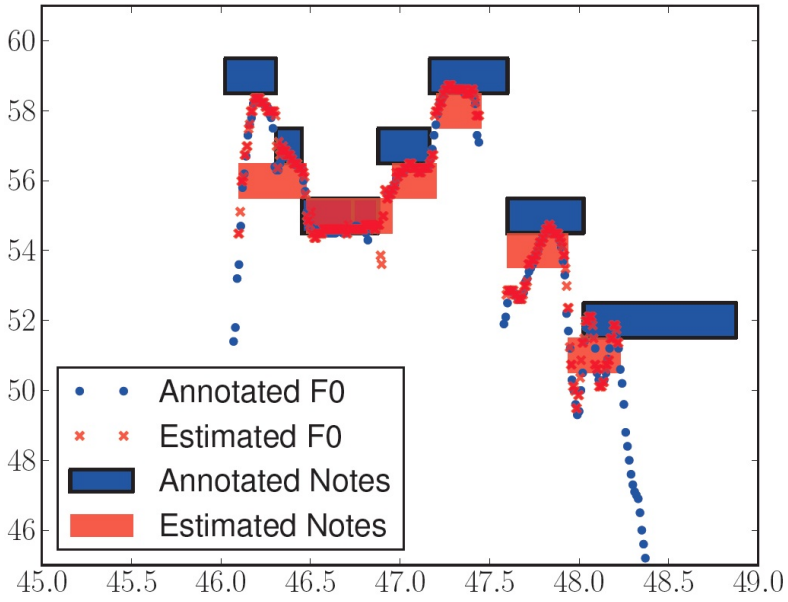


Figure 6.6: Example of result for the melody note tracking, on RWC-P009, from the RWC-Pop database.

a human singer as leading instrument. Due to wide variations of the pitch, the proposed method could only guess notes, but the duration model by itself does not seem to enable a musically relevant result, and including more musicological knowledge in the process, such as key, musical and harmonic context, may be necessary in order to improve the notewise estimation of the main melody. A closer analysis of which type of errors the system did should help in finding out what actually went wrong.

As a first step towards a more complete analysis, one can see on Figure 6.6 that, even for a rather "normal" performance from RWC-Pop (RM-P009), the annotated notes (blue rectangles, with black lining) are somewhat difficult to infer from the annotated frame-wise F0 (blue circles). While the estimated F0 (red crosses) follow rather well the ground-truth, the corresponding notes (light red rectangles) mostly miss the correct annotated note. In order to retrieve the annotated notes, which may be considered as the "intended" notes, one needs more high level information. Indeed, the intended melody notes are related to the accompaniment, and including this background information in the note decision can lead to improvements. Such a system was already proposed by Ryynänen and Klapuri [2008b]: the key is first estimated, and the note probabilities and transitions are changed accordingly. Such an improvement could also be added to the model proposed here, although this may add another high-level layer.

6.2 Audio separation of the main instrument and the accompaniment

The audio separation of the leading instrument from the accompaniment is an interesting application for MIR: it could for example be used as pre-processing for various applications such as instrument classification or other detection problems, whenever processing

separately these two streams may be useful. It can also generate the "de-soloed" accompaniment, suitable for further use in Karaoke applications.

Another interesting aspect of the separation results lies in the model validation that it allows. Indeed, it has been very useful to obtain and listen to the separation results, since they reflect how well our models fit the signal, especially for the leading instrument part. Chronologically, this separation application allowed us to validate the general source/filter approach [Durrieu et al., 2008a], then the possible refinement provided by the second pass estimation [Durrieu et al., 2009a], highlighting the "chirp effect" problem [Durrieu et al., 2008b] and at last the tentative approach to model the unvoiced components of the leading voice [Durrieu et al., 2009b].

The task definition is first recalled, then the adaptive Wiener estimation is discussed. The performance measures are defined, before the proposed leading instrument/accompaniment separation systems are further detailed. At last, the experiments are discussed. These last two sections were mostly published as [Durrieu et al., 2009a] and [Durrieu et al., 2009b].

6.2.1 Task definition

The goal of this task is to separate the leading instrument from the rest of the mixture, and to estimate the separated accompaniment. To limit ambiguity for this task, the signals are restricted to the ones with a clearly distinct leading instrument, such as a singing voice or a wind instrument (saxophone, trumpet, etc.).

6.2.2 Wiener filters

The adaptive Wiener filter [Benaroya et al., 2006] method has shown successful in obtaining separated signals with fewer artefacts, especially compared with many other separation back-end such as sinusoidal models or optimal binary masks. In this section, the optimal estimator for the separated leading instrument/accompaniment, namely the Wiener estimator, is presented and its practical application in the case of the present work is discussed. At last, some properties of the adaptive Wiener filtering approach are developed.

In the case of leading voice separation, the estimated voice STFT $\widehat{\mathbf{V}}$ is the expectation of $\mathbf{V}$, conditionally upon the observed mixture $\mathbf{X}$. The linear filter that minimizes the mean squared error between $\mathbf{V}$ and the estimated $\widehat{\mathbf{V}}$ is the Wiener filter. For stationary signals $\mathbf{v}_n$ and $\mathbf{x}_n$, at frame n, the filter is such that:

$$\widehat{\mathbf{v}}_n = E[\mathbf{v}_n|\mathbf{x}_n] = \underbrace{\frac{\mathbf{s}_n^{XV}}{\mathbf{s}_n^{X}}}_{\text{Wiener filter}} \mathbf{x}_n$$

where $\mathbf{s}_n^X$ and $\mathbf{s}_n^{XV}$ respectively are the PSD of (the time domain) x_n and the Fourier transform of the cross-correlation between x_n and v_n, and where all the operations are meant element by element.

With the mixture $\mathbf{X} = \mathbf{V}+\mathbf{M}$, the independence assumption between $\mathbf{V}$ and $\mathbf{M}$ implies that the auto-correlation term is the sum of the individual PSDs: $\mathbf{s}_n^X = \mathbf{s}_n^V + \mathbf{s}_n^M$ and the cross-correlation term becomes the PSD of $\mathbf{V}$: $\mathbf{s}_n^{XV} = \mathbf{s}_n^V$. The Wiener filtering formula therefore becomes:

$$\widehat{\mathbf{v}}_n = \frac{\mathbf{s}_n^V}{\mathbf{s}_n^V + \mathbf{s}_n^M}\mathbf{x}_n$$

With the proposed Gaussian model, under mild conditions (w.s.s.), the estimated variances of $\mathbf{V}$ and $\mathbf{M}$ can be assimilated to the corresponding PSDs (see Proposition 1). The computation of the above formula then becomes straightforward.

Although there is no proof that applying this filtering technique on a frame-by-frame basis is optimal for this problem, the experiments related later in this section tend to show that the obtained results are subjectively as well as objectively satisfying. Note however that Benaroya [2003] admitted that the STFT obtained by applying this time-frequency masking to the original mixture STFT may not be a "consistent" STFT, in the sense that it actually is not the STFT of any audio signal. Some works [Griffin and Lim, 1984, Le Roux et al., 2008] have already aimed at tackling that problem.

The Wiener filtering interpretation of the estimation proposed in [Benaroya et al., 2006] may seem quite far-fetched, especially since they require some strong assumptions such as the independence of both contributions and their stationarity. It is interesting to note that the Wiener estimate of $\mathbf{v}_n$ can however be derived directly using the Gaussian assumption by developing the posterior mean of $\mathbf{v}_n$ knowing $\mathbf{x}_n$, $E[\mathbf{v}_n|\mathbf{x}_n]$, as is proved in Appendix A.1.2. The adpative Wiener filter approach is also conservative. More precisely, adding all the estimated signals together gives the original mixture signal.

Furthermore, thanks to the Wiener filter approach, some model imprecisions can be blurred out. This is also a consequence of the Gaussian model that discards the phase information. Indeed, the correct phase of the STFTs could also be estimated, but as proved by Ephraim and Malah [1984], the best phase estimator is the original mixture phase. The phase information corresponds to the very fine structure of the sound. Trying to fit too closely to the phase, as is done typically with sinusoidal models, would lead to a higher degradation as soon as the model does not fit to the data anymore.

6.2.3 Performance measures

We evaluate our systems with the **BSS_eval** criteria, introduced by Vincent et al. [2006] and further developed in [Vincent et al., 2009], namely: the Signal to Distortion Ratio (SDR), the Image to Spatial distortion Ratio (ISR), the Source to Interference Ratio (SIR) and the Sources to Artefacts Ratio (SAR). These metrics decompose the estimated time-domain signal onto the different signal spaces formed by the individual sources. In the case of two sources, we define the "images" of the lead instrument $v^{imag} = \{v_j\}_{j=1\ldots J}$ and the accompaniment $m^{imag} = \{m_j\}_{j=1\ldots J}$, with $J \in \{1, 2\}$ channels. The toolbox provided by Vincent et al. [2009] decomposes the estimated signals $\widehat{v}^{imag}$ and $\widehat{m}^{imag}$ such that:

$$\widehat{v}_j = v_j + e_j^{v,spat} + e_j^{v,interf} + e_j^{v,artef} \tag{6.4}$$

$$\widehat{m}_j = m_j + e_j^{m,spat} + e_j^{m,interf} + e_j^{m,artef} \tag{6.5}$$

where, in the lead voice decomposition Equation (6.4) (the discussion for Equation (6.5) is of course analogous), the right hand side operands respectively are:

- the original "target" source,
- the spatial error, such that $v_j + e_j^{v,spat}$ computed as the projection of $\widehat{v}_j$ onto the space spanned by delayed version of v^{imag}, so as to take into account the potentially badly estimated mixing parameters,
- the interference error, which is the projection of $\widehat{v}_j$ onto the space spanned by delayed versions of the other source, *i.e.* m^{imag},

- and the artefact error, which corresponds to the residual error, which can not be "explained" by any of the original sources.

These separated errors may suffer from some problems, especially as concerns the way they are computed: it actually corresponds to yet another source separation problem, and is therefore still an open issue.

The ISR, SIR, SAR and SDR are then defined as:

$$ISR_v = 10\log_{10}\frac{\sum_{jt} v_{jt}^2}{\sum_{jt}(e_{jt}^{v,spat})^2} \tag{6.6}$$

$$SIR_v = 10\log_{10}\frac{\sum_{jt}(v_{jt}+e_{jt}^{v,spat})^2}{\sum_{jt}(e_{jt}^{v,interf})^2} \tag{6.7}$$

$$SAR_v = 10\log_{10}\frac{\sum_{jt}(v_{jt}+e_{jt}^{v,spat}+e_{jt}^{v,interf})^2}{\sum_{jt}(e_{jt}^{v,artef})^2} \tag{6.8}$$

$$SDR_v = 10\log_{10}\frac{\sum_{jt} v_{jt}^2}{\sum_{jt}(e_{jt}^{v,spat}+e_{jt}^{v,interf}+e_{jt}^{v,artef})^2} = 10\log_{10}\frac{\sum_{jt} v_{jt}^2}{\sum_{jt}(v_{jt}-\widehat{v}_{jt})^2} \tag{6.9}$$

At this point, one should also identify another issue: the definition of the sources themselves. Once we have a database, with the original separated sources, we can of course compute the above criteria. However, one may wonder how reliable they are, and what meaning one can give to them. For instance, let us take a song in which there is a piano, playing chords and a melody on top of it, in a single track. Of course, as a human listener, one may not discuss the fact that this track constitute one source. However, consider the proposed systems which could separate the chords from the melody: it is doubtful that the decomposition onto the original piano track would give any result, and therefore both the separated melody and chords may be decomposed as "artefacts". Of course this could be solved if one had access to a database with original sources organized such that several semantic level could be considered. Such a database may, however, neither be realistically feasible, nor desirable. This paragraph should however point the fact that, apart from the SDR, these values should be handled with care. The SDR as defined in Equation (6.9) does not depend on these projections, and is equivalent to a Signal-to-Noise Ratio (SNR), where the term "noise" means the reconstruction error between the original source and the estimated one.

The perceptual relevance of these measures has however been demonstrated in [Kornycky et al., 2008]. Another advantage is that it does not rely on any assumption about how the evaluated separation system works: such definitions for the criteria allows to use them for a wide range of separation system, regardless of what these systems actually are doing, as long as their output is some digital audio signal.

We also refer to SDR (resp. SIR) "gains" (gSDR and gSIR) as being the difference between the SDR (resp. SIR) obtained by the estimated tracks and the SDR (resp. SIR) computed by setting the original mono or stereo mixture as the estimated track.

6.2.4 Proposed source separation systems

During this study, we have designed two main separation systems. The first one was formally introduced in [Durrieu et al., 2009a], with the signal model grounded in [Durrieu

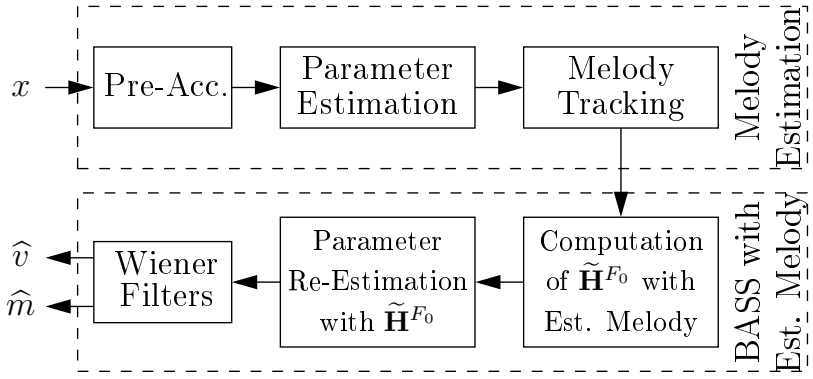


Figure 6.7: Solo/Accompaniment Separation: algorithm outline [Durrieu et al., 2009b].

et al., 2008a]. The second system was proposed in [Durrieu et al., 2009b] with several important contributions, notably the extension to stereo audio signal processing, the spectral smoothness of the filters, and the addition of the estimation of the unvoiced parts of the lead instrument.

The first system SEP-I is recalled, and the general algorithm flow is described. Then the extension to stereo signals, and the implications for the signal models are explained. The parameter estimation algorithm for stereo audio signals is then given. In the following sections, the parameter set is the SIMM set, $\mathbf{\Theta} = \mathbf{\Theta}^{\text{SIMM}}$. For the experiments on stereo signals, some additional parameters, the mixing parameters, are also included in $\mathbf{\Theta}$. For some of these experiments, we also explicitly tested the IMM set.

6.2.4.1 System SEP-I for mono music audio signals

Figure 6.7 shows the outline of the first complete blind audio source separation algorithm. It consists of two steps: the first one mainly aims at tracking the pitch contour (or melody) of the solo instrument. The second step estimates the parameters using the sequence of fundamental pitches estimated in the first step.

1. **Melody Estimation Step**

   Pre-accentuation: The mixture signal x is pre-accentuated with a conventional first order moving-average filter, with parameter $a = 0.95$.

   Model parameter estimation: A set of parameters $\mathbf{\Theta}_0$ is randomly generated. At iteration $i \leq I$, $\mathbf{\Theta}_{i-1}$ is updated to $\mathbf{\Theta}_i$ thanks to the multiplicative updates in Algorithm 5.2, with $I = 300$ the number of iterations.

   Melody tracking: We use the Viterbi smoothing algorithm 5.4 to retrieve the melody Z^{F_0}: $\widehat{Z}_n^{F_0}$ is the estimated state corresponding to the fundamental frequency of the solo instrument at frame n. It is important to note that our approach suffers from some ambiguities and this especially for the source model: the model as such allows the main instrument to be polyphonic, while we are interested in monophonic instruments. The smoothing step therefore has two goals: finding the smooth sequence of predominant pitches and limiting it to one pitch per frame.

   To circumvent some octave errors, a modified "a posteriori probability" matrix $\mathbf{G}^{F_0}$ is provided to the Viterbi algorithm: $g_{un}^{F_0} = h_{un}^{F_0} + 0.5 h_{(u+12U_{st})n}^{F_0}$. The initial algorithm tends to favor the estimation of the pitch as being the upper octave instead

of the fundamental frequency on some notes from the database. $\mathbf{G}^{F_0}$ is designed to compensate this effect. This heuristic was only introduced for the systems presented in [Durrieu et al., 2009a] and [Durrieu et al., 2009b]. The other solution proposed in this document, in Section 3.4.3, seems a more satisfying model, even though it still needs to be better studied and evaluated.

The silence frames are detected as explained in Section 3.3.2.3: the separated solo is first computed thanks to the Wiener filter masking. The energy for each frame is then computed and the frames with an energy lower than a given threshold are classified as silence frames of the solo. The threshold is chosen such that the energy of all remaining frames is above $(100 - \epsilon)\%$, where $\epsilon = 0.06$ in our system. The value of ϵ was fixed after some trials, and mainly aims at removing complete silence frames, when no instrument is playing at all, hence the chosen very low value.

2. **Source Separation Step**
Computing $\widetilde{\mathbf{H}}^{\mathbf{F_0}}$: The coefficients of $\mathbf{H}^{F_0}$ lying outside a scope of $\frac{1}{2}$ tone around the estimated melody are set to 0:

$$\begin{aligned}\widetilde{h}^{F_0}_{un} &= h^{F_0}_{un}, \text{ if } |\mathcal{F}(u) - \mathcal{F}(\widehat{Z}^{F_0}_n)| < \frac{1}{4} \text{ tone},\\ &= 0, \text{ otherwise.}\end{aligned}$$

Given this new matrix and the other parameters in $\mathbf{\Theta}$, the separated signals $\hat{v}_{(1)}$ and $\hat{m}_{(1)}$ can be computed. However, since $\mathbf{H}^{F_0}$ was modified, the estimated parameters are no longer optimal, especially for $\mathbf{W}^{\Phi}$, and a second estimation taking into account this new parameter matrix is necessary and improves the separation as shown by the results in section 6.2.5.
Parameter re-estimation: Again, $\mathbf{\Theta}_0$ is randomly drawn, except for the matrix $\mathbf{H}^{F_0}$ which is initially set to $\widetilde{\mathbf{H}}^{F_0}$. Since we use multiplicative updates, the null coefficients in $\widetilde{\mathbf{H}}^{F_0}$ do not evolve and stay null. The solo instrument is therefore limited to follow the estimated melody sequence $\widehat{Z}^{F_0}$ and the estimated parameters constrained to fit this melody (within $\frac{1}{2}$ tone).
Wiener filters: With the estimated final parameter set $\mathbf{\Theta}$, we obtain the separated signals $\hat{v}_{(2)}$ and $\hat{m}_{(2)}$. The pre-accentuation is compensated before comparison with the original sources.

6.2.4.2 Extension to stereo signals

In addition to the above signal model for SIMM, some major features have been developed: the support for stereophonic signals (and more generally for multi-channel signals) and the unvoiced components of the leading voice are also estimated thanks to a simple yet effective scheme.

First, the model needs to be extended to stereo audio signals. Let us consider an observed stereophonic sampled audio signal $[x^{\mathcal{L}}_t, x^{\mathcal{R}}_t]^T$, where t is the sample number, $x^{\mathcal{L}}$ (resp. $x^{\mathcal{R}}$) is the signal from the left (resp. right) channel. The $F \times N$ STFT of both channels are respectively denoted $\mathbf{X}^{\mathcal{R}}$ and $\mathbf{X}^{\mathcal{L}}$. The STFTs are assumed to be the instantaneous mixtures of two contributions, the solo part $\mathbf{V}$ and the accompaniment part $\mathbf{M}$.

The stereophonic aspect is modelled as a simple "panning" effect: the original sources are assumed monophonic and mixed together into the stereophonic signal by applying

different amplitude levels for each channel to simulate their spatial positions. The solo $\mathbf{V}$ is further assumed to have only one static spatial position and the accompaniment to be modelled with R several components, each of which have their own static spatial position. We assume that the STFTs, at frequency f and frame n, are given by:

$$\left\{ \begin{array}{rcl} x_{fn}^{\mathcal{R}} & = & \alpha_{\mathcal{R}} v_{fn}^{\mathcal{R}} + \sum_{r=1}^{R} \beta_r^{\mathcal{R}} m_{fn}^{\mathcal{R}r} \\ x_{fn}^{\mathcal{L}} & = & \alpha_{\mathcal{L}} v_{fn}^{\mathcal{L}} + \sum_{r=1}^{R} \beta_r^{\mathcal{L}} m_{fn}^{\mathcal{L}r} \end{array} \right. \tag{6.10}$$

where $\mathbf{V}^{\mathcal{R}}$, $\mathbf{V}^{\mathcal{L}}$, $\mathbf{M}^{\mathcal{R}r}$ and $\mathbf{M}^{\mathcal{L}r}$ are supposed to be realizations of random variables (r.v.). We assume that these r.v. are all mutually independent and individually independent across both frequency and time.

The stereophonic information is exploited only by considering that the signals for both channels (left and right) for one contribution $\mathbf{V}$ or $\mathbf{M}$ share the same statistical characteristics:

$$\left.\begin{array}{c} v_{fn}^{\mathcal{R}} \\ v_{fn}^{\mathcal{L}} \end{array}\right\} \sim \mathcal{N}_c(0, s_{fn}^{V}) \text{ and } \left.\begin{array}{c} m_{fn}^{\mathcal{R}r} \\ m_{fn}^{\mathcal{L}r} \end{array}\right\} \sim \mathcal{N}_c(0, s_{fn}^{Mr}) \tag{6.11}$$

$s_{fn}^{V} = s_{fn}^{V,(\mathrm{S})\mathrm{IMM}}$ and s_{fn}^{Mr} are the variances for the lead instrument signal and for the r^{th} component of the accompaniment, at frequency f and frame n. These variances are respectively given in Equation (3.37) and (indirectly) in Equation (3.15).

The resulting stereophonic signal therefore is distributed as follows:

$$\left\{ \begin{array}{rcl} x_{fn}^{\mathcal{R}} & \sim & \mathcal{N}_c\left(0, \alpha_{\mathcal{R}}^2 s_{fn}^{V} + \sum_{r=1}^{R} (\beta_r^{\mathcal{R}})^2 s_{fn}^{Mr}\right) \\ x_{fn}^{\mathcal{L}} & \sim & \mathcal{N}_c\left(0, \alpha_{\mathcal{L}}^2 s_{fn}^{V} + \sum_{r=1}^{R} (\beta_r^{\mathcal{L}})^2 s_{fn}^{Mr}\right) \end{array} \right. \tag{6.12}$$

6.2.4.3 Parameter estimation for stereo signals

The parameters to be estimated in this stereophonic framework are

$$\boldsymbol{\Theta} = \{\mathbf{H}^{\Gamma}, \mathbf{H}^{\Phi}, \mathbf{H}^{F_0}, \mathbf{W}^{M}, \mathbf{H}^{M}, \alpha_{\mathcal{R}}, \alpha_{\mathcal{L}}, \mathbf{B}^{\mathcal{R}}, \mathbf{B}^{\mathcal{L}}\},$$

where $\mathbf{B}_{\mathcal{R}} = \mathrm{diag}((\beta_r^{\mathcal{R}})^2)$. Thanks to the independency assumptions in time, frequency and between the channels, the log-likelihood $C(\boldsymbol{\Theta})$ of the observations writes:

$$C(\boldsymbol{\Theta}) = \sum_{fn} \log \mathbf{N}_c(x_{fn}^{\mathcal{R}}; 0, s_{fn}^{\mathrm{SIMM},\mathcal{R}}) + \log \mathbf{N}_c(x_{fn}^{\mathcal{L}}; 0, s_{fn}^{\mathrm{SIMM},\mathcal{L}}), \tag{6.13}$$

where the variances for the left and right channels are given thanks to equations (6.12), (3.37) and (3.15):

$$s_{fn}^{\mathrm{SIMM},\mathcal{R}} = \alpha_{\mathcal{R}}^2 \left[(\mathbf{W}^{\Gamma}\mathbf{H}^{\Gamma}\mathbf{H}^{\Phi}) \bullet (\mathbf{W}^{F_0}\mathbf{H}^{F_0})\right]_{fn} + \left[\mathbf{W}^{M}\mathbf{B}^{\mathcal{R}}\mathbf{H}^{M}\right]_{fn} \tag{6.14}$$

$$s_{fn}^{\mathrm{SIMM},\mathcal{L}} = \alpha_{\mathcal{L}}^2 \left[(\mathbf{W}^{\Gamma}\mathbf{H}^{\Gamma}\mathbf{H}^{\Phi}) \bullet (\mathbf{W}^{F_0}\mathbf{H}^{F_0})\right]_{fn} + \left[\mathbf{W}^{M}\mathbf{B}^{\mathcal{L}}\mathbf{H}^{M}\right]_{fn} \tag{6.15}$$

The parameter estimation is then done through a process similar to Algorithm 5.2. Indeed, the same structure can be inferred from the partial derivatives of the criterion, structure introduced in Section 5.2.2, hence leading very similar multiplicative updating rules, as shown in Algorithm 6.1, where $\mathbf{S}^{\Phi} = \mathbf{W}^{\Phi}\mathbf{H}^{\Phi}$, $\mathbf{S}^{F_0} = \mathbf{W}^{F_0}\mathbf{H}^{F_0}$, $\mathbf{D}^{\mathcal{R}} = |\mathbf{X}^{\mathcal{R}}|^{\cdot(2)}$, $\mathbf{D}^{\mathcal{L}} = |\mathbf{X}^{\mathcal{L}}|^{\cdot(2)}$, with the following convention for element-wise power notation: "$.(\omega)$". The

Algorithm 6.1 Updating rules for the SIMM on stereophonic signals:
Estimating $\boldsymbol{\Theta} = \{\mathbf{H}^{\Gamma}, \mathbf{H}^{\Phi}, \mathbf{H}^{F_0}, \mathbf{W}^{M}, \mathbf{H}^{M}, \alpha_{\mathcal{R}}, \alpha_{\mathcal{L}}, \mathbf{B}^{\mathcal{R}}, \mathbf{B}^{\mathcal{L}}\}$

$$\mathbf{H}^{F_0} \leftarrow \mathbf{H}^{F_0} \bullet \frac{(\mathbf{W}^{F_0})^T(\alpha_{\mathcal{R}}^2 \mathbf{S}^{\Phi} \bullet (\mathbf{S}^{\mathrm{SIMM},\mathcal{R}})^{\cdot(-2)} \bullet \mathbf{D}^{\mathcal{R}} + \alpha_{\mathcal{L}}^2 \mathbf{S}^{\Phi} \bullet (\mathbf{S}^{\mathrm{SIMM},\mathcal{L}})^{\cdot(-2)} \bullet \mathbf{D}^{\mathcal{L}})}{(\mathbf{W}^{F_0})^T(\alpha_{\mathcal{R}}^2 \mathbf{S}^{\Phi} \bullet (\mathbf{S}^{\mathrm{SIMM},\mathcal{R}})^{\cdot(-1)} + \alpha_{\mathcal{L}}^2 \mathbf{S}^{\Phi} \bullet (\mathbf{S}^{\mathrm{SIMM},\mathcal{L}})^{\cdot(-1)})} \tag{6.16}$$

$$\mathbf{H}^{\Phi} \leftarrow \mathbf{H}^{\Phi} \bullet \frac{(\mathbf{W}^{\Gamma}\mathbf{H}^{\Gamma})^T(\alpha_{\mathcal{R}}^2 \mathbf{S}^{F_0} \bullet (\mathbf{S}^{\mathrm{SIMM},\mathcal{R}})^{\cdot(-2)} \bullet \mathbf{D}^{\mathcal{R}} + \alpha_{\mathcal{L}}^2 \mathbf{S}^{F_0} \bullet (\mathbf{S}^{\mathrm{SIMM},\mathcal{L}})^{\cdot(-2)} \bullet \mathbf{D}^{\mathcal{L}})}{(\mathbf{W}^{\Gamma}\mathbf{H}^{\Gamma})^T(\alpha_{\mathcal{R}}^2 \mathbf{S}^{F_0} \bullet (\mathbf{S}^{\mathrm{SIMM},\mathcal{R}})^{\cdot(-1)} + \alpha_{\mathcal{L}}^2 \mathbf{S}^{F_0} \bullet (\mathbf{S}^{\mathrm{SIMM},\mathcal{L}})^{\cdot(-1)})} \tag{6.17}$$

$$\mathbf{H}^{M} \leftarrow \mathbf{H}^{M} \bullet \frac{(\mathbf{W}^{M}\mathbf{B}^{\mathcal{R}})^T(\mathbf{S}^{\mathrm{SIMM},\mathcal{R}})^{\cdot(-2)} \bullet \mathbf{D}^{\mathcal{R}} + (\mathbf{W}^{M}\mathbf{B}^{\mathcal{L}})^T(\mathbf{S}^{\mathrm{SIMM},\mathcal{L}})^{\cdot(-2)} \bullet \mathbf{D}^{\mathcal{L}}}{(\mathbf{W}^{M}\mathbf{B}^{\mathcal{R}})^T(\mathbf{S}^{\mathrm{SIMM},\mathcal{R}})^{\cdot(-1)} + (\mathbf{W}^{M}\mathbf{B}^{\mathcal{L}})^T(\mathbf{S}^{\mathrm{SIMM},\mathcal{L}})^{\cdot(-1)}} \tag{6.18}$$

$$\mathbf{H}^{\Gamma} \leftarrow \mathbf{H}^{\Gamma} \bullet \frac{(\mathbf{W}^{\Gamma})^T(\alpha_{\mathcal{R}}^2 \mathbf{S}^{F_0} \bullet (\mathbf{S}^{\mathrm{SIMM},\mathcal{R}})^{\cdot(-2)} \bullet \mathbf{D}^{\mathcal{R}} + \alpha_{\mathcal{L}}^2 \mathbf{S}^{F_0} \bullet (\mathbf{S}^{\mathrm{SIMM},\mathcal{L}})^{\cdot(-2)} . \mathbf{D}^{\mathcal{L}})(\mathbf{H}^{\Phi})^T}{(\mathbf{W}^{\Gamma})^T(\alpha_{\mathcal{R}}^2 \mathbf{S}^{F_0} \bullet (\mathbf{S}^{\mathrm{SIMM},\mathcal{R}})^{\cdot(-1)} + \alpha_{\mathcal{L}}^2 \mathbf{S}^{F_0} \bullet (\mathbf{S}^{\mathrm{SIMM},\mathcal{L}})^{\cdot(-1)})(\mathbf{H}^{\Phi})^T} \tag{6.19}$$

$$\mathbf{W}^{M} \leftarrow \mathbf{W}^{M} \bullet \frac{((\mathbf{S}^{\mathrm{SIMM},\mathcal{R}})^{\cdot(-2)} \bullet \mathbf{D}^{\mathcal{R}})(\mathbf{B}^{\mathcal{R}}\mathbf{H}^{M})^T + ((\mathbf{S}^{\mathrm{SIMM},\mathcal{L}})^{\cdot(-2)} \bullet \mathbf{D}^{\mathcal{L}})(\mathbf{B}^{\mathcal{L}}\mathbf{H}^{M})^T}{((\mathbf{S}^{\mathrm{SIMM},\mathcal{R}})^{\cdot(-1)})(\mathbf{B}^{\mathcal{R}}\mathbf{H}^{M})^T + ((\mathbf{S}^{\mathrm{SIMM},\mathcal{L}})^{\cdot(-1)})(\mathbf{B}^{\mathcal{L}}\mathbf{H}^{M})^T} \tag{6.20}$$

$$\alpha_{\mathcal{C}} \leftarrow \alpha_{\mathcal{C}} \frac{\mathrm{sum}\left(\mathbf{S}^{V,\mathrm{SIMM}} \bullet (\mathbf{S}^{\mathrm{SIMM},\mathcal{C}})^{\cdot(-2)} \bullet \mathbf{D}^{\mathcal{C}}\right)}{\mathrm{sum}\left(\mathbf{S}^{V,\mathrm{SIMM}} \bullet (\mathbf{S}^{\mathrm{SIMM},\mathcal{C}})^{\cdot(-1)}\right)}, \text{ for } \mathcal{C} \in \{\mathcal{R}, \mathcal{L}\} \tag{6.21}$$

$$\mathbf{B}^{\mathcal{C}} \leftarrow \mathbf{B}^{\mathcal{C}} \bullet \frac{(\mathbf{W}^{M})^T((\mathbf{S}^{\mathrm{SIMM},\mathcal{C}})^{\cdot(-2)} \bullet \mathbf{D}^{\mathcal{C}})(\mathbf{H}^{M})^T}{(\mathbf{W}^{M})^T((\mathbf{S}^{\mathrm{SIMM},\mathcal{C}})^{\cdot(-1)})(\mathbf{H}^{M})^T}, \text{ for } \mathcal{C} \in \{\mathcal{R}, \mathcal{L}\} \tag{6.22}$$

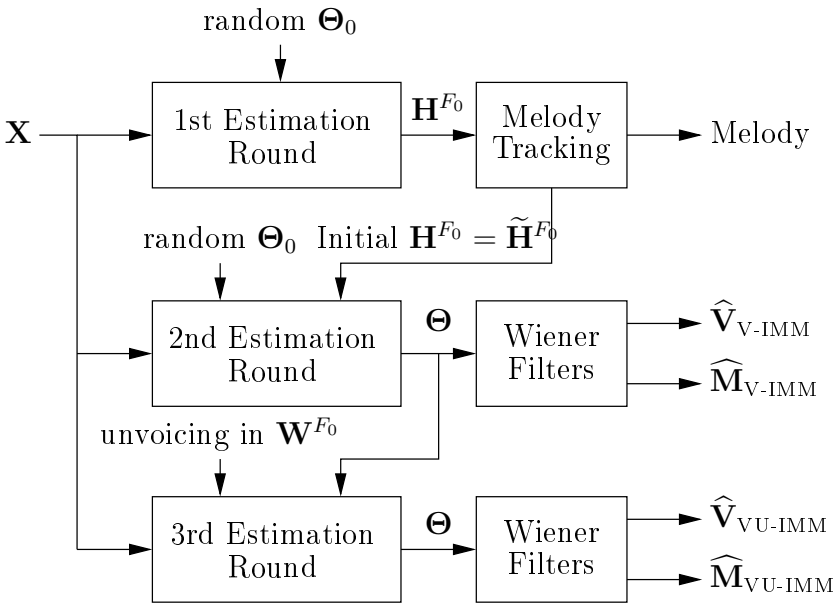


Figure 6.8: Solo/Accompaniment Separation System Flow [Durrieu et al., 2009b]

"sum" operator stands for a summation over all the elements of the argument matrix. The initial set of parameters is either randomly drawn (first round of estimation), with matrix $\widetilde{\mathbf{H}}^{F_0}$ as initial matrix for $\mathbf{H}^{F_0}$ (second round of estimation) or with previously estimated parameters (third round).

The proposed iterative system flow for the lead instrument/accompaniment separation for stereo signals is similar to our previously discussed mono signal system:

1. **1st (unconstrained) parameter estimation round** using Algorithm 6.1,
2. **Melody tracking:** a smooth path of fundamental frequencies is computed from the corresponding activation coefficients $\mathbf{H}^{F_0}$, using a Viterbi algorithm; the chosen path thus accomplishes a trade-off between its energy and the transitions between successive f_0 frequencies,
3. **2nd parameter estimation round** using Algorithm 6.1 and $\widetilde{\mathbf{H}}^{F_0}$ as initialisation for $\mathbf{H}^{F_0}$:
   - Solo and accompaniment separation using the corresponding Wiener filters, on each channel
     → "voiced"-IMM (V-IMM),
4. **3rd parameter estimation round**, including the "unvoicing" basis vector in $\mathbf{W}^{F_0}$ and with $\mathbf{W}^{\Phi}$ (i.e. $\mathbf{H}^{\Gamma}$) fixed:
   - Separation by Wiener filters, on each channel
     → "voiced+unvoiced"-IMM (VU-IMM).

Figure 6.8 also depicts the flow of the system. Each of the three rounds of parameter estimation correspond to 500 iterations of Algorithm 6.1. For the first round, the parameters are randomly initialized with a set $\mathbf{\Theta}_0$. For the second round, they are also randomly initialized, except the amplitude matrix for the solo source part which is initialized as in

Section 6.2.4.1: a matrix $\widetilde{\mathbf{H}}^{F_0}$ is obtained from the tracked main melody and the firstly estimated matrix $\mathbf{H}^{F_0}$ by setting to 0 all the coefficients that are outside a scope of a quarter tone from the estimated melody. These values remain null through the multiplicative rule in Algorithm 6.1. $\widetilde{\mathbf{H}}^{F_0}$ is then used as initial $\mathbf{H}^{F_0}$ matrix for the second estimation round. After this second round, we obtain a first solo/accompaniment separation result (V-IMM), where only the voiced parts of the solo were taken into account. We obtain stereophonic STFT "images" of the estimated solo $\widehat{\mathbf{V}}_{\text{V-IMM}}$ and accompaniment $\widehat{\mathbf{M}}_{\text{V-IMM}}$ by applying the corresponding Wiener filters, individually on each channel. These images are such that:

$$\widehat{v}_{fn}^{\text{imag}} = \begin{bmatrix} \alpha_{\mathcal{R}}\widehat{v}_{fn}^{\mathcal{R}} \\ \alpha_{\mathcal{L}}\widehat{v}_{fn}^{\mathcal{L}} \end{bmatrix} \text{ and } \widehat{m}_{fn}^{\text{imag}} = \begin{bmatrix} \widehat{m}_{fn}^{\mathcal{R}} \\ \widehat{m}_{fn}^{\mathcal{L}} \end{bmatrix},$$

where $\alpha_{\mathcal{R}}\widehat{v}_{fn}^{\mathcal{R}}$, $\alpha_{\mathcal{L}}\widehat{v}_{fn}^{\mathcal{L}}$, $\widehat{m}_{fn}^{\mathcal{R}}$ and $\widehat{m}_{fn}^{\mathcal{L}}$ respectively are the Wiener estimators of the right and left channels of the solo and of the right and left channels of the accompaniment. The audio time domain signals are then obtained by an overlap-add procedure applied individually on each channel of these STFT images.

At last, the initial parameters for the third round are the parameters estimated from the second round, except for $\mathbf{W}^{F_0}$, to which we add the unvoiced basis vector. We assume, by fixing the filter dictionary $\mathbf{W}^{\Phi}$ for this round, that the unvoiced parts of the solo instrument are generated by the same filters as for the voiced parts. The algorithm therefore catches unvoiced components whose spectral characteristics actually fit the previously estimated filter shapes. This new separation result is referred to as the VU-IMM, with the estimated images $\widehat{\mathbf{V}}_{\text{VU-IMM}}$ and $\widehat{\mathbf{M}}_{\text{VU-IMM}}$.

6.2.5 Experiments and results

The fundamental frequencies of the source part for the solo voice range from 60Hz to 1000Hz, with 96 source elements per octave. This results in $N_{F_0} = 391$ basis vectors in $\mathbf{W}^{F_0}$. For the filters, we chose $P = 30$ Hann atomic elements for $\mathbf{W}^{\Gamma}$, with an overlap rate of 75%, covering the whole frequency range. This corresponds to elementary smooth filters with a constant "bandwidth" of about 3kHz. The number of filters is fixed to $K = 9$ and the number of spectra for the accompaniment to $R = 50$.

6.2.5.1 Datasets

The dataset used for the experiments on mono audio signals is described in Appendix D.3. It is composed of 3 subsets: (A) the SiSEC 2008 development set for the "professionally produced music recordings" separation task[1], (B) some songs from Ozerov and Lagrange's private database (Ozerov et al. [2007] and Lagrange et al. [2008]) and (C) publicly available songs by S. Hurley, under Creative Commons licence. C is further divided into a pitch contour annotated set C1 and its complementary set C2. The songs are split into one-minute-long mono excerpts, discarding the ones that have no lead. The sampling rate for these songs is 11025Hz, the analysis window size is 512 samples (46.44ms) and the hopsize 64 samples (5.8ms).

The development database for the system proposed in [Durrieu et al., 2009b] is based on the multiple track recordings from the MTG MASS database [Vinyes, 2008]. We generated 13 synthetic instantaneous mixtures from the available multi-track data. The sampling rate

[1] Details and software available online at: http://sisec.wiki.irisa.fr/

for this set is 44100Hz. The STFTs are computed on analysis windows of 2048 samples (46.44ms), with a hopsize of 256 samples (5.8ms, overlapping rate of 87.5%). We use a "sinebell" weighting window, both for analysis and synthesis.

In addition, within the Quaero Project, a source separation database was proposed and developed (see Appendix D.4). It is based on some audio signals for which the multiple tracks were available on the Internet, mainly gathered by the METISS team, at IRISA. For the lead instrument separation, only a sub-set of the database could be used. The songs used for the evaluation of the lead instrument/accompaniment tasks were the songs by Another Dreamer , Fort Minor, Glen Philips, Jims big Ego, Nine Inch Nails, Shannon Hurley, and Vieux Farka Touré.

Although during the actual evaluation campaign, the proposed BSS_Eval criteria could not be computed due to technical issues, some results on the development set (Another Dreamer) are reported. Listening to the resulting files for the test set may also carry some interesting information.

6.2.5.2 Melody Tracking Performance

On subset C1, the recall in terms of main melody detection is at around 70%, while the precision only scores at 50%. This is explained by the fact that our system essentially focuses on energetic cues to track the melody line, and not on timbral cues. It therefore tracks the solo even if the solo instrument changes during the excerpt.

Note also that the files from subset C1 for this melody estimation mainly correspond to full length songs, while the evaluations at MIREX for instance have so far focussed on shorter excerpts on which there actually is a main melody most of the time. This fairly disappointing result also shows that the system designed in this thesis may require some more effort in discriminating the different contributions of a sound. Concretely, it would need to include some mechanism allowing to detect what instrument, for instance, is playing the estimated melody.

6.2.5.3 Source Separation with the True Pitch Contour

We verify that the model is able to separate the desired signals when the true pitch contour is given. We validate it on the melody-annotated mono audio subset C1. We separate the contributions by skipping the "melody estimation" step and use the annotated groundtruth of the melody pitch sequence $Z_{\mathrm{GT}}^{F_0}$ to initialize $\widetilde{\mathbf{H}}^{F_0}$ in the mono version of SEP-I.

Table 6.5 summarizes the results of the proposed system given the pitch contour, named "Melody", for these songs along with two other cases: "Mixture" characterizes the criteria computed by setting $\hat{v} = \hat{m} = x$, "Wiener" gives the results with the optimal Wiener filter computed with the original separated contributions. The first line therefore shows how difficult the task is and the last one gives the theoretical performance limit.

For the main instrument as well as the accompaniment, the separation results are satisfying. Most of the interferences and artefacts correspond to the unvoiced part of the vocal part, which is not explicit in the model for the tested mono version of SEP-I. The unvoiced parts are therefore estimated as belonging to the accompaniment. The SEP-I mono system approach is bounded in average by the result given in table 6.5, and the good performances there validate the proposed model.

	Main Instrument			Accompaniment		
Method	SDR	SIR	SAR	SDR	SIR	SAR
Mixture	-6.2	-6.0	–	6.2	6.2	–
Melody	8.1	16.1	9.0	14.3	19.4	16.6
Wiener	11.9	21.3	12.8	15.5	26.5	16.6

Table 6.5: Evaluation criteria (in dB) for the method given the pitch contour on the mono audio signal dataset C1.

	Main Instrument			Accompaniment		
Subset	SDR	SIR	SAR	SDR	SIR	SAR
All (1)	1.6	5.4	5.2	7.7	14.6	10.7
A (2)	8.2	17.4	8.9	10.8	15.4	12.9
B (2)	2.4	6.6	5.2	8.5	14.6	11.7
C (2)	2.7	9.2	5.0	9.1	14.1	12.7
C1 (2)	3.5	8.2	4.1	9.7	14.2	12.6
All (2)	2.7	8.1	5.2	8.8	14.4	12.1

Table 6.6: Evaluation criteria (in dB) for our global system averaged over each mono signal subset.

6.2.5.4 Source Separation with Estimated Melody

Table 6.6 shows the results obtained by the proposed algorithm for each set of our mono signal database. The number into brackets indicates whether the separation is directly held after the melody extraction step (1) or after the second step (2). The mean "main instrument to accompaniment" ratio is -6.1dB. In average, the SDR/SIR gains obtained by the proposed iterative method on the database respectively are 8.8/13.8 for the solo voice and 2.6/8.0 for the accompaniment.

The figures in table 6.6 first show the improvement of our iterative approach (2) compared to the direct separation after the melody estimation (1). Informal listening tests confirm that the parameter re-estimation really improves the quality and selectivity of the separation. It also seems that most of the interferences are due to estimated fundamental frequencies belonging to instruments of the "accompaniment", especially on the Celtic rock songs from subset B. In those songs, $\widehat{v}$ often corresponds to musical instruments performing solos and not to the expected singer voice. It is worth noticing that our algorithm is designed to track the main melody without assuming timbre coherence. It is therefore possible to obtain a main solo track $\widehat{v}$ played by different subsequent instruments.

In spite of these drawbacks, our results compare well with the state of the art. In Ozerov et al. [2007], the authors report, for their supervised system, a SDR gain of 10.5dB for the separated voice, while our system obtains an average of 8.5dB SDR gain on the corresponding subset B, without the voice/music automatic segmentation as pre-processing and no learning step, since our approach is unsupervised. In Ryynänen et al. [2008], the separated accompaniment obtains a SDR gain average of 0.8dB at a -5dB "main instrument to accompaniment" ratio, while we obtain a SDR gain of 2.6dB. Some separation results are available on our webpage at `http://perso.telecom-paristech.fr/grichard/icassp09/`.

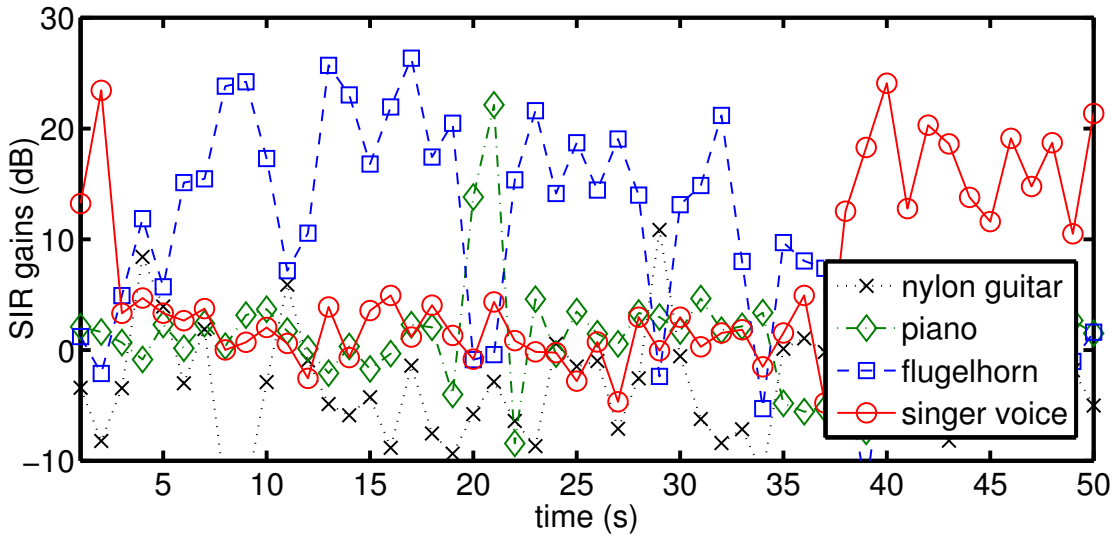


Figure 6.9: Evolution of SIR gains and "solo sections" for 4 instruments: guitar, piano, flugelhorn and singer.

6.2.5.5 Multitrack example

In order to give a deeper insight of our algorithm, we further analyze it on the 3rd excerpt of "We Are In Love" (S. Hurley), for which we have the 8 separated tracks of each of the instruments of the song. Figure 6.9 shows the evolution of the SIR gain of the estimated $\widehat{v}$ over the mixture for 4 cases, depending on the instrument we consider as the "main instrument", i.e. setting v to either of the following tracks: the guitar, the piano, the flugelhorn and the singer.

In this excerpt, the singer finishes her phrase at $t = 3$s and sings again at $t = 38$s. The flugelhorn plays from $t = 5$s to $t = 38$s. The piano and the guitar also have solo notes at $t = 20$s and $t = 29$s. As shown on Figure 6.9, the SIR gains are maximum for these instruments at the times where they are soloing: our system successfully separates the predominant instrument. These excerpts and the corresponding files, as well as some other separation examples, are available on our webpage at `http://perso.telecom-paristech.fr/grichard/icassp09/`.

6.2.5.6 Stereo signal + unvoiced extension

We report the results for 5 sub-systems from SEP-I in table 6.7: V-IMM and VU-IMM, each without (0) and with (1) the smooth filter model. The system "Mono" is the monophonic version of system SEP-I (as introduced in [Durrieu et al., 2009a]), applied separately to each channel. Some audio examples are also available on-line at `http://perso.telecom-paristech.fr/durrieu/en/eusipco09/`.

6.2.5.7 Smooth filters and unvoicing model

The results in table 6.7 first show that the performances in source separation with and without the smooth filter algorithm are not significantly different. This feature does not

Method	SDR	ISR	SIR	SAR	gSDR	gSIR
Mono	5.8/6.9	9.0/21.8	16.8/9.6	5.8/11.5	6.9/5.8	17.8/8.5
V-IMM0	7.9/8.9	12.1/23.0	19.2/12.6	8.2/12.5	8.9/7.9	20.2/11.6
V-IMM1	7.9/8.9	12.5/22.1	18.4/12.8	8.3/11.6	8.9/7.9	19.4/11.8
VU-IMM0	**8.2**/**9.3**	12.4/**23.3**	**19.9**/12.9	8.7/**12.7**	**9.3**/**8.2**	**20.9**/11.8
VU-IMM1	**8.2**/**9.3**	**13.0**/21.8	18.6/**13.2**	**8.8**/12.0	**9.3**/**8.2**	19.6/**12.2**

Table 6.7: Average results on the stereo audio signal database database, in dB. For each criterion: estimated solo/estimated accompaniment.

seem to be able to discriminate the timbre of the solo instrument from the accompaniment ones, since the extracted solo occasionally switches from the desired instrument to some instruments of the accompaniment. The main interest of obtaining smooth filters however lies in the better "semantics" that these spectral shapes may convey, rather than the direct improvement in source separation. As part of a production model for the solo instrument, the smoothness of the filters is more realistic than having unconstrained filters. It may therefore be useful for recognition purposes. In a supervised framework, it can also be used to learn spectral shapes that are characteristic for a given instrument.

The unvoicing model seems to lead to better results since VU-IMM in general obtained better results than V-IMM on our database. However, the difference between the criteria is not significantly high, and informal listening of the estimated tracks reveals that most of the unvoiced parts that are caught actually correspond to drum sounds. We also noticed that only some of the desired unvoiced solo parts are extracted: especially for the excerpts by "Tamy", from the SiSEC "Professionally Produced Music Recordings" [SiSEC, 2008], with a guitar as accompaniment instrument, some consonants are missing in the extracted solo. This may show that the unvoicing model, i.e. assuming that the unvoiced parts share the same filter shapes as the voiced parts, is not complete and may need to be further extended in order to take into account the other potential unvoiced components.

6.2.5.8 Stereophonic vs. monophonic algorithm

In order to compare the monophonic version of SEP-I and its stereo version in the same conditions, we create a "pseudo"-stereophonic result from the monophonic result by applying the algorithm on both channels, separately and independently.

The table 6.7 shows that the performances are significantly improved by the use of the stereophonic versions of SEP-I. In the stereophonic framework, the melody is estimated once for both channels and the energy variation for a single contribution, e.g. the solo, of one channel is therefore proportional to that of the other channel: the result is therefore more coherent, and this way we avoid obtaining separated signals that are randomly "floating" from one side to the other. Applying the monophonic algorithm independently on each channel does not guarantee this coherence.

Contrary to [Ryynänen et al., 2008], our approach is also general enough to deal with "truly" stereophonic signals, even when the solo instrument is not exactly panned in the middle: allowing this flexibility therefore improves the separation of the accompaniment.

System	Singer SDR	Guitar SDR
Cancela2	**9.7**	8.6
SEP-I VU-IMM	7.8	**9.4**
Cancela1	8.7	8.0
SEP-I V-IMM	6.9	8.6
Cobos	6.4	8.0
Ozerov	5.1	6.7
Ozerov/Févotte	3.6	5.3
Vinyes Raso	4.9	4.2
Ideal Binary Mask	*10.1*	*11.8*

Table 6.8: Result table for SiSEC 2008 (song "Tamy - Que Pena / Tanto Faz")

6.2.5.9 SiSEC campaign results

At last, early versions of the proposed systems were evaluated at the SiSEC evaluation campaign for "Professionally Produced Music Recordings" [SiSEC, 2008]. We provided the extracted female singer voice and the extracted background music (the guitar) for the first song by "Tamy" using two of the aforementioned methods, V-IMM and VU-IMM, both with smoothed filters.

The results in terms of SDR are given in table 6.8. Details for the other systems can be found in [Vincent et al., 2009]. We ordered the systems by decreasing mean between the obtained SDR for the singer and guitar extractions. The result from the previous sections are confirmed, since VU-IMM performs better than V-IMM. Compared to the other participants, VU-IMM achieved the second mean SDR value, after Cancela [2008], whose systems share an interesting similarity with our algorithms: **they also explicitly model the solo part using the melodic line** (with the fundamental frequencies in Hz). The results obtained during this evaluation tend to prove or at least to validate the use of this information in order to successfully separate (monophonic) melodic instruments. This type of "knowledge-based" approaches are then to be compared with more classical approaches for source separation, more "data-driven". The drawback is that the assumptions on the signals are quite strong, and it is difficult to adapt the model in order to extract some other instrument such as the guitar, for instance, from the other song by "Bearlin" in the SiSEC evaluation: such polyphonic instruments need a more complicated polyphonic pitch estimation step followed by a clustering step to determine which instrument played which estimated pitch.

6.2.5.10 Evaluation on the Quaero Source Separation Database

The above SEP-I system, with the stereo + unvoiced extension, was also tested on the database developed within the Quaero project. However, due to a lack of time and some problems with the evaluation software, the performance on the test set is not available at the time of writing this section.

The results on the development set, which was constituted of one song, Another Dreamer - *One we love*, at all the possible mixes, are rather uneven. First, for this task, the full songs were provided, and an arbitrary segmentation of these songs was performed before any processing. The segment length that was chosen was 1 minute. To discuss the

results a relevant way, the segments where the lead voice was absent were not taken into account during the computation of the average results.

For the six different mixing conditions that were provided for that song, first considering the original mixture as the estimates for both the lead and the accompaniment, the average $SDR0_v$, as computed by the BSS_eval toolbox, is -7dB. $SIR0_v$ is equal to $SDR0_v$, and, by definition, $ISR0_v$ and $SAR0_v$ are equal to infinity. The average SDR_v for the best performing combination of parameters is 0.8dB, obtaining a SDR_v gain of 7.8dB. The SIR_v gain is 14, 4dB. Note that the result for the extracted accompaniment is less optimal. The average SDR_m is 7.8dB, but that is only an improvement of SDR_m of 0.8dB. The average SIR_m is 1.7dB. This tends to show, once again, that the system SEP-I is well suited to enhance the lead instrument, without being completely successful in removing it from the mixture.

6.2.5.11 Note on the front-end melody estimation systems: F-I, F-II or F-III?

The choice of system F-II as front-end for SEP-I was first motivated by its good trade-off between the melody estimation and the speed at which the estimates are obtained. Without technical constraints, it could be interesting to try F-I or F-III as front-end and analyze the results in terms of lead instrument and accompaniment separation.

The preliminary results obtained by a system using F-I or F-III instead of F-II as the first step in SEP-I are not as promising as the ones using F-II. Since there were many approximations from F-I or F-III, with basically the (S)GSMM as frame-wise model, to F-II, with the (S)IMM frame-wise model, the advantages of the latter on the former are not obvious, especially since the performance in predominant F0 estimation is almost the same for F-I and F-II. An explanation may be the choice of keeping in $\widetilde{\mathbf{H}}^{F_0}$ the coefficients within a semi-tone of the estimated melody, and not just the coefficient corresponding to the estimated $\widehat{Z}^{F_0}$. Indeed, considering the ill-fitted estimation for the GSMM, that is for system F-I as well as system F-III, shown on Figure 6.3, the obtained time-frequency Wiener mask does probably "reject" most of the energy of high frequency lobes to the accompaniment part.

To circumvent this problem, a potential improvement can be achieved by combining the strength of both the GSMM and the IMM approaches. Indeed, the link between the models was made explicit in Equation (3.51), within a Bayesian framework. By designing some conditional prior for the parameter set, $p(\mathbf{\Theta}_n^{(\mathrm{S})\mathrm{IMM}}|Z_n^{\Phi}=k, Z_n^{F_0}=u)$, one could model chirps for instance as a linear combination of several atoms of source dictionary $\mathbf{W}^{F_0}$. Note however that this would lead to some estimation algorithm closer to the GSMM estimation. The possibility of directly including the spectra of chirped harmonic signals in $\mathbf{W}^{F_0}$ is also under investigation.

Chapter 7

Conclusion

7.1 Conclusions

We have proposed two generic models for "singing voice + accompaniment" mixture signals, for two specific applications, the main melody transcription and the separation of the lead instrument from the accompaniment.

The first model for the leading instrument is a Gaussian Scaled Mixture Model (GSMM) where the hidden states, thanks to a source/filter model, explicitly involve the fundamental frequency of the signal, hence creating a direct bound between the estimation of the state sequence and the estimation of the melody in terms of fundamental frequencies. The second model for the lead voice is, to a certain extent, a generalization of the GSMM: the signal is assumed to be the combinations of all the hidden states of the GSMM, hence the name Instantaneous Mixture Model (IMM).

In each model, the accompaniment part is always modelled as a combination of independent Gaussian components, for which the estimation process turns out to be equivalent to a Non-negative Matrix Factorization (NMF) problem, where the matrix to be factorized is the power spectrogram of the signal and the reconstruction error measure is the Itakura-Saito (IS) divergence. The parallel between our framework and NMF methodology is essential for the derivation of the proposed algorithms.

For all the models, the temporal relations that command the pitch sequence are incorporated within two additional layers: a physical layer, such that the pitches form a smooth melody line, and a more "musicological" layer, that constrains the melody to form realistic notes, in terms of their durations. Two additional improvements to our original source/filter models, the GSMM and the IMM, were also proposed: first the filter part smoothness was structurally constrained by parameterizing the filters as linear combinations of smooth functions. Second, for the signal separation purposes, an element corresponding to unvoiced parts of the leading instrument was included in the source spectral dictionary.

Five systems have been proposed, F-I, F-II, F-III, MUS-I and SEP-I. The first three systems aim at estimating the fundamental frequency sequence, as a frame-wise result. MUS-I is designed to return a sequence of notes corresponding to the melody and SEP-I separates the input signal into two audio signals, the estimated leading instrument and the estimated accompaniment.

The different experiments on F-I and F-II tend to show that the estimations by these systems are good and reliable, especially since the use of system F-II as a pre-processing for system SEP-I leads to excellent performance in terms of source separation. F-I and

F-II were evaluated within international evaluation campaigns, MIREX 2008 and MIREX 2009, where they also proved to be at the state of the art. F-III, which relies on the GSMM with the temporal constrains (HMM) directly included in the parameter estimation, still needs to be evaluated and further studied. Surprisingly, preliminary results are not as positive as expected: while F-III should be an improved version of F-I, it does not seem to perform as well as that system. More experiments are necessary before any conclusion can be drawn and before any solution can be found.

MUS-I allowed to provide encouraging results on a synthetic database, but obtained disappointing results on the RWC Popular subset. In real audio signals, when the lead instrument is a singing voice, the great variability of the pitch and the not-so-obvious link between the note and the actual corresponding physical fundamental frequency may explain the relatively disappointing results.

At last, the results of SEP-I, in its stereo version, were among the best in the SiSEC 2008 international evaluation campaign. An interesting result from that campaign was that the most promising systems were the ones using the melody line of the instrument to be separated, a female singer, performed best. This tends to prove that using music related information in order to separate signals leads to approaches that are not only viable, but also the best performing ones.

7.2 Potential improvements

Improving the results may be done by further studying different parts of the systems, from the signal models to the estimation process, or even modifying the general algorithm flows.

7.2.1 Even more "Musicological" model for note duration

The duration model can be informed using more musical knowledge, such as tempo, or rhythm. The *prior* densities of the durations could then depend on the estimated rhythms and tempi, with several modes instead of one, as proposed in this thesis. Each mode could correspond to a musical unit such as whole note, half note, quarter note and so forth. The tempo would then be used to convert these units into frame durations.

The tempo estimation could also be included within the different systems, as another layer, on top of the note layer E. However, the parameter estimation of such a model may not be feasible, since we discussed previously that the addition of the note layer and the note sequence estimation were already very challenging. Such a model could however be used in particular cases, for instance in score following, or alignment. Indeed, when the musical score is available, the note sequence estimation is not necessary, and the search space is much reduced, making it possible to perform the aforementioned tasks.

7.2.2 A more complex physical layer

The proposed HMM structure for the physical layer may also seem insufficient, regarding the complexity of the production process for natural instruments. Indeed, phenomena such as the vibrato may be modelled with longer temporal dependencies than the HMM frame-to-frame approach. Such a mechanism could be coupled with the note estimation, as the vibrato is closely related to the musical content.

7.2.3 Accompaniment model: towards more supervision?

The chosen simplicity of the accompaniment model may seem to contradict the desire to model a rather complex mixture of sounds. This part could be further processed, especially in order to highlight its differences with the "signal of interest", namely the singing voice.

Indeed, one of the worst drawback of the model as we have designed it so far is the inherent ambiguity between both representations: it is easy to note that the singing voice model can be included in the accompaniment model as soon as the dimension of the accompaniment spectral matrix $\mathbf{W}^M$ is big enough. As the results obtained by our algorithms show, the opposite, for instance a lead guitar being detected as the main melody, is more likely to happen. However, this only seems to be a matter of initial conditions for the algorithms and deserves further improvements.

A first improvement could be to learn the spectral shapes corresponding to the accompaniment. This of course requires that we know which instruments are playing and that what they are playing is sufficiently different from the lead instrument.

The above approach would also lead to a rather different conception of the task at hand, since it would then involve supervision. Another approach that would be closer to what was proposed in this thesis can be found in works by Févotte et al. [2009a]. By explicitly modelling the temporal transitions, that is to say the temporal evolution of each spectral shape, we could further characterize the corresponding instruments. Instead of the parametric model used in [Févotte et al., 2009a] one could also think of learning these evolutions for different instruments, including more complex models with hidden states corresponding to different steps in the "process" of a note, as in [Ryynänen and Klapuri, 2005]. Although it may sound illogical to assume that the leading instrument and the accompaniment are independent, this assumption is not completely irrealistic.

At last, as is done in [Vincent, 2004], the accompaniment could be modelled in the same manner as the lead instrument is in our work, with a description layer (fundamental frequency layer, or more general layer as in [Vincent, 2004]), and a note layer. This independence between the Fourier transform of the lead instrument $\mathbf{v}_n$ and the accompaniment $\mathbf{m}_n$, at frame n, assumed in Section 3.3, should then be stated as the independence of the signals, conditionally upon these higher description levels. The signals were generated by different instruments, different physical systems, such that their generation processes can be considered independent. However, the instruments are usually playing their parts from a common musical score, in the same scale, involving spectra sharing common partials. This is typically a musicological dependency, which could be further modelled within higher level layers (such as those proposed in Section 3.3.4). This would be a way of including more interaction between the lead instrument and the accompaniment, allowing to define the former in comparison to the latter, and not only in an "absolute" way. This would however be equivalent to performing a multiple note tracking within the excerpt, which may not always provide results that are flexible enough for the music excerpts our methods have so far been able to process.

7.2.4 Decidedly perfectible models...

The above propositions for improvements are only a very small portion of the directions in which our models could be further studied. Many of the assumptions we made could be discussed and other assumptions or other solutions could lead to even better performances.

First, the Gaussian statistical framework may not completely hold for audio signals. Indeed, why should we assume so much random variables when music signals are essentially

composed of sinusoids? There might be some hybrid model in which deterministic and stochastic parts appear and can be estimated, probably at the cost of the flexibility that our framework allows.

The mono-pitch assumption of the lead instrument may also be discussed, especially the underlying assumption that the pitch is constant within each frame. What about vibrato sections? It seems obvious that during the frames between the maxima of the fundamental frequencies of a vibrato (the "transition frames"), the frequency content can not be constant. The fundamental frequency could therefore be explicitly modelled, along with some more parameters to estimate the "slope" of the chirp. How can we however estimate that slope? Which approximation order is better suited to this task? If the order 0 (constant F0) is too coarse, how about order 1 (linear chirp) or 2 (quadratic chirp)?

The silence model we chose may also be improved in many ways, although the statistical signal model may not be the ideal framework in which to take silences into account, since this basically means taking variances equal to 0. One could also think of adding yet another high level description layer, whose states would reflect the instrumental content of the frames: is the lead instrument present or not? As we discussed previously, such classification schemes were already used in other works, but including the estimations directly within the systems within a unified framework with the other parameters could be an interesting challenge.

At last, the models we proposed, especially the source/filter GSMM with all the layers, is a very complex model. We have proposed several systems that approximately estimate the desired parameters and sequences. We have also sought to approach the joint estimation of the sequences and the parameters, as shown from system F-I (GSMM) to F-III (HM-GSMM). The note level could probably also be included in the estimation loop, although this may be quite demanding in terms of computational ressources. Seeking such a joint estimation scheme may also be questioned, since the systems MUS-I and SEP-I, who provide sequentially estimated parameters and state sequences, obtained results that were among the best ones during international evaluation campaigns. This shows that, although approximate and sub-optimal, the chosen sequential approaches may be good enough for the tasks we are considering.

Glossary

Fundamental frequency: for a periodic sound, with period T_0, the fundamental frequency f_0 is defined as the inverse of the period of that signal:

$$f_0 = \frac{1}{T_0}$$

Harmonic sound: this notion is related to the sinusoidal model [McAulay and Quatieri, 1986]: for a harmonic sound, the frequencies of the partials are all entire multiples of the fundamental frequency. Let H be the maximum number of harmonics for the considered sound, f_h the frequency for the h^{th} harmonic, with $h \in [0, H-1]$ and a_h the amplitude associated with the harmonic h. The harmonicity of the sound implies that $f_h = (h+1)f_0$. The harmonic signal s is then given by:

$$s(t) = \sum_{h=0}^{H-1} a_h \cos(2\pi f_h t) = \sum_{h=0}^{H-1} a_h \cos(2\pi(h+1)f_0 t)$$

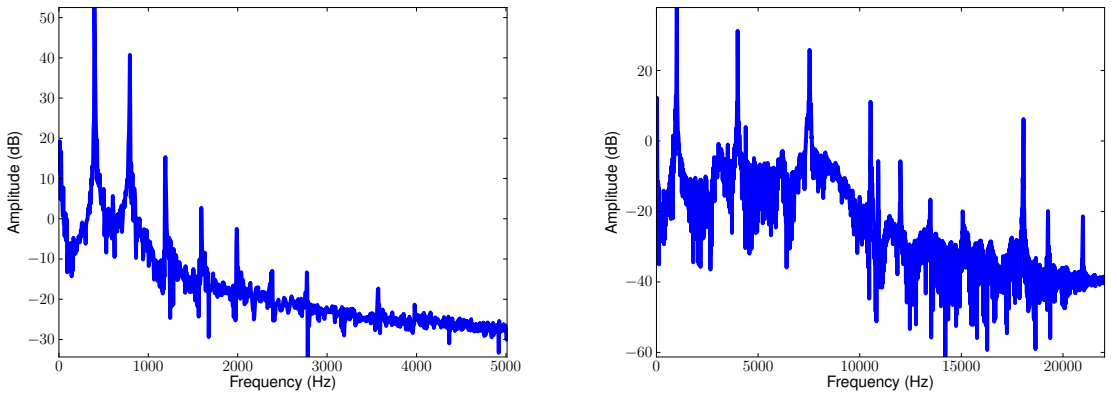
Harmonic sounds have a characteristic spectrogram, as can be seen on Figure 7.1(a): it shows regularly spaced peaks. This spacing actually equals the fundamental frequency of the corresponding sound.

Inharmonic sound: an inharmonic sound is usually understood to be a pitched sound which is not harmonic. Some instruments are inherently inharmonic, such as percussive instruments. An example of a glockenspiel spectrum is given on Figure 7.1(b). Some instruments like the piano or the guitar are also slightly inharmonic, but the frequencies of the partials are almost in harmonic relation. They can be considered as "quasi-harmonic" instruments.

MIDI: Musical Instrument Digital Interface, widely spread numeric annotation format for music scores.

Period: the period of a time-domain signal y, if it exists, is the smallest duration $T \in \mathbb{R}$ such that, for all $t \in \mathbb{R}$, $y(t) = y(t+T)$. A signal for which such a quantity exists is called a periodic signal.

"Pitch": the pitch of a complex sound is the frequency to which a listener needs to tune a sinusoid such that it sounds at the same "height". It is therefore a sensation related to psychoacoustics. When two sounds have the same height, they have the same pitch.



(a) Horn excerpt (harmonic sound)

(b) Glockenspiel excerpt (inharmonic sound)

Figure 7.1: Spectrum of a harmonic sound and of an inharmonic sound.

For a sinusoid or a periodic signal, the pitch can very often be identified with the fundamental frequency of the signal. However, in some cases, even when the signal is not strictly periodic, a human listener may also be able to identify a pitch.

In this work, the "psychoacoustic pitch" will not be sought for. This study mainly focuses on the analysis of the sounds themselves, and not on the perceptual effects they can have. In the remainder of this document, the term "pitch" will be used with the meaning of "fundamental frequency", as an objective quantity which can be associated with a periodic sound without ambiguity. Note that, as will be seen later, the proposed model for the lead instrument relies on the harmonicity of the sound, for which the (psychoacoustic) pitch usually coincides with the fundamental frequency.

SMF: Standard MIDI File. See "MIDI".

Appendix A

Probability density function definitions

In this Chapter, we define the probability density functions used in this document, and discuss some of their properties.

A.1 Complex proper Gaussian distribution $\mathcal{N}_c$

A.1.1 Complex proper Gaussian distribution definition

A complex proper Gaussian random variable is a complex random variable whose real part and imaginary part are independent one from the other, each of which following a (real) Gaussian distribution, with the same parameters: mean equal to $\mathbf{0}$ and identical variance (co-variance matrix in the multi-variate case).

For the proposed framework, the expression of the likelihood of complex proper Gaussian random variable is needed. Let $\mathbf{z} = \mathbf{x} + j\mathbf{y}$ be such a complex random vector of size F, with $\mathbf{x}, \mathbf{y} \in \mathbb{R}^F$. Then $\mathbf{x}$ and $\mathbf{y}$ follow a Gaussian distribution such that:

$$\mathbf{x}, \mathbf{y} \sim \mathcal{N}(\mathbf{0}_F, \frac{1}{2}\mathbf{\Sigma}^Z)$$

With the independence between $\mathbf{x}$ and $\mathbf{y}$, $\mathbf{x} \perp\!\!\!\perp \mathbf{y}$, we obtain the following equation:

$$\begin{aligned} p(\mathbf{x}, \mathbf{y}) &= \mathbf{N}(\mathbf{x}; \mathbf{0}_F, \frac{1}{2}\mathbf{\Sigma}^Z)\mathbf{N}(\mathbf{y}; \mathbf{0}_F, \frac{1}{2}\mathbf{\Sigma}^Z) \\ &= \left((2\pi)^{-\frac{F}{2}}|\mathbf{\Sigma}^Z|^{-\frac{1}{2}}\right)^2 \exp\left[-\frac{1}{2}\left(\mathbf{x}^T\left(\frac{\mathbf{\Sigma}^Z}{2}\right)^{-1}\mathbf{x} + \mathbf{y}^T\left(\frac{\mathbf{\Sigma}^Z}{2}\right)^{-1}\mathbf{y}\right)\right] \\ &= \pi^{-F}|\mathbf{\Sigma}^Z|^{-1}\exp\left(-\mathbf{x}^T(\mathbf{\Sigma}^Z)^{-1}\mathbf{x} - \mathbf{y}^T(\mathbf{\Sigma}^Z)^{-1}\mathbf{y}\right) \end{aligned}$$

In the proposed framework, the covariance matrix is assumed to be diagonal: $\mathbf{\Sigma}^Z = \text{diag}(\mathbf{s}^Z)$. This leads to:

$$p(\mathbf{x}, \mathbf{y}) = \prod_f \frac{1}{\pi s_f^Z}\exp\left(-\frac{x_f^2 + y_f^2}{s_f^Z}\right) \tag{A.1}$$

It is also interesting to express this likelihood in polar coordinates. $\mathbf{z}$ is then written with its modulus and argument such that $\mathbf{z} = \boldsymbol{\rho} \bullet \exp j\boldsymbol{\theta}$, with $\boldsymbol{\rho}, \boldsymbol{\theta} \in \mathbb{R}^+ \times [0, 2\pi[$. The joint

likelihood of $\boldsymbol{\rho}$ and $\boldsymbol{\theta}$ is then obtained by change of variables from the cartesian coordinates (x, y) to (ρ, θ). Using the following scalar relations:

$$p(x,y)dxdy = p(\rho,\theta)d\rho d\theta$$
$$dxdy = \rho d\rho d\theta$$

We can derive the expression of the desired likelihood:

$$p(\boldsymbol{\rho},\boldsymbol{\theta}) = \prod_f \frac{\rho_f}{\pi s_f^Z} \exp\left(-\frac{\rho_f^2}{s_f^Z}\right) \tag{A.2}$$

From Equation (A.2), two properties can be identified: first, the likelihood does not depend on the phase of $\mathbf{z}$, which means that the definition of a complex proper Gaussian random variable implies a uniformly distributed phase of the complex variable. Second, integrating Equation (A.2) for $\boldsymbol{\theta} \in [0, 2\pi[^F$ shows that the proper Gaussian assumption is equivalent to assuming that the modulus at entry f ρ_f of Z follows a Rayleigh distribution $\mathcal{R}(s_f^Z/2)$:

$$p(\boldsymbol{\rho}) = \prod_f \frac{2\rho_f}{s_f^Z} \exp\left(-\frac{\rho_f^2}{s_f^Z}\right)$$

The slight differences between the expressions of the likelihood in the cartesian domain and in the polar domain also deserves to be highlighted. For our application, within the ML or MAP estimations, there is no consequence when choosing one coordinate system or the other. One should however bear in mind that the value of the likelihood that is maximized actually depends on the coordinate system. In our case, this is not a problem since we essentially use gradient methods which aim at increasing the likelihood, and the above equivalence shows that increasing one form of the likelihood also corresponds to increasing the likelihood for the other form.

A.1.2 Complex proper Gaussian distribution properties

First we prove Property 1. Indeed, although not a direct result on Gaussian processes, this property, which holds for w.s.s. processes, with continuous time-continuous frequency Fourier transform, motivates the choice of diagonal covariance for the Gaussian vector models:

Proof We can write:

$$\begin{aligned}
E[\tilde{x}_{f+\xi}\tilde{x}_f^*] &= E\left[\int_t x_t \exp(-j2\pi(f+\xi)t)dt \int_\theta x_\theta^* \exp(j2\pi f\theta)d\theta\right] \\
&= \int_{t,\theta} E[x_t x_\theta^*] \exp\left(-j2\pi((f+\xi)t - f\theta)\right) dtd\theta \\
&= \int_{t,\theta} E[x_t x_\theta^*] \exp\left(-j2\pi(f(t-\theta) + \xi t)\right) dtd\theta \\
&= \int_{\tau,\theta} E[x_{\theta+\tau} x_\theta^*] \exp\left(-j2\pi(f\tau + \xi(\theta+\tau))\right) d\tau d\theta \\
&= \int_\tau r^X(\tau) \exp(-j2\pi(f+\xi)\tau)d\tau \int_\theta \exp(-j2\pi\xi\theta)d\theta \qquad \text{(A.3)} \\
&= s^X_{f+\xi}\delta_\xi = \delta_\xi s^X_f \qquad \blacksquare
\end{aligned}$$

Note that Proposition 1 also holds with the discrete time Fourier transform (DTFT), but that within a discrete Fourier transform (DFT), the above sum of complex exponential in Equation (A.3) still depends on τ, and the final equation does not hold anymore. The determining condition is therefore the limitation over the length of the observation, which is why we talk about the windowing effect.

Another interesting yet classical result can be observed with Gaussian variables in the source separation case, as was sketched in Section 6.2.2. Let $\mathbf{x}_n = \mathbf{v}_n + \mathbf{m}_n$ be the n^{th} frame of the mixture of two complex proper Gaussian, centered, components $\mathbf{v}_n$ and $\mathbf{m}_n$. Their diagonal covariance matrices respectively are $\text{diag}(\mathbf{s}_n^V)$ and $\text{diag}(\mathbf{s}_n^M)$. $\mathbf{v}_n$ and $\mathbf{m}_n$ are assumed independent.

Proposition 3 (Posterior mean of sum of Gaussians) *The posterior mean of* $\mathbf{v}_n$*, knowing* $\mathbf{x}_n$*, is given by the following relation:*

$$E[\mathbf{v}_n|\mathbf{x}_n] = \frac{\mathbf{s}_n^V}{\mathbf{s}_n^V + \mathbf{s}_n^M} \bullet \mathbf{x}_n \qquad \text{(A.4)}$$

Proof This result is a classic result for real Gaussian random variables. The result is also rather direct to obtain from the definition of the posterior mean. In the following equations, the super-script r is used to denote the real part of a complex vector or number. Note also that the fractions between vectors are meant element by element of these vectors. For the sake of simplicity, the covariance matrices of the multivariate Gaussians, since they

are assumed diagonal, are only denoted by their diagonal vectors.

$$\hat{\mathbf{v}}_n = E[\mathbf{v}_n|\mathbf{x}_n] = \int_{\mathbf{v}} \mathbf{v} p(\mathbf{v}|\mathbf{x}_n) d\mathbf{v} \tag{A.5}$$

$$= \int_{\mathbf{v}} \mathbf{v} \frac{p(\mathbf{v}, \mathbf{x}_n)}{p(\mathbf{x}_n)} d\mathbf{v} \tag{A.6}$$

$$= \int_{\mathbf{v},\mathbf{m}|\mathbf{v}+\mathbf{m}=\mathbf{x}_n} \mathbf{v} \frac{p(\mathbf{v}, \mathbf{m}, \mathbf{x}_n)}{p(\mathbf{x}_n)} d\mathbf{v} d\mathbf{m} \tag{A.7}$$

$$= \int_{\mathbf{v},\mathbf{m}|\mathbf{v}+\mathbf{m}=\mathbf{x}_n} \mathbf{v} \frac{p(\mathbf{v})p(\mathbf{m})}{p(\mathbf{x}_n)} d\mathbf{v} d\mathbf{m} \tag{A.8}$$

$$= \frac{1}{p(\mathbf{x}_n^r)} \int_{\mathbf{v}^r,\mathbf{m}^r|\mathbf{v}^r+\mathbf{m}^r=\mathbf{x}_n^r} \mathbf{v}^r p(\mathbf{v}^r) p(\mathbf{m}^r) d\mathbf{v}^r d\mathbf{m}^r + j \dots \tag{A.9}$$

$$= \frac{1}{p(\mathbf{x}_n^r)} \int_{\mathbf{v}^r} \mathbf{v}^r \mathbf{N}(\mathbf{v}^r; \mathbf{0}_F, \mathbf{s}_n^V/2) \mathbf{N}(\mathbf{x}_n^r - \mathbf{v}^r; \mathbf{0}_F, \mathbf{s}_n^M/2) d\mathbf{v}^r + j \dots \tag{A.10}$$

$$= \frac{1}{p(\mathbf{x}_n^r)} \int_{\mathbf{v}^r} \mathbf{v}^r \left(\prod_f \frac{1}{\pi \sqrt{s_{fn}^V s_{fn}^M}} \exp\left(- \frac{\left(v_{fn}^r - \frac{s_{fn}^V}{s_{fn}^V + s_{fn}^M} x_{fn}^r \right)^2}{\frac{s_{fn}^V s_{fn}^M}{s_{fn}^V + s_{fn}^M}} - \frac{(x_{fn}^r)^2}{s_{fn}^V + s_{fn}^M} \right) \right) d\mathbf{v}^r + j \dots \tag{A.11}$$

$$= \frac{\mathbf{N}(\mathbf{x}_n^r; \mathbf{0}_F, \mathbf{s}_n^V/2 + \mathbf{s}_n^M/2)}{p(\mathbf{x}_n^r)} \int_{\mathbf{v}^r} \mathbf{v}^r \mathbf{N}(\mathbf{v}^r; \frac{\mathbf{s}_n^V}{\mathbf{s}_n^V + \mathbf{s}_n^M} \bullet \mathbf{x}_n^r, \frac{1}{2} \frac{\mathbf{s}_n^V \bullet \mathbf{s}_n^M}{\mathbf{s}_n^V + \mathbf{s}_n^M}) d\mathbf{v}^r + j \dots \tag{A.12}$$

$$= \int_{\mathbf{v}} \mathbf{v} \mathbf{N}_c(\mathbf{v}; \frac{\mathbf{s}_n^V}{\mathbf{s}_n^V + \mathbf{s}_n^M} \bullet \mathbf{x}_n, \frac{\mathbf{s}_n^V \bullet \mathbf{s}_n^M}{\mathbf{s}_n^V + \mathbf{s}_n^M}) d\mathbf{v} \tag{A.13}$$

$$\hat{\mathbf{v}}_n = \frac{\mathbf{s}_n^V}{\mathbf{s}_n^V + \mathbf{s}_n^M} \bullet \mathbf{x}_n \tag{A.14}$$

From Equation (A.9), the derivations for the imaginary parts are hidden, since they are exactly the same as the ones for the real part. Between Equation (A.10) and Equation (A.11), many steps have been skipped. One can however easily verify that the equality holds between these equations. Equation (A.13) is obtained by recombining the real and the imaginary parts of the above equations, using the definition of the complex proper Gaussian. Note that this result also provides the posterior covariance matrix, and also the posterior power of $\mathbf{v}$:

$$\widehat{\mathbf{v}_n^2} = E[\mathbf{v}_n^2|\mathbf{x}_n] = \left(\frac{\mathbf{s}_n^V}{\mathbf{s}_n^V + \mathbf{s}_n^M} \bullet \mathbf{x}_n \right)^2 + \frac{\mathbf{s}_n^V \bullet \mathbf{s}_n^M}{\mathbf{s}_n^V + \mathbf{s}_n^M} \tag{A.15}$$

which is a result used for instance in [Ozerov et al., 2007] and [Févotte et al., 2009a]. ∎

When the signals $\mathbf{v}_n$ and $\mathbf{m}_n$ are stationary, then the diagonal of their covariance matrix equals there PSDs, and the posterior mean is equal to the Wiener estimator of $\mathbf{v}_n$ given $\mathbf{x}_n$.

A.2 Gamma distribution $\mathcal{G}$

Let $X \in \mathbb{R}$ be a random variable which is Gamma distributed $\mathcal{G}(\alpha, \beta)$, with shape parameter α and scale parameter β. Then the likelihood writes:

$$p(X|\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} X^{\alpha-1} \exp(-\beta X) = \mathbf{G}(X; \alpha, \beta)$$

where Γ is the Gamma function defined as:

$$\Gamma(y) = \int_{t=0}^{\infty} t^{y-1} \exp(-t) dt \tag{A.16}$$

Appendix B

Derivation of the algorithms

In this Chapter, we first derive the multiplicative algorithm presented in Section 5.2. The Expectation-Maximisation (EM) algorithm is then presented.

Both algorithms use the same multiplicative gradient approach, which is explained in Section B.1.1. We give the multiplicative rules in the case of Itakura-Saito divergence for the parameters of our model, with detailed calculus for some of them. The EM algorithm presents an additional difficulty since it also deals with hidden states. We will show in Section B.2 how to apply a GEM algorithm and avoid some numerical issues that one may come across when implementing it. Some insights about the behaviour of the multiplicative gradient method are given, especially as concerns their "convergence" rate.

B.1 (S)IMM multiplicative algorithm derivations

For the IMM[1], the criterion to maximize is:

$$C_{\mathrm{IMM}}(\mathbf{\Theta}^{\mathrm{IMM}}) = \sum_{f,n} \log \frac{|x_{fn}|}{\pi s_{fn}^{\mathrm{IMM}}} - \frac{|x_{fn}|^2}{s_{fn}^{\mathrm{IMM}}} \tag{B.1}$$
$$\text{with } s_{fn}^{\mathrm{IMM}} = \left[(\mathbf{W}^{\Phi}\mathbf{H}^{\Phi}) \bullet (\mathbf{W}^{F_0}\mathbf{H}^{F_0}) + \mathbf{W}^{M}\mathbf{H}^{M}\right]_{fn}$$

In a first section, the general principle leading to the desired algorithm for estimation of

$$\mathbf{\Theta}^{\mathrm{IMM}} = \{\mathbf{H}^{F_0}, \mathbf{W}^{\Phi}, \mathbf{H}^{\Phi}, \mathbf{H}^{M}, \mathbf{W}^{M}\}$$

is given. Then the actual algorithm and the necessary computations are presented.

B.1.1 Multiplicative gradient principle

As seen in Chapter 4, maximizing the criterion given in Equation (B.1) is equivalent to minimizing the Itakura-Saito (IS) divergence between the power spectrum $|\mathbf{X}|^2$ and the variance parameter $\mathbf{S}^{\mathrm{IMM}}$.

Let $\theta \in \mathbf{\Theta}^{\mathrm{IMM}}$. Minimizing the desired criterion may be done by finding the value θ^* at which the derivative with respect to θ is zero. The partial derivative of the IS divergence

[1] For the sake of simplicity, we only keep the "IMM" notation in this section. The extension to the SIMM model is rather straightforward and changes from the IMM model are clearly indicated.

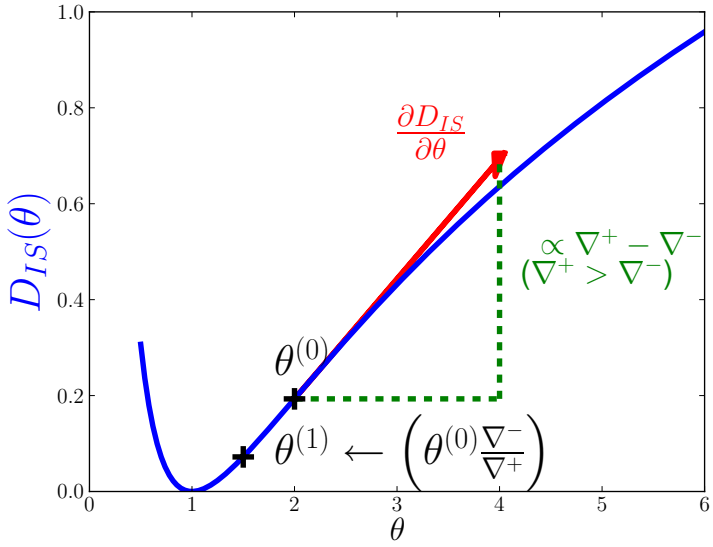


Figure B.1: Multiplicative gradient principle. The blue line represents the cost function to be minimized, the red arrow represents the gradient at $\theta^{(0)}$ and $\theta^{(1)}$ is the updated value of θ following the multiplicative gradient method. See text for details.

between $|\mathbf{X}|^2$ and $\mathbf{S}^{imm}$ is:

$$\frac{\partial D_{IS}(|\mathbf{X}|^2||\mathbf{S}^{\mathrm{IMM}}(\theta))}{\partial \theta} = \underbrace{\sum_{fn} \frac{\partial s_{fn}^{\mathrm{IMM}}(\theta)}{\partial \theta} \frac{1}{s_{fn}^{\mathrm{IMM}}(\theta)}}_{\nabla^+} - \underbrace{\sum_{fn} \frac{\partial s_{fn}^{\mathrm{IMM}}(\theta)}{\partial \theta} \frac{|x_{fn}|^2}{s_{fn}^{\mathrm{IMM}}(\theta)^2}}_{\nabla^-} \tag{B.2}$$

In Equation (B.2), ∇^+ and ∇^- are positive terms, as will be explicitly seen in Section B.1.2. Using this relation, and given the need for a factor γ which parameterizes the gradient descent $\theta_{(i)} \leftarrow \theta_{(i-1)}\gamma$, the desired update direction and orientation are found with $\gamma = \frac{\nabla^-}{\nabla^+}$, giving the general multiplicative rule:

$$\theta^{(i+1)} \leftarrow \theta^{(i)} \frac{\nabla^-}{\nabla^+} \tag{B.3}$$

Figure B.1 indeed shows the graphical interpretation of the multiplicative gradient approach. When the gradient at $\theta^{(0)}$ is positive, i.e. $\nabla^+ > \nabla^-$, as is the case on Figure B.1, then the updated value for θ should be lower than $\theta^{(0)}$ - we need to go the "opposite" way of the gradient. $\theta^{(1)} = \theta^{(0)}\frac{\nabla^-}{\nabla^+} < \theta^{(0)}$ is therefore a good candidate for this new value. The principle stays valid for a negative gradient, with $\nabla^+ < \nabla^-$.

Note that the desired value minimizing the IS divergence, θ^*, if it exists, verifies the condition: $\frac{\partial D_{IS}}{\partial \theta}(\theta^*) = 0$. When $\theta^{(i)}$ is close to θ^*, the expected behaviour is thus that $\theta^{(i+1)}$ stays close to θ^* (if not closer), and that $\theta^{(i)} = \theta^* \Rightarrow \theta^{(i+1)} = \theta^*$, which means that the solution θ^* is stable. The second condition trivially holds for the updating rule

Equation (B.3). Indeed, we have:

$$\frac{\partial D_{IS}}{\partial \theta}(\theta^{(i)}) = 0 \tag{B.4}$$

$$\Leftrightarrow \nabla^+ = \nabla^- \tag{B.5}$$

$$\Leftrightarrow \theta^{(i+1)} = \theta^{(i)} \tag{B.6}$$

However, the multiplicative updates do not guarantee the convergence *per se*, and nothing prevents them from following erratic paths. This behaviour can however be controlled by adopting the following updating rule, instead of Equation (B.3):

$$\theta^{(i+1)} \leftarrow \theta^{(i)} \left(\frac{\nabla^-}{\nabla^+}\right)^{\omega} \tag{B.7}$$

where the parameter ω allows to modify the convergence speed, with values typically between 0 and 1. Note that convergence studies have been recently held for the IS divergence by Badeau et al. [2009].

B.1.2 IMM and Itakura-Saito multiplicative rules

For the IMM, the criterion to maximize is, as in Equation (5.18):

$$C_{\text{IMM}}(\mathbf{\Theta}^{\text{IMM}}) = \sum_{f,n} \log \frac{|x_{fn}|}{\pi s_{fn}^{\text{IMM}}} - \frac{|x_{fn}|^2}{s_{fn}^{\text{IMM}}} \tag{B.8}$$

$$\text{with } \mathbf{S}^{\text{IMM}} = \left(\mathbf{W}^{\Phi}\mathbf{H}^{\Phi}\right) \bullet \left(\mathbf{W}^{F_0}\mathbf{H}^{F_0}\right) + \mathbf{W}^{M}\mathbf{H}^{M}$$

The parameter set is $\mathbf{\Theta}^{\text{IMM}} = \{\mathbf{H}^{F_0}, \mathbf{H}^{\Phi}, \mathbf{W}^{\Phi}, \mathbf{W}^{M}, \mathbf{H}^{M}\}$. We differentiate the criterion w.r.t. each of the coefficients of $\mathbf{\Theta}^{\text{IMM}}$ to obtain the desired updating rules.

Generally, we can note that, for a given parameter $\theta \in \mathbf{\Theta}^{\text{IMM}}$, we have:

$$\frac{\partial C_{\text{IMM}}(\mathbf{\Theta}^{\text{IMM}})}{\partial \theta} = \sum_{f,n} \frac{\partial s_{fn}^{\text{IMM}}}{\partial \theta} \frac{|x_{fn}|^2}{(s_{fn}^{\text{IMM}})^2} - \sum_{f,n} \frac{\partial s_{fn}^{\text{IMM}}}{\partial \theta} \frac{1}{s_{fn}^{\text{IMM}}}$$

We give below all the partial derivatives for each type of parameter in $\mathbf{\Theta}^{\text{IMM}}$.

$$\frac{\partial s_{\xi\tau}^{\text{IMM}}}{\partial h_{kn}^{\Phi}} = w_{\xi k}^{\Phi} \left[\mathbf{W}^{F_0}\mathbf{H}^{F_0}\right]_{\xi n} \delta_{\tau=n}$$

$$\frac{\partial s_{\xi\tau}^{\text{IMM}}}{\partial w_{fk}^{\Phi}} = h_{k\tau}^{\Phi} \left[\mathbf{W}^{F_0}\mathbf{H}^{F_0}\right]_{f,\tau} \delta_{\xi=f}$$

$$\frac{\partial s_{\xi\tau}^{\text{SIMM}}}{\partial h_{pk}^{\Gamma}} = w_{\xi p}^{\Gamma} h_{k\tau}^{\Gamma} \left[\mathbf{W}^{F_0}\mathbf{H}^{F_0}\right]_{\xi\tau}$$

$$\frac{\partial s_{\xi\tau}^{\text{IMM}}}{\partial h_{un}^{F_0}} = w_{\xi u}^{F_0} \left[\mathbf{W}^{\Phi}\mathbf{H}^{\Phi}\right]_{\xi n} \delta_{\tau=n}$$

$$\frac{\partial s_{\xi\tau}^{\text{IMM}}}{\partial h_{rn}^{M}} = w_{\xi r}^{M} \delta_{\tau=n}$$

$$\frac{\partial s_{\xi\tau}^{\text{IMM}}}{\partial w_{fr}^{M}} = h_{r\tau}^{M} \delta_{\xi=f}$$

In detail, for $h_{un}^{F_0}$, we obtain:

$$\begin{aligned}\frac{\partial C_{\text{IMM}}(\boldsymbol{\Theta})}{\partial h_{un}^{F_0}} &= \sum_f \frac{w_{fu}^{F_0}\left[\mathbf{W}^{\Phi}\mathbf{H}^{\Phi}\right]_{f,n}|x_{fn}|^2}{(s_{fn}^{\text{IMM}})^2} - \sum_f \frac{w_{fu}^{F_0}\left[\mathbf{W}^{\Phi}\mathbf{H}^{\Phi}\right]_{f,n}}{s_{fn}^{\text{IMM}}} \\ &= \sum_f w_{fu}^{F_0}\left[\frac{(\mathbf{W}^{\Phi}\mathbf{H}^{\Phi})\bullet|\mathbf{X}|^2}{\mathbf{S}^2}\right]_{f,n} - \sum_f w_{fu}^{F_0}\left[\frac{\mathbf{W}^{\Phi}\mathbf{H}^{\Phi}}{\mathbf{S}}\right]_{f,n} \\ &= \underbrace{\left[(\mathbf{W}^{F_0})^T\left(\frac{(\mathbf{W}^{\Phi}\mathbf{H}^{\Phi})\bullet|\mathbf{X}|^2}{\mathbf{S}^2}\right)\right]_{u,n}}_{p_{un}^{F_0}} - \underbrace{\left[(\mathbf{W}^{F_0})^T\left(\frac{\mathbf{W}^{\Phi}\mathbf{H}^{\Phi}}{\mathbf{S}}\right)\right]_{u,n}}_{q_{un}^{F_0}}\end{aligned}$$

Note that $\mathbf{P}^{F_0}$ and $\mathbf{Q}^{F_0}$ are matrices of the same size as $\mathbf{H}^{F_0}$, and the updating rule can be written as a Hadamard product of the previous $\mathbf{H}^{F_0}$ and the multiplicative gradient matrix $\mathbf{P}^{F_0}/\mathbf{Q}^{F_0}$ (element-wise division):

$$\mathbf{H}^{F_0} \leftarrow \mathbf{H}^{F_0} \bullet \frac{\mathbf{P}^{F_0}}{\mathbf{Q}^{F_0}} \tag{B.9}$$

Similar developments can be done for the other parameters, and the obtained updating rules are given in Algorithm 5.1. Note that the updating rules for the accompaniment parameters, $\mathbf{W}^M$ and $\mathbf{H}^M$, are the same as classical NMF updating rules (for instance in [Dhillon and Sra, 2005]), except for the computation of $\mathbf{S}^{\text{IMM}}$, which includes the proposed source/filter model for the leading voice.

B.2 (S)GSMM: Expectation-Maximisation algorithm derivations

We recall the ML criterion that we want to optimize:

$$\begin{aligned}C_{\text{GSMM}}(\boldsymbol{\Theta},\boldsymbol{\Theta}^{(i-1)}) = \sum_{n,k,u}\Bigg[\sum_f\left(\log\frac{|x_{fn}|}{\pi s_{fn}^{\text{GSMM}|ku}} - \frac{|x_{fn}|^2}{s_{fn}^{\text{GSMM}|ku}}\right) + \log\pi_{ku}\Bigg] \\ \times\, p(k,u|\mathbf{x}_n;\boldsymbol{\Theta}^{(i-1)}) - \lambda\left(\sum_{k,u}\pi_{ku} - 1\right) + \text{CST}\end{aligned} \tag{B.10}$$

where "CST" is a constant which is independent from the parameters in

$$\boldsymbol{\Theta}^{\text{GSMM}} = \{\mathbf{B}, \mathbf{W}^{\Phi}, \mathbf{H}^M, \mathbf{W}^M\}$$

and $s_{fn}^{\text{GSMM}|ku}$ is given by:

$$\mathbf{s}_n^{\text{GSMM}|ku} = b_{kun}\mathbf{w}_k^{\Phi} \bullet \mathbf{w}_u^{F_0} + \mathbf{W}^M\mathbf{h}_n^M \tag{B.11}$$

We give in detail all the computations needed to obtain the formulas in this manuscript as well as in Durrieu et al. [2010].

As was discussed in Section 5.3, this criterion is a classical criterion for Expectation-Maximisation (EM) algorithm. The detail of both steps, namely the E-step and the M-step, is given in the following sections.

B.2.1 E step: Computing the posterior $p(k, u|\mathbf{x}_n; (\mathbf{\Theta}^{\mathbf{GSMM}})^{(i-1)})$

Following the Bayes theorem, we have the following relation:

$$\begin{aligned} p(k,u|\mathbf{x}_n;\mathbf{\Theta}^{\mathrm{GSMM}}) &= \frac{p(\mathbf{x}_n|k,u;\mathbf{\Theta}^{\mathrm{GSMM}})p(k,u)}{p(\mathbf{x}_n)} \\ &= \frac{\left(\prod_f p(x_{fn}|k,u;\mathbf{\Theta}^{\mathrm{GSMM}})\right)\pi_{ku}}{\sum_{k,u}\left(\prod_f p(x_{fn}|k,u;\mathbf{\Theta}^{\mathrm{GSMM}})\right)\pi_{ku}} \end{aligned}$$

To obtain the *a posteriori* probabilities for each state, we need to compute the conditional likelihood of the observations for the given state, then multiply it with the *a priori* probability of the state and at last normalize the result over the states. Note that the likelihood is given by:

$$p(x_{fn}|k,u;\mathbf{\Theta}^{\mathrm{GSMM}}) = \frac{|x_{fn}|}{\pi s_{fn}^{\mathrm{GSMM}|ku}} \exp\left(-\frac{|x_{fn}|^2}{s_{fn}^{\mathrm{GSMM}|ku}}\right)$$

where $s_{fn}^{\mathrm{GSMM}|ku}$ is given by Equation (B.11). This conditional probability will be denoted $p(x_{fn}|k,u)$, when there is no ambiguity about the parameter set used to compute $s_{fn}^{\mathrm{GSMM}|ku}$. In practice, some numerical issues can arise, leading to *a posteriori* probabilities all equal to 0, for a given frame and all the states. This can happen when the model does not fully fit the observations, with very low likelihood values. To solve this problem, we compute the log-likelihoods instead, and remove a certain quantity from all of the values for all the states. This quantity is chosen such that, for a given frame, the maximum over the states for $\left(\prod_f p(x_{fn}|k,u)\right)\pi_{ku}$ is arbitrarily set to 1. It does not affect the result, since this quantity, removed from the logarithm and therefore divided to the likelihood, would anyway disappear in the normalization over the states

Let us introduce $\mathbf{L}$ the joint log-probability of the observation and the source/filter states such that: $L_{kun} = \log p(\mathbf{x}_n|k,u) + \log \pi_{ku}$. The full process needed to compute this E-step is given in Algorithm B.1.

B.2.2 M step: amplitude coefficients B

The criterion, without the parts not depending on $\mathbf{B}$, is:

$$\begin{aligned} C_{\mathrm{GSMM}}(\mathbf{B},(\mathbf{\Theta}^{\mathrm{GSMM}})^{(i-1)}) = \sum_{n,k,u}\left[\sum_f\left(\log\frac{|x_{fn}|}{\pi s_{fn}^{\mathrm{GSMM}|ku}} - \frac{|x_{fn}|^2}{s_{fn}^{\mathrm{GSMM}|ku}}\right)\right] \\ \times\,\gamma_n^{(i-1)}(k,u) \end{aligned}$$

Algorithm B.1 Computing $\gamma_n(k,u) = p(k,u|\mathbf{x}_n)$

Initialisation: $L_{kun} = 0, \forall n,k,u$
for $n \in [1,N]$ **do**
  for $(k,u) \in [1,K] \times [1,U]$ **do**
    Prior probabilities: $L_{kun} \leftarrow \log \pi_{ku}$
  end for
  for $f \in [1,F]$ **do**
    for $(k,u) \in [1,K] \times [1,U]$ **do**
      Adding contribution of frequency bin f: $L_{kun} \leftarrow L_{kun} + \log p(x_{fn}|k,u)$
    end for
    Computing the maximum: $maxL_{fn} \leftarrow \max_{ku} L_{kun}$
    for $(k,u) \in [1,K] \times [1,U]$ **do**
      Removing the maximum value: $L_{kun} \leftarrow L_{kun} - maxL_{fn}$
    end for
  end for
  Computing normalizing factor: $norm \leftarrow \sum_{ku} \exp(L_{kun})$
  for $(k,u) \in [1,K] \times [1,U]$ **do**
    Normalizing: $p(k,u|\mathbf{x}_n) = \dfrac{\exp(L_{kun})}{norm}$
  end for
end for

W.r.t. a given coefficient b_{kun}, we have:

$$\frac{\partial C_{\text{GSMM}}(\mathbf{B}, (\mathbf{\Theta}^{\text{GSMM}})^{(i-1)})}{\partial b_{kun}} = \left[-\sum_f \frac{\frac{\partial s_{fn}^{\text{GSMM}|ku}}{\partial b_{kun}}}{s_{fn}^{\text{GSMM}|ku}} + \sum_f \frac{\partial s_{fn}^{\text{GSMM}|ku}}{\partial b_{kun}} \frac{|x_{fn}|^2}{(s_{fn}^{\text{GSMM}|ku})^2} \right] \times \gamma_n^{(i-1)}(k,u)$$

$$\text{with: } \frac{\partial s_{fn}^{\text{GSMM}|ku}}{\partial b_{kun}} = w_{fk}^{\Phi} w_{fu}^{F_0}$$

We note the characteristic form of the gradient $p_{kun}^B - q_{kun}^B$ where both p_{kun}^B and q_{kun}^B are positive quantities, as explained in Section B.1.1:

$$p_{kun}^B = \gamma_n^{(i-1)}(k,u) \sum_f \frac{w_{fk}^{\Phi} w_{fu}^{F_0} |x_{fn}|^2}{(S_{fn}^{\text{GSMM}|KU})^2}$$

$$q_{kun}^B = \gamma_n^{(i-1)}(k,u) \sum_f \frac{w_{fk}^{\Phi} w_{fu}^{F_0}}{s_{fn}^{\text{GSMM}|ku}}$$

The multiplicative updating rule is then found with:

$$b_{kun} \leftarrow b_{kun} \times \frac{p_{kun}^B}{q_{kun}^B} \tag{B.12}$$

Note at last that if $\gamma_n^{(i-1)}(k,u) \neq 0$, the resulting updating rule does not depend on the posterior probabilities. if $\gamma_n^{(i-1)}(k,u) = 0$, that means that state (k,u) is probably not

active at frame n, such that the corresponding amplitude b_{kun} could be set to an arbitrary value. The term $\gamma_n^{(i-1)}(k,u)$ can therefore be discarded from Equation (B.12) and updating $\mathbf{B}$ can be done without the E-step.

B.2.3 M step: w_{fk}^{Φ}

The considered criterion is here also equal to:

$$C_{\mathrm{GSMM}}(\mathbf{W}^{\Phi},(\mathbf{\Theta}^{\mathrm{GSMM}})^{(i-1)}) = \sum_{n,k,u}\left[\sum_{f}\left(\log\frac{|x_{fn}|}{\pi s_{fn}^{\mathrm{GSMM}|ku}} - \frac{|x_{fn}|^2}{s_{fn}^{\mathrm{GSMM}|ku}}\right)\right] \times \gamma_n^{(i-1)}(k,u)$$

and the partial derivative, for a given w_{fk}^{Φ}:

$$\frac{\partial C_{\mathrm{GSMM}}(\mathbf{W}^{\Phi},(\mathbf{\Theta}^{\mathrm{GSMM}})^{(i-1)})}{\partial w_{fk}^{\Phi}} = \underbrace{-\sum_{n,u}\frac{b_{kun}w_{fu}^{F_0}}{s_{fn}^{\mathrm{GSMM}|ku}}\times\gamma_n^{(i-1)}(k,u)}_{q_{fk}^{\Phi}} + \underbrace{\sum_{n,u}\frac{b_{kun}w_{fu}^{F_0}|x_{fn}|^2}{(s_{fn}^{\mathrm{GSMM}|ku})^2}\times\gamma_n^{(i-1)}(k,u)}_{p_{fk}^{\Phi}}$$

B.2.4 M step: h_{pk}^{Γ} (SGSMM)

The considered criterion is here also equal to:

$$C_{\mathrm{GSMM}}(\mathbf{H}^{\Gamma},(\mathbf{\Theta}^{\mathrm{SGSMM}})^{(i-1)}) = \sum_{n,k,u}\left[\sum_{f}\left(\log\frac{|x_{fn}|}{\pi s_{fn}^{\mathrm{SGSMM}|ku}} - \frac{|x_{fn}|^2}{s_{fn}^{\mathrm{SGSMM}|ku}}\right)\right] \times \gamma_n^{(i-1)}(k,u)$$

and the partial derivative, for a given h_{pk}^{Γ}:

$$\frac{\partial C_{\mathrm{GSMM}}(\mathbf{H}^{\Gamma},(\mathbf{\Theta}^{\mathrm{GSMM}})^{(i-1)})}{\partial h_{pk}^{\Gamma}} = \underbrace{-\sum_{f,n,u}\frac{b_{kun}w_{fp}^{\Gamma}w_{fu}^{F_0}}{s_{fn}^{\mathrm{GSMM}|ku}}\times\gamma_n^{(i-1)}(k,u)}_{q_{pk}^{\Gamma}} + \underbrace{\sum_{f,n,u}\frac{b_{kun}w_{fp}^{\Gamma}w_{fu}^{F_0}|x_{fn}|^2}{(s_{fn}^{\mathrm{GSMM}|ku})^2}\times\gamma_n^{(i-1)}(k,u)}_{p_{pk}^{\Gamma}}$$

B.2.5 M step: h_{rn}^M

The partial derivative, for a given h_{rn}^M, is:

$$\frac{\partial C_{\text{GSMM}}(\mathbf{H}^M, (\mathbf{\Theta}^{\text{GSMM}})^{(i-1)})}{\partial h_{rn}^M} = \underbrace{-\sum_{f,k,u} \frac{w_{fr}^M}{s_{fn}^{\text{GSMM}|ku}} \times \gamma_n^{(i-1)}(k,u)}_{q_{rn}^H} + \underbrace{\sum_{f,k,u} \frac{w_{fr}^M |x_{fn}|^2}{s_{fn}^{\text{GSMM}|ku}} \times \gamma_n^{(i-1)}(k,u)}_{p_{rn}^H}$$

B.2.6 M step: w_{fr}^M

The partial derivative, for a given w_{fr}^M, is:

$$\frac{\partial C_{\text{GSMM}}(\mathbf{W}^M, (\mathbf{\Theta}^{\text{GSMM}})^{(i-1)})}{\partial w_{fr}^M} = \underbrace{-\sum_{n,k,u} \frac{h_{rn}^M}{s_{fn}^{\text{GSMM}|ku}} \times \gamma_n^{(i-1)}(k,u)}_{q_{fr}^W} + \underbrace{\sum_{n,k,f_0} \frac{h_{rn}^M |x_{fn}|^2}{s_{fn}^{\text{GSMM}|ku}} \times \gamma_n^{(i-1)}(k,u)}_{p_{fr}^W}$$

B.2.7 M step: Derivations for the *a priori* probabilities π

The criterion, reduced to the parts that depend on π, is:

$$C_{\text{GSMM}}(\pi, (\mathbf{\Theta}^{\text{GSMM}})^{(i-1)}) = \sum_{n,k,u} \left[\log \pi_{ku}\right] \gamma_n^{(i-1)}(k,u) - \lambda \left(\sum_{k,u} \pi_{ku} - 1 \right)$$

By differentiating the above equation, w.r.t. π_{ku}, $\forall (k,u) \in [1,K] \times [1,U]$ and also w.r.t. λ, we obtain the following equations:

$$\begin{cases} \frac{\partial C_{\text{GSMM}}(\pi, (\mathbf{\Theta}^{\text{GSMM}})^{(i-1)})}{\partial \pi_{k_a,u_a}} & = \frac{\sum_n \gamma_n^{(i-1)}(k_a,u_a)}{\pi_{k_a,u_a}} - \lambda, \, \forall (k_a,u_a) \in [1,K] \times [1,U] \\ \frac{\partial C_{\text{GSMM}}(\pi, (\mathbf{\Theta}^{\text{GSMM}})^{(i-1)})}{\partial \lambda} & = \sum_{k,u} \pi_{ku} - 1 \end{cases}$$

By equating these partial derivatives with 0, we obtain, with the second equation, the normalization condition for the *a priori* probabilities. The first equation successively is

equivalent to:

$$\frac{\sum_n \gamma_n^{(i-1)}(k_a,u_a)}{\pi_{k_a u_a}} - \lambda = 0$$

$$\sum_n \gamma_n^{(i-1)}(k_a,u_a) = \lambda \times \pi_{k_a u_a}, \forall (k_a,u_a) \in [1,K]\times[1,U]$$

$$\sum_{k,u}\sum_n \gamma_n^{(i-1)}(k,u) = \lambda \times \sum_{k,u} \pi_{ku}$$

$$\sum_n \sum_{k,u} \gamma_n^{(i-1)}(k,u) = \lambda \times 1, \text{ thanks to the normalization condition}$$

$$\sum_n 1 = \lambda, \text{ by property of conditional probabilities}$$

and therefore: $\lambda = N$

At last, we have:

$$\pi_{ku} = \frac{1}{N}\sum_n \gamma_n^{(i-1)}(k,u) \tag{B.13}$$

B.2.8 Temporal constraint with HMM during the estimation: adaptation of E-step

When the temporal constraints are included during the estimation process, the EM algorithm needs to be adapted. Since the transition probabilities $p(Z_n^{F_0}|Z_{n-1}^{F_0})$ are set in advance, they are not to be estimated: only the E-step needs to be modified.

The HMM criterion verifies:

$$C_{\text{HMM}}(\boldsymbol{\Theta},\boldsymbol{\Theta}^{(i-1)}) = \sum_{n,k,u}\left[\sum_f \left(\log \frac{|x_{fn}|}{\pi s_{fn}^{\text{GSMM}|ku}} - \frac{|x_{fn}|^2}{s_{fn}^{\text{GSMM}|ku}}\right) + \log \pi_{ku}\right] \times p(Z_n^{\Phi}=k, Z_n^{F_0}=u|\mathbf{X};\boldsymbol{\Theta}^{(i-1)}) - \lambda\left(\sum_{k,u}\pi_{ku} - 1\right) + \text{CST} \tag{B.14}$$

As for the criterion Equation (B.10), the term "CST" does not depend on the parameters of interest. Note also that the criterion for the HMM is the same as for the GSMM, except for the posterior probability, since in the HMM framework, the state at a given frame n depends on the whole sequence $\mathbf{X}$ and not only on the observation $\mathbf{x}_n$ at that particular frame.

The E-step then requires the computation of $p(Z_n^{\Phi}=k, Z_n^{F_0}=u|\mathbf{X};\boldsymbol{\Theta}^{(i-1)})$. This can be done thanks to the forward-backward procedure [Rabiner, 1989]. However, due to the dimension of the "feature space" of the proposed model, that is to say the size of the Fourier transform, the numerical issues are even more difficult to solve than for the GSMM.

The complete forward-backward procedure is then given by Algorithm B.2. Note that the last equation in Algorithm B.2 may also require some attention, especially when dividing by the sum: one should indeed avoid to divide by 0.

Algorithm B.2 Computing $p(Z_n^\Phi = k, Z_n^{F_0} = u|\mathbf{X})$ for the HMM

Compute the conditional probabilities: $p(\mathbf{x}_n|k,u)$, $\forall n,k,u$
Initialisation of forward/backward variables:
for $(k,u) \in [1,K] \times [1,U]$ **do**
$L^\alpha_{ku1} = \log \pi_{ku} + \log p(\mathbf{x}_1|k,u)$
$L^\beta_{kuN} = 0$
end for
Scaling α: $L^\alpha_{::1} \leftarrow L^\alpha_{::1} - \max_{ku} L^\alpha_{::1}$,
$\alpha_{::1} = \exp L^\alpha_{::1}$
$\beta_{::1} = \exp L^\beta_{::1}$
Forward variables:
for n from 2 to N **do**

$$L^\alpha_{kun} \leftarrow \log p(\mathbf{x}_n|ku) - \max_{k'u'} \log p(\mathbf{x}_n|k'u') + \log\left(\sum_{lv} \alpha_{lv(n-1)} p(ku|lv)\right) - \max_{k'u'} \log\left(\sum_{lv} \alpha_{lv(n-1)} p(k'u'|lv)\right)$$

$$L^\alpha_{kun} \leftarrow L^\alpha_{kun} - \max_{k'u'} L^\alpha_{k'u'n}$$

$\alpha_{kun} = \exp L^\alpha_{kun}$
end for
Backward variables:
for n from $N-1$ to 1 **do**
$K^\beta_{ku,lv} = L^\beta_{lv(n+1)} + \log p(ku|lv) + p(\mathbf{x}_{n+1}|lv)$
$\beta_{kun} = \sum_{lv} \exp(K^\beta_{ku,lv} - \max K^\beta_{ku,lv})$
$\beta_{kun} \leftarrow \frac{\beta_{kun}}{\max_{k'u'} \beta_{k'u'n}}$
$L^\beta_{kun} = \log \beta_{kun}$
end for
Compute the posterior probabilities:
$p(Z_n^\Phi = k, Z_n^{F_0} = u|\mathbf{X}) = \dfrac{\alpha_{kun}\beta_{kun}}{\sum_{k'u'} \alpha_{k'u'n}\beta_{k'u'n}}$

B.3 Multiplicative algorithm behaviour

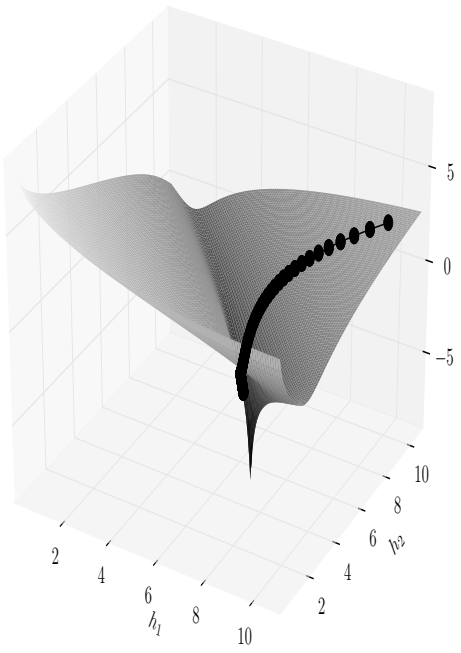
To better picture how the multiplicative gradient algorithm works, and how the power ω can affect the resulting estimation, some informal experiments with a toy example of NMF with IS divergence are presented in this section.

Let $\mathbf{W}$ and $\mathbf{H}$ be 2 matrices of respective sizes $F \times R$ and $R \times N$. For the experiment, in order to be able to visualise the IS surface and the evolution of the estimates in the parameter space, it is convenient to set $R = 2$ and $N = 1$, and to assume that we know the matrix $\mathbf{W}$. The vector $\mathbf{h}$ is then estimated using the "usual" multiplicative updating rules, as given in [Badeau et al., 2009], with different values of ω. Additionally, F is set to 100. The column vectors in $\mathbf{W}$ are the same, equal to 1, except for one value of the first vector. The vector $\mathbf{h}$ is equal to $[7, 5]^T$. All the algorithms are initialized with the same value ($[10, 10]^T$).

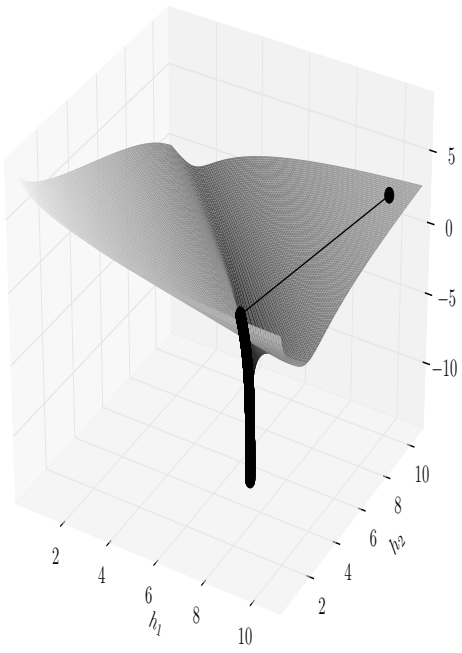
Figures B.3(a), B.3(b) and B.3(c) show the evolution of the parameters and the logarithm of the corresponding IS divergence surface, for ω values strictly under 2, which seems to be an upper bound for the convergence of the algorithm [Badeau et al., 2009]. With increasing values of ω, we observe increased updating steps for the parameters. The result for smooth target functions (as the pictured IS divergence) is the same, whatever the value for ω. Note that Figure B.3(d) corresponds to $\omega = 2$, and shows that this value may indeed be the upper bound as suggested by Badeau et al. [2009].

With a given number of iterations, these strategy however achieve quite different convergence rates: indeed, with $\omega = 0.1$, in our experiment, the estimated $\widehat{\mathbf{h}}$ was admittedly quite "far" from the true vector $\mathbf{h}$, while both the algorithms with $\omega = 1$ and 1.9, the estimated $\widehat{\mathbf{h}}$ was about equal to the desired vector. The choice of ω may however ultimately also depend on the cost function and the initialization: if one assumes that the cost function is smooth enough, with few or no local minima, then any ω value would do. In the case where this cost function has many local minima, and that we are interested in finding a global minimum, it might be a good option to use a rather high value of ω. If we are more interested in finding a solution close to the initialization, then a smaller value of ω seems more appropriate, albeit the very slow convergence.

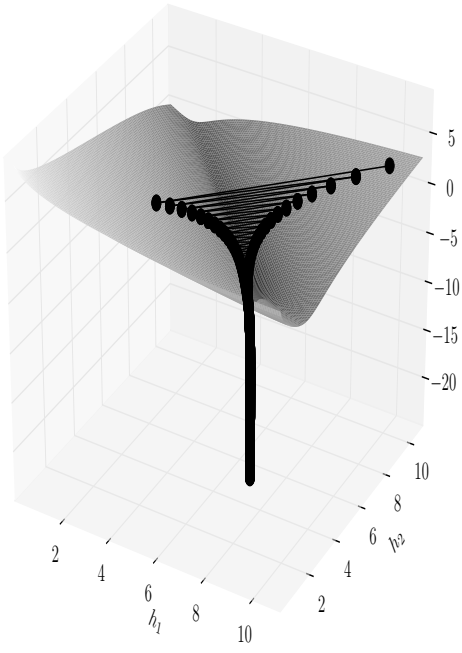
Figure B.2: IS-NMF experiments: IS divergence (in dB), w.r.t. the amplitude coefficients h_1 and h_2. The evolution of these parameters over the iterations of the multiplicative gradient algorithm are represented, for different values of ω.



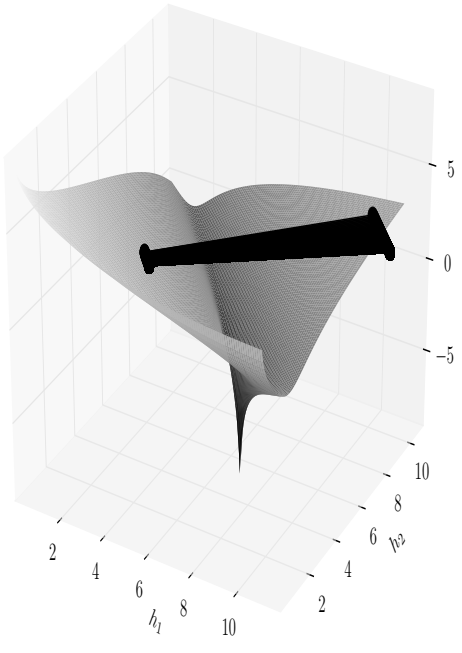
(a) $\omega = 0.1$



(b) $\omega = 1$



(c) $\omega = 1.9$



(d) $\omega = 2$

Appendix C

KLGLOTT88 : a glottal source model

We initialize each column $\mathbf{w}_u^{F_0}$ of the matrix $\mathbf{W}^{F_0}$ such that it corresponds to a specific fundamental frequency $\mathcal{F}(u)$ (in Hz). In our study, we consider the frequency range $[f_0^{\min}, f_0^{\max}]$ Hz. We discretize this frequency axis such that there are U_{st} elements of the dictionary per semitone:

$$\mathcal{F}(u) = f_0^{\min} * 2^{\frac{u-1}{12U_{st}}}$$

We thus obtain U available fundamental frequencies.

The source spectra are generated following a glottal source model: KLGLOTT88 Klatt and Klatt [1990]. We first generate the corresponding derivative of the glottal flow waveform $e_u(t)$, and then perform its Fourier transform $E_u(f)$ with the same parameters as the STFT of the observation signal: namely with the same window length, same Fourier transform size and same analysis window.

The original formula Klatt and Klatt [1990] is a continuous time function. To avoid aliasing when sampling such a formula, we use the complex amplitude for all the harmonics of the signal up to the Nyquist frequency. Let c_h be the amplitude of the h^{th} harmonic, $h \in [1, h_{\max}]$, we have Henrich [2001]:

$$c_h = \mathcal{F}(u)\frac{27}{4}\left(\exp(-i2\pi hO_q) + 2\frac{1+2\exp(-i2\pi hO_q)}{i2\pi hO_q} - 6\frac{1-\exp(-i2\pi hO_q)}{(i2\pi hO_q)^2}\right)$$

where O_q is the "open quotient" parameter , which we fixed at $O_q = 0.5$[1]. $e_u(t)$ is then the sum of the harmonics with the above amplitudes:

$$e_u(t) = \sum_h c_h \exp(i2\pi h\mathcal{F}(u)tT_s)$$

where T_s is the sampling period and $t \in \mathbb{N}^+$. We then compute $E_u(f)$. The variance $w_{fu}^{F_0}$ is then set to the squared magnitude of this Fourier transform: $w_{fu}^{F_0} = |E_u(f)|^2, \forall f \in [1, F]$.

[1] The spectral envelope of the harmonic comb depends on O_q. This parameter may therefore actually be important to correctly initialise the dictionary, and further studies may be needed to estimate the best value for our application.

Appendix D

Databases

D.1 MIREX AME databases

The databases used for MIREX are described online at the following address `http://www.music-ir.org/mirex/2009/index.php/Audio_Melody_Extraction`. It reads, firsts for the test datasets:

- **MIREX09 database:** dataset for Mirex, 374 Karaoke recordings of Chinese "Karaoke" songs. Instruments: singing voice (male, female), synthetic accompaniment, mixed at different SNR conditions, *i.e.* -5dB, 0dB and +5dB. The length of the excerpts is in average 10 seconds.
- **MIREX08 database:** 4 excerpts of 1 minute each from "north Indian classical vocal performances", instruments: singing voice (male, female), *tanpura* (Indian instrument, perpetual background drone), *harmonium* (secondary melodic instrument) and *tablas* (pitched percussions).
- **MIREX05 database:** 25 excerpts of 10-40 seconds from the following genres: Rock, R&B, Pop, Jazz, and Solo classical piano.
- **ISMIR04 database (or Audio Description Contest - ADC - 2004):** 20 excerpts of about 20 seconds each.

Second, the development sets are given by several sources, as described below:

- **MIR-1K database:** collection of 1000 excerpts of Chinese Karaoke songs, generated the same way as the test evaluation set.
- **MIREX05 database:** A development set was proposed and distributed by G. Poliner at the following address: `http://www.ee.columbia.edu/~graham/mirex_melody/`, additionally described at `http://labrosa.ee.columbia.edu/projects/melody/`.
- **ISMIR04 database:** The ADC 2004 collection was made publicly available after the 2004 contest, and used since then as development set for all the MIREX evaluation campaigns.

The specifications on the digital audio signals are:

- CD-quality (PCM, 16-bit, 44100 Hz)

- single channel (mono)

The ground-truth for the main melody was manually annotated. All the ground-truth annotations have been either generated (from MIREX 2005) or down-sampled (for ADC 2004) to a 10 ms time grid.

D.2 Quaero Main Melody Database

In order to evaluate multiple F0 estimation and note tracking within the project Quaero, several options were taken: first some songs from the RWC-Pop database [Goto et al., 2002], for which the maintainers of the database provided aligned MIDI files were cross-validated. Some offsets were to be corrected, as well as some octave "errors", especially on the melody line track. Second, the multiple track songs that constitute the source separation database are meant, in the long term, to be annotated in terms of notes and F0 lines, when possible. Only the first dataset could be developed so far. The second set is under development. Furthermore, the songs from the MTG MASS [Vinyes, 2008] have also been annotated with respect to the main melody line as well as lead instrument notes.

The database was used for the first internal evaluation campaign of the Quaero project, for the Work Package (WP) 6.2, devoted to audio content analysis. Half of the database was used as a training set, the other half being hidden from the evaluation participants. The list of the different songs that were selected is given in Table D.1. The octave corrections that we have found are also reported in that table. The offsets between the aligned MIDI files and the audio are machine dependent, and it would not make sense to report them here. The convention adopted for the octave correction is [number | correction], where "correction" is a number giving the number of octave one should move the MIDI track "number" to obtain the same octave as in the audio signal. For instance, [5:-1] means that, in track 5, an A4 should be converted to A3 to fit to the audio signal. The song numbering follows the naming convention of RWC.

D.3 Leading instrument / accompaniment separation mono database

The mono audio signal database for the experiments of article [Durrieu et al., 2009a], which was used for the experiments of Sections 6.2.5.2, 6.2.5.3, 6.2.5.4 and 6.2.5.5, is composed of 3 subsets: (A) the SiSEC 2008 development set for the "professionally produced music recordings" separation task[1], (B) some songs from Ozerov and Lagrange's private database (Ozerov et al. [2007] and Lagrange et al. [2008]) and (C) publicly available songs by S. Hurley, under Creative Commons licence. C is further divided into a pitch contour annotated set C1 and its complementary set C2.

- SiSEC professionally produced material dataset (subset A): **bearlin-roads_85-99_with_effects** (14", piano, bass, drums, male singer), **tamy-que_pena_tanto_faz_6-19** (13", female singer and guitar).
- From A. Ozerov and M. Lagrange database (subset B): **Joyce** (37", synthetic accompaniment, male singer voice), **Katzen_jammer__Clipsinvegas** (4 excerpts of 1' each, rock, strong effects on the male singer voice), **Katzen_jammer__Darkeyed**

[1] Details and software available online at: http://sisec.wiki.irisa.fr/

Table D.1: Quaero Multiple F0 and note tracking database, RWC part.

Training set		Development set	
Song number	Octave correction	Song number	Octave correction
4	[5:-1]	9	[5:-1]
10	[5:-1]	11	[5:-1]
17	[5:-1]	18	[5:-1]
20	[5:-1]	22	[5:-1]
25	[5:-1] [16:1]	27	[5:-1]
44	[5:-1]	47	[5:-1] [10:-1]
49	[5:-1]	51	[5:-1]
52	[5:-1]	53	
54		55	[5:-1]
56	[5:-1]	58	[4:-1]
60	[5:-1]	63	[5:-1]
65	[5:-1]	67	[5:-1]
69		70	[5:-1]
72	[4:-1]	75	
76	[3:-1]	78	[5:-1]
79	[3:-1]	80	[3:-1]
81		83	[5:-1]
84	[5:-1]	86	[4:-1][5:-1][7:-1]
90	[5:-1]	97	[5:-1]
98	[5:-1] [6:-2]	100	[5:-1]

Table D.2: Quaero Source Separation Database: abbrevations for Table D.3

Abbreviation	Description
Bal	volume balance between the tracks
Pan	panoramic effect ("spatialization")
Eq	equalization
Comp	dynamic compression (tracks treated individually)
Fx	various effects
Comp+	dynamic compression on the "master" (all tracks already mixed down)

(3 ex., 2 x 1' + 18", female singer), **Sting__Every_Breath_You_Take** (4 ex. : 3 x 1' + 17", karaoke with male singer), **bentOutOfShape** (3 ex. : 2 x 1' + 40", rock, male singer), **chevalierBran** (4 ex. : 4 x 1', Celtic rock, male singer + strong presence of violin and "biniou"), **intoTheUnknown** (3 ex. : 2 x 1' + 39", rock, male singer), **lePub** (4 ex. : 4 x 1', Celtic rock, same as chevalierBran), **schizosonic** (3 ex., rock, male singer).

- Shannon Hurley's songs (creative common licence) (subset C) (with melody-annotated set C1): **Silence** (C1) (4 excerpts : 4 x 1'), **Sunrise** (C1) (4 ex. : 3 x 1' + 16"), **We Are in Love** (C1) (4 ex. : 3 x 1' + 42"), **Matter of Time** (5 ex. : 4 x 1' + 37"), **Shame** (5 ex. : 4 x 1' + 40").

These files are also described on the companion webpage of [Durrieu et al., 2009a], `http://perso.telecom-paristech.fr/grichard/icassp09/`.

D.4 Quaero Source Separation Database

In details, the sound engineer of Telecom ParisTech provided the different mixing conditions for the differents songs given in Table D.3. The abbreviations for the different effects that were applied are described in Table D.2.

Table D.3: Quaero Source Separation Database.

Song name	Bal Pan	Bal Pan Eq	Bal Pan Eq Comp	Bal Pan Eq Comp Fx	Bal Pan Eq Fx	Bal Pan Eq Fx Comp+	Bal Pan Eq Comp Fx Comp+
Another Dreamer - One We Love	X	X	X	X			X
Carl Leth - the world is under attack	X	X			X		X
Alexq - carol of the bells	X	X	X	X			X
Emily Hurst - parting friends	X	X	X	X			X
Fort Minor - remember the name	X	X	X	X			X
Glen Philips - the spirit of shackleton	X	X	X	X			X
Jims Big Ego - mix tape	X	X	X	X			X
MIREX - var5	X	X			X	X	
mokamed - styltriady	X	X			X	X	
Nine Inch Nails - Good Soldier	X	X			X	X	
Shannon Hurley - Sunrise	X	X	X	X			X
Ultimate nz tour	X	X	X (Comp+)				
Vieux Farka Touré - Ana	X	X	X	X			X

Bibliography

S. A. Abdallah and M. D. Plumbley. Polyphonic music transcription by non-negative sparse coding of power spectra. In *Proceedings of the International Conference on Music Information Retrieval*, pages 318–325, Barcelona, Spain, October 10-14 2004.

T. Abe and M. Honda. Sinusoidal model based on instantaneous frequency attractors. *IEEE Transactions on Audio, Speech and Language Processing*, 14(4):1292 – 1300, July 2006.

M. Alonso, G. Richard, and B. David. Accurate tempo estimation based on harmonic+noise decomposition. *EURASIP Journal on Applied Signal Processing*, 2007, 2007.

S. Arberet, R. Gribonval, and F. Bimbot. A robust method to count and locate audio sources in a multichannel underdetermined mixture. *IEEE Transactions on Signal Processing*, 58(1):121 –133, jan. 2010.

R. Badeau, N. Bertin, and E. Vincent. On the stability of multiplicative update algorithms. application to non-negative matrix factorization. Technical report, Institut TELECOM; TELECOM ParisTech; CNRS LTCI, November 2009.

J. P. Bello and J. Pickens. A robust mid-level representation for harmonic content in music signals. In *Proceedings of the International Conference on Music Information Retrieval*, pages 204 – 311, London, U.K., 11 - 15 September 2005.

L. Benaroya. *Séparation de plusieurs sources sonores avec un seul microphone*. PhD thesis, Universite de Rennes 1, 2003.

L. Benaroya, L. Donagh, F. Bimbot, and R. Gribonval. Non negative sparse representation for Wiener based source separation with a single sensor. *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 6:613–16, 2003.

L. Benaroya, F. Bimbot, and R. Gribonval. Audio source separation with a single sensor. *IEEE Transactions on Audio, Speech and Language Processing*, 14(1):191–199, January 2006.

N. Bertin, R. Badeau, and E. Vincent. Enforcing harmonicity and smoothness in Bayesian non-negative matrix factorization applied to polyphonic music transcription. *IEEE Transactions on Audio, Speech and Language Processing*, 18(3):538 – 549, March 2010.

J. Brown. Calculation of a constant Q spectral transform. *The Journal of the Acoustical Society of America*, 89(1):425–434, 1991.

P. Cancela. Tracking melody in polyphonic audio. mirex 2008. *Music Information Retrieval Evaluation eXchange*, 2008.

C. Cao and M. Li. Multiple f0 estimation in polyphonic music (mirex 2008). *extended abstract for* the Music Information Retrieval Evaluation eXchange, 2008.

J.-F. Cardoso, M. Martin, J. Delabrouille, M. Betoule, and G. Patnachon. Component separation with flexible models. application to the separation of astrophysical emissions. *IEEE Journal of Selected Topics in Signal Processing*, October 2008.

A. T. Cemgil. *Bayesian Music Transcription*. PhD thesis, Radboud University of Nijmegen, 2004.

A. T. Cemgil and H. J. Kappen. Monte Carlo methods for Tempo Tracking and Rhythm Quantization. *Journal of Artificial Intelligence Research*, 18:45–81, 2003.

A. T. Cemgil, P. Desain, and H. J. Kappen. Rhythm Quantization for Transcription. In *Proceedings of the AISB'99 Symposium on Musical Creativity*, pages 140–146, Edinburgh, UK, April 1999. AISB.

Z. Chen, A. Cichocki, and T. M. Rutkowski. Constrained non-negative matrix factorisation method for eeg analysis in early detection of alzheimer's disease. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 893–896, Toulouse, France, May 2006.

M. G. Christensen and A. Jakobsson. *Multi-Pitch Estimation*. Morgan & Claypool Publishers, 2009.

A. Cichocki. Generalized component analysis and blind source separation methods for analyzing multichannel brain signals. 2004.

A. Cichocki. Generalized independent component analysis and its applications in processing of multisensory biomedical data. In *Proceedings of IVth International Workshop Computational Problems of Electrical Engineering*, pages 13–24, Warsaw, Poland, 2002. Institute of the Theory of Electrical Engineering and Electrical Measurements, Warsaw University of Technology, Warsaw University of Technology.

P. Comon et al. Independent component analysis, a new concept. *Signal Processing*, 36 (3):287–314, 1994.

A. Daniel, V. Emiya, and B. David. Perceptually-based evaluation of the errors usually made when automatically transcribing music. In *ISMIR*, 2008.

M. Davy and S. J. Godsill. Bayesian harmonic models for musical signal analysis (with discussion). In J. Bernardo, J. Berger, A. Dawid, and A. Smith, editors, *Bayesian Statistics VII*. Oxford University Press, 2003.

M. Davy, S. Godsill, and J. Idier. Bayesian analysis of polyphonic western tonal music. *Journal of the Acoustic Society of America*, 119(4):2498–2517, April 2006.

A. de Cheveigné. Separation of concurrent harmonic sounds: Fundamental frequency estimation and a time-domain cancellation model of auditory processing. *Journal of the Acoustical Society of America*, 93:3271–3290, 1993.

A. de Cheveigné and H. Kawahara. Yin, a fundamental frequency estimator for speech and music. *J Acoust Soc Am*, 111(4):1917–1930, Apr 2002.

A. Dempster, N. Laird, and D. Rubin. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977.

I. Dhillon and S. Sra. Generalized nonnegative matrix approximations with Bregman divergences. *Proceeding of the Neural Information Processing Systems (NIPS) Conference*, 2005.

K. Dressler. Extraction of the Melody Pitch Contour from Polyphonic Audio. *extended abstract for the Music Information Retrieval Evaluation eXchange*, 2005.

K. Dressler. Audio melody extraction for MIREX 2009. *extended abstract for the Music Information Retrieval Evaluation eXchange*, 2009.

Z. Y. Duan, J. Y. Han, and B. Pardo. Harmonically informed multi-pitch tracking. In *Proceedings of the International Society on Music Information Retrieval conference*, pages 333 – 338, Kobe, Japan, October 26 - 30 2009.

J.-L. Durrieu, G. Richard, and B. David. Singer melody extraction in polyphonic signals using source separation methods. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 169–172, Las Vegas, Nevada, USA, March 31-April 4 2008a.

J.-L. Durrieu, G. Richard, and B. David. Single sensor singer/music separation using a source/filter model of the singer voice. *ACOUSTICS*, 2008b.

J.-L. Durrieu, G. Richard, and B. David. Main melody extraction from polyphonic music excerpts using a source/filter model of the main source. *extended abstract for* the Music Information Retrieval Evaluation eXchange, August 2008c.

J.-L. Durrieu, G. Richard, and B. David. An iterative approach to monaural musical mixture de-soloing. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 105–108, Taipei, Taiwan, April 19-24 2009a.

J.-L. Durrieu, A. Ozerov, C. Févotte, G. Richard, and B. David. Main instrument separation from stereophonic audio signals using a source/filter model. In *European Signal Processing Conference (EUSIPCO)*, Glasgow, Scotland, August 24-28 2009b.

J.-L. Durrieu, G. Richard, and B. David. A source/filter approach to audio melody extraction. *extended abstract for* the Music Information Retrieval Evaluation eXchange, September 2009c.

J.-L. Durrieu, G. Richard, B. David, and C. Févotte. Source/filter model for unsupervised main melody extraction from polyphonic audio signals. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(3):564 –575, March 2010.

D. Ellis. Beat Tracking by Dynamic Programming. *Journal of New Music Research*, 36 (1):51–60, 2007.

D. Ellis and G. Poliner. Classification-based melody transcription. *Machine Learning*, 65 (2-3):439–456, December 2006.

D. Ellis and G. Poliner. Identifying 'cover songs' with chroma features and dynamic programming beat tracking. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, volume 4, pages 1429 – 1432, Hawaii, April 2007.

V. Emiya, R. Badeau, and B. David. Multipitch estimation of piano sounds using a new probabilistic spectral smoothness principle. *IEEE Transactions on Audio, Speech and Language Processing*, PP(99), december 2009.

Y. Ephraim and D. Malah. Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 32(6):1109–1121, 1984.

S. Essid, G. Richard, and B. David. Instrument recognition in polyphonic music based on automatic taxonomies. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(1):68–80, January 2006a.

S. Essid, G. Richard, and B. David. Musical instrument recognition by pairwise classification strategies. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(4): 1401–1412, July 2006b.

G. Fant. *Acoustic Theory of Speech Production*. Mouton, 1970.

C. Févotte, N. Bertin, and J.-L. Durrieu. Nonnegative matrix factorization with the Itakura-Saito divergence: With application to music analysis. *Neural Computation*, 21(3), March 2009a.

C. Févotte, N. Bertin, and J.-L. Durrieu. Nonnegative matrix factorization with the Itakura-Saito divergence: With application to music analysis. *Neural Computation*, 21(3), March 2009b.

D. Fitzgerald, M. Cranitch, and M. Cychowski. Towards an inverse constant Q transform. In *120th Audio Engineering Society Convention*, Paris, France, May 20-23 2006.

D. FitzGerald, M. Cranitch, and E. Coyle. Extended nonnegative tensor factorisation models for musical sound source separation. *Computational Intelligence and Neuroscience, Hindawi Publishing Corporation*, 2008.

R. Foucard, J.-L. Durrieu, M. Lagrange, and G. Richard. Multimodal similarity between musical streams for cover version detection. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Dallas, Texas, USA, March 14 - 19 2010.

H. Fujihara, T. Kitahara, M. Goto, K. Komatani, T. Ogata, and H. Okuno. F0 estimation method for singing voice in polyphonic audio signal based on statistical vocal model and viterbi search. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 5, pages V–V, 14-19 May 2006.

H. Fujihara, M. Goto, and H. G. Okuno. An F0 estimation method of vocal part in polyphonic music by using statistical modelling of singing voice and Viterbi search. *Information Processing Society of Japan Journal*, 49(10):3682 – 3693, October 2008. (in Japanese).

O. Gillet and G. Richard. Transcription and Separation of Drum Signals From Polyphonic Music. *Audio, Speech, and Language Processing, IEEE Transactions on [see also Speech and Audio Processing, IEEE Transactions on]*, 16(3):529–540, 2008.

E. Gómez. *Melodic Description of Audio Signals for Music Content Processing*. PhD thesis, Doctoral Pre-Thesis Work. UPF, 2002.

E. Gómez, S. Streich, B. Ong, R. P. Paiva, S. Tappert, J. M. Batke, G. Poliner, D. Ellis, and J. P. Bello. A quantitative comparison of different approaches for melody extraction from polyphonic audio recordings. Technical report, Universitat Pompeu Fabra, Barcelona, Spain, 2006.

M. Goto. Robust predominant-F0 estimation method for real-time detection of melody and bass lines in CD recordings. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 2, pages 757–760, Istanbul, Turkey, June 2000.

M. Goto. A real-time music-scene-description system: predominant-f0 estimation for detecting melody and bass lines in real-world audio signals. *ISCA Speech Communication*, 43(5):311 – 329, September 2004.

M. Goto. PreFEst: A Predominant-F0 Estimation method for polyphonic musical audio signals. In *Proceedings of the 2nd Music Information Retrieval Evaluation eXchange*, September 2005.

M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka. RWC music database: Popular, classical, and jazz music databases. In *Proceedings of the International Conference on Music Information Retrieval*, pages 287–288, 2002.

D. W. Griffin and J. S. Lim. Signal estimation from modified short-time Fourier transform. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP-32:236 – 242, April 1984.

Y. S. Han and C. Raphael. Desoloing monaural audio using mixture models. In *Proceedings of the International Conference on Music Information Retrieval*, Vienna, Austria, September 23 - 27 2007.

T. Heittola, A. Klapuri, and T. Virtanen. Musical instrument recognition in polyphonic audio using source-filter model for sound separation. In *Proceedings of the International Society for Music Information Retrieval Conference*, pages 327 – 332, Kobe, Japan, October 26 - 30 2009.

N. Henrich. *Etude de la source glottique en voix parlée et chantée*. PhD thesis, Universite de Paris 6, 2001.

D. J. Hermes. Measurement of pitch by subharmonic summation. *The Journal of the Acoustical Society of America*, 83(1):257–264, 1988.

C. Hsu, L. Chen, J. Jang, and H. Li. Singing pitch extraction from monaural polyphonic songs by contextual audio modeling and singing harmonic enhancement. In *Proceedings of the International Society for Music Information Retrieval conference*, Kobe, Japan, 26-30 October 2009.

C. Joder, S. Essid, and G. Richard. Temporal integration for audio classification with application to musical instrument classification. *IEEE Transactions on Audio, Speech and Language Processing*, 17(1):174–186, January 2009.

A. Jourjine, S. Rickard, and O. Yilmaz. Blind separation of disjoint orthogonal signals: demixing N sources from 2 mixtures. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 5, pages 2985–2988, Istanbul, Turkey, June 2000.

C. Jutten and J. Herault. Blind separation of sources, Part 1: an adaptive algorithm based on neuromimetic architecture. *Signal Processing*, 24(1):1–10, 1991.

A. Klapuri. Multipitch estimation and sound separation by the spectral smoothness principle. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 5, pages 3381–3384, 7-11 May 2001.

A. Klapuri. Multipitch analysis of polyphonic music and speech signals using an auditory model. *IEEE Transactions on Audio, Speech and Language Processing*, 16(2):255–266, February 2008.

D. Klatt and L. Klatt. Analysis, synthesis, and perception of voice quality variations among female and male talkers. *JASA*, 87:820–857, 1990.

J. Kornycky, B. Gunel, and A. Kondoz. Comparison of subjective and objective evaluation methods for audio source separation. *Journal of the Acoustical Society of America*, 123 (5):3569, 2008.

H. W. Kuhn and A. W. Tucker. Nonlinear programming. In *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, pages 481–492, 1951.

M. Lagrange, L. G. Martins, J. Murdoch, and G. Tzanetakis. Normalized cuts for predominant melodic source separation. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(2):278–290, February 2008. ISSN 1558-7916.

E. Large and J. Kolen. Resonance and the perception of musical meter. *Connection Science*, 6(2):3, 1994.

J. Le Roux, H. Kameoka, N. Ono, A. de Cheveigné, and S. Sagayama. Single channel speech and background segregation through harmonic-temporal clustering. In *Proceedings of the WASPAA 2007 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 279–282, 2007.

J. Le Roux, N. Ono, and S. Sagayama. Explicit consistency constraints for STFT spectrograms and their application to phase reconstruction. In *Proceedings of the SAPA 2008 ISCA Workshop on Statistical and Perceptual Audition*, pages 23–28, 2008.

D. Lee and H. Seung. Algorithms for Non-negative Matrix Factorization. *Advances in Neural Information Processing Systems*, pages 556–562, 2001.

D. D. Lee and H. S. Seung. Learning the parts of objects by nonnegative matrix factorization. *Nature*, 401:788 – 791, October 1999.

P. Leveau. *Décompositions parcimonieuses structurées : application à la représentation objet de la musique : modèles de signaux, algorithmes et applications.* PhD thesis, Université Pierre et Marie Curie (Paris VI), 2007.

Y. P. Li and D. L. Wang. Separation of singing voice from music accompaniment for monaural recordings. *IEEE Transactions on Audio, Speech And Language Processing*, 15(4):1475, 2007.

S. Mallat. *A Wavelet tour of signal processing, 3rd edition.* Academic Press, December 2008.

S. Mallat and Z. Zhang. Matching pursuits with time-frequency dictionaries. *IEEE transactions on Signal Processing*, 41(12):3397–3415, 1993.

M. Marolt. A connectionist approach to automatic transcription of polyphonic piano music. *IEEE Transactions on Multimedia*, 6(3):439–449, June 2004. doi: 10.1109/TMM.2004.827507.

M. Marolt. Audio melody extraction based on timbral similarity of melodic fragments. In *Proceedings of Eurocon*, Belgrade, SMN, 2005.

R. McAulay and M. Malpass. Speech enhancement using a soft-decision noise suppression filter. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 28(2):137–145, Apr 1980. ISSN 0096-3518.

R. J. McAulay and T. F. Quatieri. Speech analysis/synthesis based on a sinusoidal representation. *IEEE Transasctions on Acoustics, Speech and Signal Processing*, ASSP-34 (4):744 – 754, August 1986.

R. Meddis. Simulation of mechanical to neural transduction in the auditory receptor. *The Journal of the Acoustical Society of America*, 79(3):702–711, 1986. URL `http://link.aip.org/link/?JAS/79/702/1`.

G. H. Mohimani, M. Babaie-Zadeh, and C. Jutten. Complex-valued sparse representation based on smoothed l^0 norm. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 3881–3884, March 31 -April 4 2008.

M. Mørup, L. K. Hansen, S. M. Arnfred, L. Lim, and K. H. Madsen. Shift invariant multilinear decomposition of neuroimaging data. *accepted for publication NeuroImage*, 42(4):1439–1450, 2008. URL `http://www2.imm.dtu.dk/pubdb/p.php?5551`.

L. Oudre, Y. Grenier, and C. Févotte. Template-based chord recognition: influence of the chord types. In *Proceedings of the International Society for Music Information Retrieval conference*, volume 1, pages 153 – 158, Kobe, Japan, October 26 - 31 2009.

A. Ozerov. *Adaptation de modèles statistiques pour la séparation de sources mono-capteur. Application à la séparation voix / musique dans les chansons.* PhD thesis, University of Rennes 1, 2006.

A. Ozerov and C. Févotte. Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation. *IEEE Transactions on Audio, Speech and Language Processing*, 18(3):550 – 563, March 2010.

A. Ozerov, P. Philippe, F. Bimbot, and R. Gribonval. Adaptation of Bayesian Models for Single-Channel Source Separation and its Application to Voice/Music Separation in Popular Songs. *IEEE Transactions on Audio, Speech and Language Processing*, 15(5): 1564–1578, 2007.

R. Paiva. *Melody Detection in Polyphonic Audio*. PhD thesis, University of Coimbra, 2006.

R. P. Paiva, T. Mendes, and A. Cardoso. On the detection of melody notes in polyphonic audio. In *Proceedings of the International Conference on Music Information Retrieval*, London, UK, 11 - 15 September 2005.

H. Papadopoulos and G. Peeters. Simultaneous estimation of chord progression and downbeats from an audio file. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 121–124, Las Vegas, Nevada, USA, March 31 - April 4 2008.

S. Pauws. Cubyhum: A fully operational query by humming system. *ISMIR 2002 Conference Proceedings*, pages 187–196, 2002.

K. Pearson. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2(6):559 – 572, 1901.

G. Peeters. Template-based estimation of time-varying tempo. *EURASIP Journal on Advances in Signal Processing*, 2007, 2007. doi: 10.1155/2007/67215.

G. Peeters. Beat-marker location using a probabilistic framework and linear discriminant analysis. In *Proceedings of the Digital Audio Effects (DAFX) conference*, Come, Italy, 2009.

M. D. Plumbley. Algorithms for nonnegative independent component analysis. *IEEE Transactions on Neural Networks*, 4(3):534–543, May 2003.

G. Poliner, D. Ellis, A. Ehmann, E. Gómez, S. Streich, and B. Ong. Melody transcription from music audio: Approaches and evaluation. *IEEE Transactions on Audio, Speech, Language Processing*, 14(4):1247–1256, May 2007.

P. Ponce de León, D. Rizo, R. Ramirez, and J. Iñesta. Melody characterization by a genetic fuzzy system. In M. Supper and S. Weinzierl, editors, *Proceedings of the 5th Sound and Music Computing Conference*, pages 15–23. Universitätsverlag der TU Berlin, July 2008.

L. Rabiner. A tutorial on hidden Markov models and selected applications inspeech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

B. Raj, P. Smaragdis, M. Shashanka, and R. Singh. Separating a foreground singer from background music. *International Symposium on Frontiers of Research on Speech and Music (FRSM)*, January 2007. Mysore, India.

V. Rao and P. Rao. Melody extraction using harmonic matching. *Music Information Retrieval Evaluation eXchange*, 2008.

D. Rizo, P. J. Ponce de León, C. Pérez-Sancho, A. Pertusa, and J. M. Iñesta. A pattern recognition approach for melody track selection in MIDI files. In *Proceedings of the Internation Society for Music Information Retrieval conference*, Victoria, Canada, 8 - 12 October 2006.

S. Roweis. One microphone source separation. *Advances in Neural Information Processing Systems*, 13:793–799, 2001.

M. Ryynänen and A. Klapuri. Query by Humming of MIDI and audio using locality sensitive hashing. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 2249–2252, Las Vegas, Nevada, USA, April 2008a.

M. Ryynänen and A. Klapuri. Polyphonic music transcription using note event modeling. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 319–322, 16-19 October 2005.

M. Ryynänen and A. Klapuri. Modelling of note events for singing transcription. *Proceedings of ISCA Tutorial and Research Workshop on Statistical and Perceptual Audio Processing*, 2004.

M. Ryynänen, T. Virtanen, J. Paulus, and A. Klapuri. Accompaniment separation and karaoke application based on automatic melody transcription. *IEEE International Conference on Multimedia and Expo*, pages 1417–1420, 2008.

M. P. Ryynänen and A. P. Klapuri. Automatic transcription of melody, bass line, and chords in polyphonic music. *Computer Music Journal*, 32(3):72–86, 2008b.

M. P. Ryynänen and A. P. Klapuri. Transcription of the singing melody in polyphonic music. *Proceedings of the International Conference on Music Information Retrieval*, pages 222–227, 2006.

E. D. Scheirer. Tempo and beat analysis of acoustic musical signals. *Journal of the Acoustic Society of America*, 103(1):588–601, January 1998.

J. Serrà, E. Gómez, P. Herrera, and X. Serra. Chroma binary similarity and local alignment applied to cover song identification. *IEEE Transactions on Audio, Speech and Language Processing*, 16:1138–1151, August 2008.

SiSEC. Professionally produced music recordings. Internet page: `http://sisec.wiki.irisa.fr/tiki-index.php?page=Professionally+produced+music+recordings`, 2008.

M. Slaney, D. Naar, and R. Lyon. Auditory model inversion for sound separation. In *IEEE Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, volume 2, Adelaide, Australia, April 1994.

P. Smaragdis and J. C. Brown. Non-negative matrix factorization for polyphonic music transcription. *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 177–180, 2003.

P. Smaragdis, B. Raj, and M. Shashanka. Sparse and shift-invariant feature extraction from non-negative data. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, Las Vegas, Nevada, USA, April 2008.

S. S. Stevens. A scale for the measurement of the psychological magnitude: loudness. *Psychological Review, APA Journals*, 43(5):405–416, 1936.

C. Sutton, E. Vincent, M. Plumbley, and J. Bello. Transcription of vocal melodies using voice characteristics and algorithm fusion. *Extended abstract for* the Music Information Retrieval Evaluation eXchange, September 2006.

University of Illinois Urbana Champaign USA. Music Information Retrieval Evaluation eXchange. online: http://www.music-ir.org/mirex/2008/, September 2008.

E. Vincent. *Modeles d'instruments pour la separation de sources et la transcription d'enregistrements musicaux*. PhD thesis, Université Paris VI and IRCAM, Paris, France, 2004.

E. Vincent. Musical source separation using time-frequency source priors. *IEEE Transactions on Audio, Speech and Language Processing*, 14(1):91–98, 2006.

E. Vincent, R. Gribonval, and C. Févotte. Performance measurement in blind audio source separation. *IEEE Transactions on Audio, Speech and Language Processing*, 14(4):1462–1469, July 2006.

E. Vincent, N. Bertin, and R. Badeau. Harmonic and inharmonic Nonnegative Matrix Factorization for Polyphonic Pitch transcription. *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, pages 109–112, 2008.

E. Vincent, S. Araki, and P. Bofill. The 2008 signal separation evaluation campaign: A community-based approach to large-scale evaluation. In *Proc. Int. Conf. on Independent Component Analysis and Blind Source Separation (ICA)*, pages 734–741, Paraty, Brazil, 15-18 March 2009.

M. Vinyes. MTG MASS database. `http://www.mtg.upf.edu/static/mass/resources`, 2008.

T. Virtanen. Monaural Sound Source Separation by Nonnegative Matrix Factorization With Temporal Continuity and Sparseness Criteria. *IEEE Transactions on Audio, Speech and Language Processing*, 15(3):1066–1074, 2007.

T. Virtanen. *Sound Source Separation in Monaural Music Signals*. PhD thesis, Tampere University of Technology, November 2006.

T. Virtanen and A. Klapuri. Analysis of polyphonic audio using source-filter model and non-negative matrix factorization. In *Advances in Models for Acoustic Processing, Neural Information Processing Systems Workshop*, 2006.

A. Viterbi. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE transactions on Information Theory*, 13(2):260–269, 1967.

R. M. Warren. Elimination of biases in loudness judgments for tones. *The Journal of the Acoustical Society of America*, 48(6B):1397–1403, 1970. URL `http://link.aip.org/link/?JAS/48/1397/1`.

J. Weil, J.-L. Durrieu, G. Richard, and T. Sikora. Beat tracking using the delta-phase matrix. Technical report, Département Traitement du Signal et des Images, Groupe AAO : Audio, Acoustique et Ondes, Télécom ParisTech, Paris, France, August 2009a.

J. Weil, T. Sikora, J.-L. Durrieu, and G. Richard. Automatic generation of lead sheets from polyphonic music signals. In *Proceedings of International Society fo Music Information Retrieval Conference*, Kobe, Japan, 26-30 October 2009b.

R. Weiss and D. Ellis. A variational EM algorithm for learning eigenvoice parameters in mixed signals. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, pages 113 – 116, Taipei, Taiwan, April 2009.

R. Weiss and D. Ellis. Speech separation using speaker-adapted eigenvoice speech models. *Computer Speech & Language*, 24(1):16 – 29, 2010. ISSN 0885-2308. URL `http://www.sciencedirect.com/science/article/B6WCW-4S2F5KK-1/2/fc22a4fea7b29ca989d96914741205e2`. Speech Separation and Recognition Challenge.

M. Wendelboe. Using OQSTFT and a modified SHS to detect the melody in polyphonic music (mirex 2009). Extended abstract for the Music Information Retrieval Evaluation eXchange, September 2009.

Index

List of Tables

List of Figures

List of Algorithms