Source/Filter Model for Unsupervised Main Melody Extraction From Polyphonic Audio Signals

Jean-Louis Durrieu, Gael Richard, Bertrand David, Cédric Févotte

Source/Filter Model for Unsupervised Main Melody Extraction From Polyphonic Audio Signals

Jean-Louis Durrieu, Gaël Richard, Bertrand David and Cédric Févotte

***Abstract*— Extracting the main melody from a polyphonic music recording seems natural even to untrained human listeners. To a certain extent it is related to the concept of source separation, with the human ability of focusing on a specific source in order to extract relevant information. In this article, we propose a new approach for the estimation and extraction of the main melody (and in particular the leading vocal part) from polyphonic audio signals. To that aim, we propose a new signal model where the leading vocal part is explicitly represented by a specific source/filter model. The proposed representation is investigated in the framework of two statistical models: a Gaussian Scaled Mixture Model (GSMM) and an extended Instantaneous Mixture Model (IMM). For both models, the estimation of the different parameters is done within a maximum likelihood framework adapted from single-channel source separation techniques. The desired sequence of fundamental frequencies is then inferred from the estimated parameters. The results obtained in a recent evaluation campaign (MIREX08) show that the proposed approaches are very promising and reach state-of-the-art performances on all test sets.**

***Index Terms*— Music, Source/Filter Model, Main Melody Extraction, Blind Audio Source Separation, Spectral Analysis, Maximum Likelihood, Expectation-Maximization (EM) algorithm, Gaussian Scaled Mixture Model (GSMM), Non-negative Matrix Factorization (NMF)**

I. INTRODUCTION

THE "main melody" of a polyphonic music excerpt commonly refers to the sequence of notes played by a single monophonic instrument (including singing voice) over a potentially polyphonic accompaniment. If humans have a natural ability to identify and, to a certain extent, isolate this main melody from a polyphonic music recording, its automatic extraction and transcription by a machine remains a very challenging task despite the recent efforts of the research community.

The main melody sequence is a feature of great interest since it carries a significant amount of semantically rich information about a music piece and appears to be particularly useful for a number of Music Information Retrieval (MIR) applications. For instance, it can be directly used in systems such as Query-By-Humming or Query-By-Singing systems [1]. It can also be exploited for music structuring [2], music similarity search such as cover version detection [3], and to a certain extent in copyright protection.

This work was partly supported by the European Commission under contract *FP6-027026-K-SPACE* and the OSEO project *QUAERO*.

J.-L. Durrieu, G. Richard and B. David are with Institut TELECOM ; TELECOM ParisTech ; CNRS LTCI - 46 rue Barrault - 75634 Paris Cedex 13 - France. C. Févotte is with CNRS LTCI ; TELECOM ParisTech - 46 rue Barrault - 75634 Paris Cedex 13 - France. e-mails: firstname.lastname@telecom-paristech.fr.

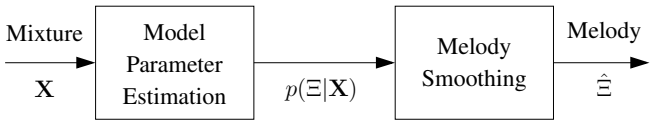


Fig. 1. Proposed system outline: $\mathbf{X}$ is the short-time Fourier transform (STFT) of the mixture signal, $p(\Xi|\mathbf{X})$ the posterior probability of a given melody sequence Ξ, and $\hat{\Xi}$ the desired smooth melody sequence.

Several types of methods have been proposed to address the problem, and most of them are parametric. The estimation then relies on a signal model, e.g. a probabilistic modeling of the spectrogram in [4] or using more classical signal processing solutions as in [5] or [6]. These systems are not limited to these categories, and often use several heuristics and statistical methods to achieve their goal. Another possibility is the use of classification schemes, such as [7]. The first kind of methods usually introduce generative models for the signal, while the latter method is related to perceptive aspects of the task.

The common underlying concept followed by these systems is a two step process: first, the signal is mapped onto a feature space, and then these features are post-processed to track the melody line. The feature space can directly be a mapping on the Fourier domain [7] but most of the approaches aim at obtaining higher level features or objects, such as pitch candidates as in [5] and [6]. As depicted in Figure 1, the hereafter proposed system is a two-step melody tracker as well and relies on a parameterization of the power spectrogram. The parameters are first estimated and the posterior probabilities of potential melody sequences are then computed. At last, the melody smoothing block outputs the desired sequence $\hat{\Xi}$.

Our approach includes several original contributions. First, specific (and different) models are used for each component (leading instrument vs accompaniment) of the music mixture to take into account their specificities and/or 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 production characteristics of the human voice compared to any other instrument as in [8]. 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). On the other hand, the accompaniment includes 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, etc.). To exploit this relative pitch stability and temporal repetitive structure, the model for the accom-

paniment is inspired by Non-negative Matrix Factorization (NMF) with the Itakura-Saito divergence [9]. The proposed systems 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.

Second, the leading voice is modeled in a statistical framework in which two different generative models are proposed, both of them including the previously mentioned source/filter parameterization. The first model is a source/filter Gaussian Scaled Mixture Model (GSMM) [10] while the second one is a more general Instantaneous Mixture Model (IMM). Our generative model is essentially inspired by single-channel blind source separation approaches presented in [10] and [11]. We can therefore also proceed to the actual separation of the estimated solo part and background part which can be useful for other applications such as audio remixing, karaoke or polyphonic music transcription. The proposed methods are unsupervised, and thus differ from the supervised techniques of [10] and [11].

Third, it is commonly accepted that most melody lines exhibit a limited variation from one note to the next in terms of relative energy and interval. To take into account this property, it is then proposed to exploit a smoothing strategy based on an adapted Viterbi algorithm to track, among the most probable sequences of fundamental frequencies obtained in the first step, the sequence that reaches the best trade-off between the energy of the path and its regularity. This strategy relaxes the assumption that, in each analysis frame, the fundamental frequency is the most energetic one. The resulting melody sequence is then physically more relevant.

The results obtained are very promising and the evaluation conducted in the framework of the international Music Information Retrieval Evaluation eXchange (MIREX) 2008 campaign on the audio melody extraction task[1] has shown that our algorithms achieve state-of-the-art performances on various sets of music material.

This article is organized as follows: the different signal models introduced are detailed in section II. The estimation of the model parameters is discussed in section III. The smoothing post-processing stage which allows to obtain the desired melody sequence is described in section IV. The results of audio main melody extraction are presented in section V, where we also give some insights about two applications of our approach, namely source separation and multipitch tracking. Finally, some conclusions and future extensions are suggested in section VI.

II. Signal Models

A. Notations

The short-time Fourier transform (STFT) of a time-domain signal y is denoted by the $F \times N$ matrix $\mathbf{Y}$, F being the Fourier transform size and N the number of analysis frames. $\mathbf{S}_Y$ denotes the $F \times N$ matrix whose columns are the power spectrum densities (PSD) of consecutive frames of a signal y.

[1] http://www.music-ir.org/mirex/2008/

For a matrix $\mathbf{A}$, we define the notation for the element at the i-th row and j-th column $a_{ij} = [\mathbf{A}]_{ij}$, convenient for matrix products. The j-th column of $\mathbf{A}$ is denoted as the vector $\mathbf{a}_j$.

B. Modeling the spectra of the signals

We assume that the signals are wide-sense stationary (w.s.s.) within each analysis frame. For frame n, the Fourier transform $\mathbf{y}_n$ of signal y is considered as a centered proper complex Gaussian variable. We further assume that the covariance matrix of $\mathbf{y}_n$ is diagonal, with diagonal coefficients equal to the PSD $\mathbf{s}_{Y,n}$, as in [10]: this is equivalent to neglecting the correlation between two frequency channels of the Fourier transform, i.e. ignoring the spectral spread due to windowing.

A (scalar) complex variable is centered proper Gaussian if both its real and imaginary parts are independent centered Gaussian variables, with the same variance. The likelihood of the STFT $y_{fn} = \rho_{fn} \exp(i\phi_{fn})$ at frequency bin f and frame n is therefore defined as:

$$p(y_{fn}) = p(\rho_{fn}, \phi_{fn}) = \frac{\rho_{fn}}{\pi s_{Y,fn}} \exp\left(-\frac{\rho_{fn}^2}{s_{Y,fn}}\right) \quad (1)$$

We denote a random variable following (1) with the following convention: $y_{fn} \sim \mathcal{N}_c\left(0, s_{Y,fn}\right)$, and for the vector $\mathbf{y}_n \sim \mathcal{N}_c\left(0, \mathrm{diag}(\mathbf{s}_{Y,n})\right)$. Note that such a definition also implies that the phase of the complex variable is uniformly distributed.

The models we propose essentially put spectral and temporal constraints on the PSD $\mathbf{s}_{Y,n}$. As shown in [9], estimating the PSD in this framework is equivalent to fitting the power spectrogram $|\mathbf{y}_n|^2$ with the (constrained) PSD $\mathbf{s}_{Y,n}$, using the Itakura-Saito divergence as cost function.

C. Mixture signal

The observed musical mixture signal x is the sum of two contributions v, the leading instrument, and m, the musical accompaniment. Therefore, their STFTs verify:

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

In this paper, we consider musical pieces or excerpts where such a leading instrument is clearly identifiable and unique. The latter assumption particularly implies that the melody line is not harmonized with multiple voices. We assume that its energy is mostly predominant over the other instruments of the mixture. These can thus be assimilated to the accompaniment. This implies that we are tracking an instrument with a rather high average energy in the processed song and a continuous fundamental frequency line. In this section and in section III, the parameters mainly reflect the spectral shapes and the amplitudes, in other words the energy. In Section IV, we focus more on the melody tracking and therefore propose a model for the continuity of the melodic line.

Figure 2 shows the general principle of the parameterization of the mixture signal: a source/filter model is fitted to the main instrument part (Section II-D), while the residual accompaniment is modeled in an NMF framework (Section II-E).

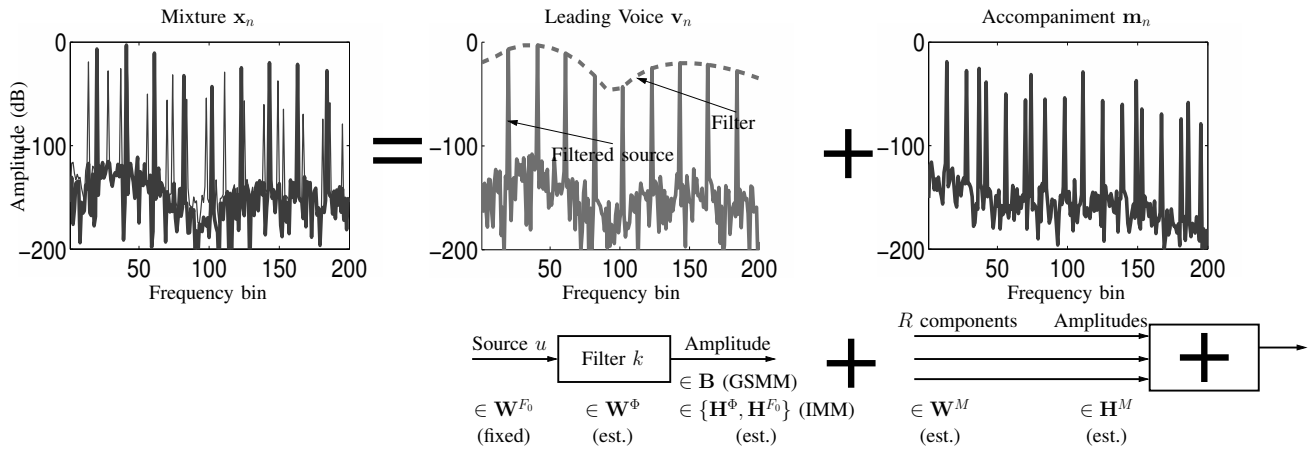


Fig. 2. Principle for the decomposition of one frame of the mixture STFT into leading voice and accompaniment spectra. The parameters indicated here are presented in Section II. The source spectral shapes are fixed as explained in Appendix I and the other parameters are estimated directly from the audio signal.

D. A source/filter model to fit the main instrument parts

Let v and $\mathbf{V}$ respectively denote the main voice time-domain signal and its STFT. Unlike in previous works on speech/music separation [10] and singer/music separation [11], the pitched aspect of the spectral shapes used to identify the main part is here fundamental. We are interested in transcribing the melody itself, i.e. the fundamental frequencies that are sung or played, which are closely related to the pitched components of the signal. Therefore, in order to obtain pitch constrained spectra, and inspired by speech processing modeling techniques, we propose a conventional source/filter model of the principal instrument signal [12] for which the source part is harmonic (voiced source) and fixed.

Only the pitched segments of the main part are modeled, unpitched or unvoiced segments are therefore rejected as belonging to the accompaniment. In source/filter modeling, the voiced speech signal is produced by an excitation, depending on a fundamental frequency, which is then filtered by a vocal tract shape, providing the pronounced vowel. At first, the model presented in this paper was designed for singer signals as a realistic production model. It can also be extended to some music instruments, for which the filter part is then interpreted as shaping the timbre of the sound, while the source part mainly consists in a more generic harmonic signal driven by the fundamental frequency.

Our strategies rely on a decomposition of the main voice signal onto several hidden states or elementary components. In practice, the decomposition of the STFT is done onto a limited number of spectral components. In our source/filter model, the filter is independent from the source and its fundamental frequency, and the filter and source parts can therefore be modeled independently. The range of the source spectra corresponds to the range of notes the singer or instrument can play. The discrete range of filters corresponds to a limited number of possible timbres or vowels pronounced in the main voice. Under certain assumptions, we could for example consider that each of the estimated filters represents a specific vowel such as [a], [e] and so on.

Let U be the number of possible fundamental frequencies (notes) for the main part and K the number of "vocal tract" filters. The elementary variance for a filter-source couple $(k, u) \in [1, K] \times [1, U]$ is the product $w^{\Phi}_{fk} w^{F_0}_{fu}$ for $f \in [1, F]$: $w^{F_0}_{fu}$ is the variance of the source for a fundamental frequency number u and w^{Φ}_{fk} is the squared magnitude of the frequency response of filter k at frequency bin f. The $F \times U$ matrix $\mathbf{W}^{F_0}$ is the source spectra dictionary. Each source spectrum is parameterized by a fundamental frequency $f_0 = \mathcal{F}(u)$, where the function $\mathcal{F}$ maps the number u of the spectrum to a given frequency f_0 in Hz. Some more details are given in Appendix I. For the filters, we assume that they have real frequency responses, since equation (1) shows that our model discards the phase information from the likelihood $p(\mathbf{X})$[2]. $\mathbf{W}^{\Phi}$ is the $F \times K$ filter spectral shape matrix. $\mathbf{W}^{\Phi}$ is normalized such that each of its columns sums to 1 and $\mathbf{W}^{F_0}$ such that the maximum value of each column is equal to 1.

From this general framework, we derive two different models. The first one is the GSMM framework [10] adapted to our source/filter model; the second one relaxes the generative condition on the number of sources per frame. This latter model was motivated by the need of faster estimation schemes, as well as a more flexible model, inspired by NMF methodology. We investigate and compare these models in the following sections.

1) Gaussian Scaled Mixture Model (GSMM): Following [10], we define a GSMM for which the states are all the couples $(k, u) \in [1, K] \times [1, U]$. Under the conditions discussed in section II-B for signal v and its STFT $\mathbf{V}$, the likelihood of $\mathbf{v}_n$, for frame n, conditionally upon the state pair $Z_n = (k, u)$, is:

$$\mathbf{v}_n | Z_n \sim \mathcal{N}_c(0, b_{kun} \mathrm{diag}(\mathbf{w}^{\Phi}_k \bullet \mathbf{w}^{F_0}_u)) \tag{2}$$

where b_{kun} is the amplitude coefficient for state pair (k, u) at frame n and $\bullet$ denotes the Hadamard (entry-wise) product.

[2]For a given set of parameter θ, the likelihood should write $p(\mathbf{X}|\theta)$. However, for simplicity, and since there is no ambiguity in our context, the likelihood is here denoted $p(\mathbf{X})$. Note in particular that it is not the marginal likelihood, defined as the integration of the likelihood over all the possible parameter sets θ.

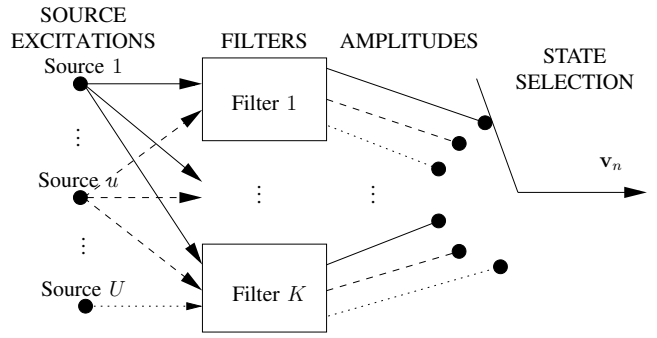


Fig. 3. Schematic principle of the generative GSMM for the main instrument part. Each source u is filtered by each filter k. For frame n, the signal is then multiplied by a given amplitude and a "state selector" then chooses the active state.

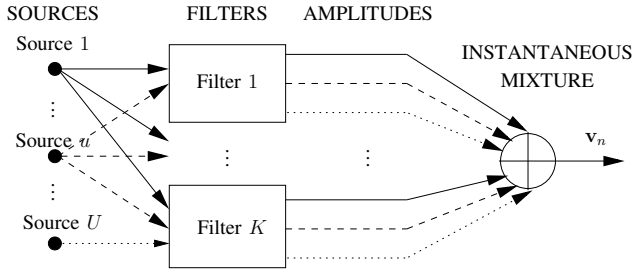


Fig. 4. Schematic principle of the generative IMM for the main instrument part. At each frame, all the U sources, each filtered by the K filters, are multiplied by amplitudes and added together to produce the leading voice signal.

Then the observation likelihood verifies:

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

$$\Leftrightarrow \qquad \mathbf{v}_n \sim \sum_{k,u} \pi_{ku} \mathcal{N}_c(0, b_{kun} \text{diag}(\mathbf{w}_k^\Phi \bullet \mathbf{w}_u^{F_0})) \quad (3)$$

where the prior probability of state $Z = (k,u)$ is denoted π_{ku}. These probabilities verify $\sum_{k,u} \pi_{ku} = 1$. For convenience, from now on, the conditional likelihoods $p\left(.|Z_n = (k,u)\right)$ are abbreviated to $p(.|k,u)$. We denote the variance for the main instrument, given the state pair (k,u), at frequency f and frame n as follows:

$$s_{V,fn|ku} = b_{kun} w_{fk}^\Phi w_{fu}^{F_0} \quad (4)$$

Such a model is formally very similar to a Gaussian Mixture Model (GMM), with an additional degree of freedom: at each frame n, the non-negative amplitude coefficient b_{kun} corresponding to state (k,u) allows the scaling of the variance to the actual energy of the frame (source and filter spectra are normalized). As a generative model, if (k,u) differs from the active state Z_n, then b_{kun} can take any value. In the Maximum Likelihood (ML) estimation explained in section III, there is however no ambiguity for these parameters. We compute b_{kun} as being the amplitude maximizing the likelihood (2), as if (k,u) were, at frame n, the active state.

Figure 3 shows the diagram of the GSMM model for the main voice part. Each source excitation u is filtered by each filter k. The amplitudes for a frame n and for all the couples (k,u) are then applied to each of the output signals. At last a "state selector" sets the active state for the given frame n.

2) Instantaneous Mixture Model (IMM): Models like the GSMM have a heavy computational load and the second model we propose aims at reducing this load while staying close to the original generative GSMM model. Here, the random variable $\mathbf{v}_n$ is obtained as a weighted sum of sub-spectra $\boldsymbol{\nu}_{kun}$, each corresponding to the combination of the filter k with the source u: $\mathbf{v}_n = \sum_{k,u} \boldsymbol{\nu}_{kun}$. Each sub-spectrum is assumed to be Gaussian such that:

$$\boldsymbol{\nu}_{kun} \sim \mathcal{N}_c\left(0, h_{kn}^\Phi h_{un}^{F_0} \text{diag}(\mathbf{w}_k^\Phi \bullet \mathbf{w}_u^{F_0})\right)$$

where $\mathbf{H}^\Phi$ and $\mathbf{H}^{F_0}$ are the amplitudes matrices for the filters and the sources such that $h_{kn}^\Phi \geq 0$ (resp. $h_{un}^{F_0} \geq 0$) is the amplitude factor associated with the filter component k (resp. source element u), for frame n. We normalize the columns of $\mathbf{H}^\Phi$ such that they sum to 1. Since both matrices $\mathbf{W}^\Phi$ and $\mathbf{W}^{F_0}$ are also normalized, the energy for the main instrument part is mostly represented by the amplitudes in H^{F_0}.

The sub-spectra are mutually independent. Their sum $\mathbf{v}_n$ is therefore also Gaussian and verifies:

$$\mathbf{v}_n \sim \mathcal{N}_c(0, \sum_{k,u} h_{kn}^\Phi h_{un}^{F_0} \text{diag}(\mathbf{w}_k^\Phi \bullet \mathbf{w}_u^{F_0})) \quad (5)$$

$$\mathbf{v}_n \sim \mathcal{N}_c\left(0, \text{diag}(\left[\mathbf{W}^\Phi \mathbf{h}_n^\Phi\right] \bullet \left[\mathbf{W}^{F_0} \mathbf{h}_n^{F_0}\right])\right) \quad (6)$$

Note how Eq. (5) differs from Eq. (3): in the GSMM, the likelihood of the voice signal is a weighted sum of likelihoods, while in the IMM, it is the variance that is a weighted sum of variances. The variance of the likelihood of an individual time-frequency bin of the vocal signal can be written with matrix factors:

$$s_{V,fn} = \left[(\mathbf{W}^\Phi \mathbf{H}^\Phi) \bullet (\mathbf{W}^{F_0} \mathbf{H}^{F_0})\right]_{fn} \quad (7)$$

This highlights the link between this parameterization and NMF.

Furthermore, from a generative point of view, the IMM diagram Fig. 4 clearly shows how the IMM differs from the GSMM . Instead of selecting only one output in the end, all the filtered outputs are added together to form $\mathbf{v}_n$. There however exists an implicit link between these two models in our framework which we discuss in the next section.

3) Bridging the models: The GSMM is closer to modeling a monophonic voice, since by construction only one state, i.e. one source and one filter, is active at each frame. The IMM, under certain circumstances, can also fit a monophonic voice, but does not inherently do so.

From a generative point of view, the second model can be reduced to the first one by constraining the amplitudes in $\mathbf{H}^\Phi$ and $\mathbf{H}^{F_0}$. For a frame n, to generate $\mathbf{v}_n$ from the GSMM, we need to draw the active state $Z_n = (\gamma, \mu)$ from the prior densities π_{ku}. In this case, we know exactly that $p(\mathbf{v}_n) = p(\mathbf{v}_n | \gamma, \mu)$ and the variance, or equivalently the PSD, of v_{fn} is $s_{V,fn} = b_{\gamma\mu n} w_{f\gamma}^\Phi w_{f\mu}^{F_0}$. Assuming the estimated filters $\mathbf{W}^\Phi$ for the IMM are the same as for the GSMM, the same PSD $\mathbf{s}_{V,n}$ is obtained for the IMM if we constrain the amplitudes

such that:

$$h^{\Phi}_{kn} h^{F_0}_{un} = \begin{cases} b_{\gamma\mu n}, \text{ if } k=\gamma \text{ and } u=\mu \\ 0, \text{ otherwise} \end{cases} \quad (8)$$

where $\delta_{a=b} = 1$ if $a = b$ and 0 otherwise. The above equation, with the normalization of the columns of $\mathbf{H}^{\Phi}$ yields to:

$$\begin{cases} h^{\Phi}_{kn} &= \delta_{k=\gamma} \\ h^{F_0}_{un} &= b_{\gamma\mu n}\delta_{u=\mu} \end{cases}$$

However, during the estimation step, the IMM is not constrained, in order to be more flexible and allow the model to better adapt to the signal.

E. Background music model

The accompaniment STFT $\mathbf{M}$ is the weighted instantaneous mixture of R elementary sources STFT $\mathbf{M}_r$, $r \in [1, R]$. Each of these signals is Gaussian, centered, with variance at frequency bin f and frame n equal to $w^M_{fr} h^M_{rn}$. $\mathbf{W}^M$ is the $F \times R$ matrix of accompaniment spectral shapes. The amplitudes form a $R \times N$ matrix $\mathbf{H}^M$. $\mathbf{m}_n$ is also a centered Gaussian, and the covariances add up such that:

$$\begin{aligned} \mathbf{m}_n &\sim \mathcal{N}_c\left(0, \sum_{r=1}^{R} h^M_{rn} \text{diag}(\mathbf{w}^M_r)\right) \\ &\sim \mathcal{N}_c\left(0, \text{diag}(\mathbf{W}^M \mathbf{h}^M_n)\right) \end{aligned} \quad (9)$$

where the PSD of $\mathbf{m}_n$, $\mathbf{s}_{M,n}$ can be identified with the diagonal of the covariance matrix of the Gaussian:

$$s_{M,fn} = \sum_{r=1}^{R} w^M_{fr} h^M_{rn} = \left[\mathbf{W}^M \mathbf{H}^M\right]_{fn} \quad (10)$$

F. Statistics of the mixture signal

In our model, the temporal dimension is not taken into account, and the frames are assumed to be independent realizations. Therefore:

$$p(\mathbf{X}) = \prod_n p(\mathbf{x}_n) \quad (11)$$

1) Statistics of the mixture signal with the GSMM for v: The likelihood of $\mathbf{x}_n$ is the weighted sum of the conditional likelihoods, sum over the states of the vocal part:

$$p(\mathbf{x}_n) = \sum_{k,u} \pi_{ku} p(\mathbf{x}_n | k, u) \quad (12)$$

where $p(\mathbf{x}_n|k,u)$ is the likelihood of the STFT $\mathbf{X}$ conditional upon the state pair (k, u) of the mixture signal. We have assumed that the Fourier transforms $\mathbf{v}_n$ for the main voice and $\mathbf{m}_n$ for the accompaniment are centered Gaussians. We also assume that, conditionally upon the state Z_n for the main instrument, $\mathbf{v}_n$ and $\mathbf{m}_n$ are independent. Therefore, their sum is also Gaussian, centered, with the covariance matrix equal to the sum of the corresponding diagonal covariances $\text{diag}(\mathbf{s}_{V,n|ku})$ and $\text{diag}(\mathbf{s}_{M,n})$. The resulting matrix is therefore diagonal, with on the diagonal the PSD $\mathbf{s}_{X,n|ku}$ such that:

$$\begin{aligned} \mathbf{s}_{X,n|ku} &= \mathbf{s}_{V,n|ku} + \mathbf{s}_{M,n} \quad (13) \\ &= b_{kun} \mathbf{w}^{\Phi}_k \bullet \mathbf{w}^{F_0}_u + \mathbf{W}^M \mathbf{h}^M_n \\ s_{X,fn|ku} &= b_{kun} w^{\Phi}_{fk} w^{F_0}_{fu} + \left[\mathbf{W}^M \mathbf{H}^M\right]_{fn} \quad (14) \end{aligned}$$

where we have used equations (4) and (10). The conditional likelihood at frame n follows:

$$p(\mathbf{x}_n|k,u) = \prod_f \frac{|x_{fn}|}{\pi s_{X,fn|ku}} \exp\left(-\frac{|x_{fn}|^2}{s_{X,fn|ku}}\right) \quad (15)$$

We denote the $K \times U \times N$ tensor of the amplitudes b_{kun} by $\mathbf{B}$ and $\mathbf{\Pi} = \{\pi_{ku}; (k,u) \in [1,K] \times [1,U]\}$. We estimate the set of parameters $\theta_{\text{GSMM}} = \{\mathbf{\Pi}, \mathbf{B}, \mathbf{W}^{\Phi}, \mathbf{H}^M, \mathbf{W}^M\}$ for this GSMM formulation in a Maximum Likelihood framework using an EM algorithm detailed in section III. $\mathbf{W}^{F_0}$ is fixed as explained in Appendix I and is therefore not estimated.

2) Instantaneous mixture model: For the IMM, the signals $\mathbf{v}_n$ and $\mathbf{m}_n$ are also assumed independent. Hence, we obtain a relation between the signal PSDs similar to (13), at frame n:

$$\mathbf{s}_{X,n} = \mathbf{s}_{V,n} + \mathbf{s}_{M,n}$$

With the equations (7) and (10), for frequency f and frame n, it leads to:

$$s_{X,fn} = \left[(\mathbf{W}^{\Phi}\mathbf{H}^{\Phi}) \bullet (\mathbf{W}^{F_0}\mathbf{H}^{F_0}) + \mathbf{W}^M\mathbf{H}^M\right]_{fn} \quad (16)$$

And the observation likelihood is then directly obtained from (1):

$$p(\mathbf{x}_n) = \prod_f \frac{|x_{fn}|}{\pi s_{X,fn}} \exp\left(-\frac{|x_{fn}|^2}{s_{X,fn}}\right) \quad (17)$$

The following section explicits how we estimate the different parameters of the IMM,
$\theta_{\text{IMM}} = \{\mathbf{W}^{\Phi}, \mathbf{H}^{\Phi}, \mathbf{H}^{F_0}, \mathbf{W}^M, \mathbf{H}^M\}$. $\mathbf{W}^{F_0}$ is also fixed as explained in Appendix I.

III. Parameter estimation by Maximum Likelihood

A. Maximum Likelihood principle

The proposed model for the mixture sound x is a probabilistic model. We can therefore estimate the set of parameters $\theta = \theta_{\text{GSMM}}$ or θ_{IMM} by a ML method:

$$\hat{\theta} = \arg\max_{\theta} p_{\theta}(\mathbf{X}) \quad (18)$$

B. Expectation-Maximization algorithm for the GSMM

The Expectation-Maximization (EM) algorithm is based on the maximization of the expectation of the joint log-likelihood for the observations and the hidden states, conditionally upon the observations. In this section, we consider the GSMM set of parameters $\theta = \theta_{\text{GSMM}}$. Let $i \in [1, I]$ the iteration number, $\theta^{(i)}$ the set of parameters updated at iteration i, $Z = \{Z_n = (k_n, u_n); n \in [1, N]\}$ the sequence of active states for the whole observation sequence. A Lagrangian term is added to the criterion, to express the condition over the prior probabilities $\mathbf{\Pi}$ in equation (3).

For $i > 0$, we define the GSMM criterion:

$$\begin{aligned} C_{\text{GSMM}}(\theta, \theta^{(i-1)}) &= E_{\theta^{(i-1)}}\left[\log p_{\theta}(\mathbf{X}, Z) | \mathbf{X}\right] \\ &\quad - \lambda\left(\sum_{k,u} \pi_{ku} - 1\right) \end{aligned} \quad (19)$$

One can show that maximizing $\theta^{(i)}$ such that:

$$\theta^{(i)} = \arg\max_{\theta} C_{\text{GSMM}}(\theta, \theta^{(i-1)}) \tag{20}$$

is equivalent to a non-decreasing observation likelihood [13]. The EM algorithm at least allows us to obtain a local maximum of the target likelihood. Here, we have:

$$\begin{aligned}\log p(\mathbf{X}, Z) &= \sum_n \log p(\mathbf{x}_n, Z_n) \\ &= \sum_n \log p(\mathbf{x}_n|k_n, u_n) + \log \pi_{k_n,u_n} \\ &= \sum_{n,k,u} \left[\log p(\mathbf{x}_n|k,u) + \log \pi_{ku}\right] \delta_{\{k=k_n,u=u_n\}} \end{aligned} \tag{21}$$

The first equation comes from the mutual independence of the observations over the frames, as expressed in equation (11). The second equation is a classical result for conditional probabilities, and where Z_n was replaced by the corresponding active states k_n and u_n. At last, equation (21) is a false sum over the states. This equation allows us to find a convenient way of expressing the criterion (19):

$$\begin{aligned} C_{\text{GSMM}}(\theta, \theta^{(i-1)}) = \sum_{n,k,u} &\left[\log p_\theta(\mathbf{x}_n|k,u) + \log \pi_{ku}\right] \\ &\times E_{\theta^{(i-1)}}\left[\delta_{\{k=k_n,u=u_n\}}|\mathbf{X}\right] \\ &- \lambda\left(\sum_{k,u} \pi_{ku} - 1\right) \end{aligned}$$

Furthermore, by definition of the expectation,

$$E_{\theta^{(i-1)}}\left[\delta_{\{k=k_n,u=u_n\}}|\mathbf{X}\right] = p_{\theta^{(i-1)}}(k,u|\mathbf{x}_n)$$

where we used the fact that the couple state (k_n, u_n) 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 the Bayes theorem:

$$p_{\theta^{(i-1)}}(k,u|\mathbf{x}_n) \propto p_{\theta^{(i-1)}}(\mathbf{x}_n|k,u)\pi_{ku}^{(i-1)} \tag{22}$$

The conditional likelihood of the observations upon the states is given by equations (14) and (15), using the parameters in $\theta^{(i-1)}$. The expression of the criterion is at last given in equation (23), where $s_{X,fn|ku}$ is calculated from the model parameters in θ, with equation (14). The term "CST" is a constant independent from the parameter set θ.

The M step then consists in updating the parameter set $\theta^{(i-1)}$ to obtain $\theta^{(i)}$ such that the criterion (23) is maximized. In order to find the updating rules for a parameter $\theta_j \in \theta$, we derivate the criterion with respect to θ_j and set $\theta_j^{(i)}$ such that it is a zero of the partial derivative.

Here, we adopt multiplicative updating rules, inspired by Non-negative Matrix Factorization (NMF) methodology [14]. The updated parameter is derived from the previous one by the equation

$$\theta_j^{(i)} = \alpha\theta_j^{(i-1)}$$

where α is the multiplicative updating factor. The partial derivatives of the criterion have the following interesting form:

$$\frac{\partial C_{\text{GSMM}}(\theta, \theta^{(i-1)})}{\partial \theta_j} = P - Q$$

where P and Q are both positive quantities. An appropriate direction of maximization is then found by setting α to $\frac{P}{Q}$ as in [15]. For each parameter in θ we derive the updating rules which we report in algorithm 1.

Additionally, one can note that updating the tensor of amplitudes $\mathbf{B}$ does not require the computation of the posterior probabilities and can be computed before each E step. We chose to update the other parameter matrices alternatively, namely one matrix of parameters for one M step. We arbitrarily adopted the following order: first $\mathbf{W}^{\Phi}$, then $\mathbf{H}^{M}$, $\mathbf{W}^{M}$ and $\mathbf{\Pi}$, then $\mathbf{W}^{\Phi}$ again and so forth. 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.

C. Multiplicative gradient method for IMM

For the IMM, since there are no hidden states, the criterion is directly chosen as the log-likelihood of the observations, for the parameter set $\theta = \theta_{\text{IMM}}$:

$$\begin{aligned} C_{\text{IMM}}(\theta) &= \log p_\theta(\mathbf{X}) \\ C_{\text{IMM}}(\theta) &= \sum_{f,n} \log \frac{|x_{fn}|}{\pi s_{X,fn}} - \frac{|x_{fn}|^2}{s_{X,fn}} \end{aligned} \tag{24}$$

The expression of the variance $s_{X,fn}$ in equation (24) is given by equation (16) and depends on θ. Here again, we use a multiplicative gradient method. The obtained updating rules are given in algorithm 2, where '$/$' and the divisions between matrices are meant element by element and 'T' as a superscript stands for matrix transposition. The power operations are element-wise.

As for the GSMM, and for the same reasons, we chose to update the parameters in the following order, for each iteration: first $\mathbf{H}^{F_0}$, $\mathbf{H}^{\Phi}$, $\mathbf{H}^{M}$, $\mathbf{W}^{\Phi}$ and $\mathbf{W}^{M}$.

IV. Main Melody Sequence Estimation

With the proposed models, the time dependency is not taken into account: each frame is independent from the other ones. The desired main melody is however expected to be rather smooth and regular, with respect to the energy of the instrument playing it as well as its frequency range and evolution. We also have to determine whether the main voice is present or not for each frame. We focus on these issues in this section.

A. Viterbi smoothing for the GSMM framework

In the probabilistic framework of the GSMM model, during the EM algorithm, we estimate the posterior probabilities $\mathrm{p}(k,u|\mathbf{x}_n)$ for each couple (k,u) and each frame n. In order to retrieve the desired melody, we use the posterior probability of the source state u for each frame: $p(u|\mathbf{x}_n) = \sum_k p(k,u|\mathbf{x}_n)$. A first strategy consists in taking the Maximum A Posteriori (MAP) for each frame. This leads to fairly good but noisy results. Instead, we propose an algorithm that smooths the melody line.

To model the regularity of the melody, we define a transition function which aims at penalizing transitions between notes

$$C_{\text{GSMM}}(\theta,\theta^{(i-1)}) = \sum_{n,k,u}\left[\sum_f\left(\log\frac{|x_{fn}|}{\pi s_{X,fn|ku}} - \frac{|x_{fn}|^2}{s_{X,fn|ku}}\right) + \log\pi_{ku}\right] p_{\theta^{(i-1)}}(k,u|\mathbf{x}_n) - \lambda\left(\sum_{k,u}\pi_{ku} - 1\right) + \text{CST} \tag{23}$$

Algorithm 1 EM algorithm for the GSMM: Estimating $\theta_{\text{GSMM}} = \{\mathbf{\Pi}, \mathbf{B}, \mathbf{W}^\Phi, \mathbf{H}^M, \mathbf{W}^M\}$

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

- $\forall k,u,n,\ \ b_{kun} \leftarrow b_{kun}\dfrac{P^B_{kun}}{Q^B_{kun}}$, where $\left\{\begin{array}{rcl} P^B_{kun} &=& \sum\limits_f \dfrac{w^\Phi_{fk}w^{F_0}_{fu}|x_{fn}|^2}{s^2_{X,fn|ku}} \\ Q^B_{kun} &=& \sum\limits_f \dfrac{w^\Phi_{fk}w^{F_0}_{fu}}{s_{X,fn|ku}} \end{array}\right.$

E step: thanks to (22), (15) and (14), compute $\gamma_n^{(i-1)}(k,u) = p_{\theta^{(i-1)}}(k,u|\mathbf{x}_n)$

$$\gamma_n^{(i-1)}(k,u) \propto p_{\theta^{(i-1)}}(\mathbf{x}_n|k,u)\pi_{ku}^{(i-1)}$$

where $p_{\theta^{(i-1)}}(\mathbf{x}_n|k,u)$ is given by Eq. (14) and (15).

M step: update the parameters (one sub-set of parameters per M step):

- $\forall f,k,\ \ w^\Phi_{fk} \leftarrow w^\Phi_{fk}\dfrac{P^\Phi_{fk}}{Q^\Phi_{fk}}$, where $\left\{\begin{array}{rcl} P^\Phi_{fk} &=& \sum\limits_{u,n}\gamma_n^{(i-1)}(k,u) \times \dfrac{b_{kun}w^{F_0}_{fu}|x_{fn}|^2}{s^2_{X,fn|ku}} \\ Q^\Phi_{fk} &=& \sum\limits_{u,n}\gamma_n^{(i-1)}(k,u)\dfrac{b_{kun}w^{F_0}_{fu}}{s_{X,fn|ku}(f)} \end{array}\right.$
- $\forall r,n,\ \ h^M_{rn} \leftarrow h^M_{rn}\dfrac{P^H_{rn}}{Q^H_{rn}}$, where $\left\{\begin{array}{rcl} P^H_{rn} &=& \sum\limits_{k,u,f}\gamma_n^{(i-1)}(k,u)\dfrac{w^M_{fr}|x_{fn}|^2}{s^2_{X,fn|ku}} \\ Q^H_{rn} &=& \sum\limits_{k,u,f}\gamma_n^{(i-1)}(k,u)\dfrac{w^M_{fr}}{s_{X,fn|ku}} \end{array}\right.$
- $\forall f,r,\ \ w^M_{fr} \leftarrow w^M_{fr}\dfrac{P^W_{fr}}{Q^W_{fr}}$, where $\left\{\begin{array}{rcl} P^W_{fr} &=& \sum\limits_{k,u,n}\gamma_n^{(i-1)}(k,u)\dfrac{h^M_{rn}|x_{fn}|^2}{s^2_{X,fn|ku}} \\ Q^W_{fr} &=& \sum\limits_{k,u,n}\gamma_n^{(i-1)}(k,u)\dfrac{h^M_{rn}}{s_{X,fn|k,u}} \end{array}\right.$
- $\forall k,u,\ \ \pi_{ku} \leftarrow \dfrac{1}{N}\sum\limits_n \gamma_n^{(i-1)}(k,u)$

end for

that are far apart. In the case of a singer, this is realistic, since singers often use glissandi when changing notes, yielding to almost continuous pitch changes in the melody. We chose a parametric penalization function, from state u_1 to u_2:

$$q(u_1,u_2) \propto \exp(-\beta\text{round}(|n_1 - n_2|))$$

where n_i is the MIDI code mapping[3] for the fundamental frequency number u_i, $i \in [1,2]$:

$$n_i = 12\log_2\left(\frac{\mathcal{F}(u_i)}{440}\right) + 69$$

440Hz is the frequency for A4 and 69 its MIDI code number. $\mathcal{F}(u_i)$ is the frequency in Hz corresponding to the source state u_i, i.e. the fundamental frequency of state u_i (see appendix I). β is a parameter arbitrarily set: it controls the trade-off between melody continuity (i.e. minimizing the distance between consecutive notes in pitch) and the "local" probability of the path (i.e. maximizing the posterior probabilities of the states on the path). Thereafter, to derive the Viterbi smoothing algorithm, we define a Hidden Markov Model (HMM) on the data as follows:

1) The observed signal is the signal STFT $\mathbf{X}$,
2) the sequence of hidden states is $\Xi = \{\xi(n) \in [1,U]; n \in [1,N]\}$ where the states are the possible notes $u \in [1,U]$,
3) the *a priori* distribution of those states is uniform, such that:
$$p_0(u) = \frac{1}{U}, \forall u \in [1,U]$$
4) the transition probabilities from state $\xi(n-1) = u_1$ to $\xi(n) = u_2$ are :
$$p(\xi(n) = u_2|\xi(n-1) = u_1) = q(u_1,u_2) \tag{25}$$

[3]This is a mapping and not a conversion, since the resulting n_i are real numbers, and not integers.

Algorithm 2 Updating rules for the IMM:
Estimating $\theta_{\text{IMM}} = \{\mathbf{W}^{\Phi}, \mathbf{H}^{\Phi}, \mathbf{H}^{F_0}, \mathbf{W}^{M}, \mathbf{H}^{M}\}$

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

- Vocal source parameters:

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

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

- Vocal filter parameters:

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

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

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

- Background music parameters:

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

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

end for

The desired sequence $\hat{\Xi}$ is such that the posterior probability of the whole sequence given the signal is the highest:

$$\hat{\Xi} = \arg\max_{\Xi} \text{p}(\Xi|\mathbf{X})$$

For the GMM, the EM algorithm directly outputs the $p(k, u|\mathbf{x}_n)$, from which we compute the $p(u|\mathbf{x}_n)$. These probabilities along with the penalization function q are the only inputs necessary for the Viterbi smoothing.

B. Viterbi smoothing in the IMM case

The previous Viterbi algorithm can be adapted to the IMM model, for which we however do not have the probabilities $\text{p}(u|\mathbf{x}_n)$. As we stated in section II, there is a link between the two models and the coefficients associated to the frequency u in the IMM, $h_{un}^{F_0}$, are ideally equal to zero if u is not active at frame n and proportional to the energy of the signal otherwise.

In practice, the amplitudes of these coefficients on one frame reflect whether the corresponding basis are present or not. They can therefore be considered as proportional to the posterior probability of the corresponding GMM: $h_{un}^{F_0} \propto p(u|\mathbf{x}_n)$. We compute a posterior "pseudo" distribution $p_{\text{IMM}}(u|\mathbf{x}_n)$ by normalizing the amplitudes $h_{un}^{F_0}$ over each frame n so that they sum to 1. The Viterbi algorithm is applied on this distribution matrix, with the same penalization function q as the GSMM, to obtain the desired regular melody line.

C. Silence Modeling

In the GSMM framework, it is possible to model silences in the main voice with a new state $Z_n =$ "silence" for which the spectrum is considered as null. The posterior probability of having a silent vocal part at frame n is denoted $\gamma_n(0) = \text{p}(\text{"silence"}|\mathbf{x}_n)$. The E step of algorithm 1 is modified to take into account this new state, for which the PSD of the vocal part, $\mathbf{s}_{V,n|\text{"silence"}}$ is fixed to 0. Both the estimation and the Viterbi algorithm can be done as explained in section III and IV.

For the IMM, 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.

V. Evaluation and results

A. Evaluation metrics and corpora

The proposed algorithms were evaluated with other systems at the MIREX 2008 Audio Melody Extraction task. The metrics that were used are the same as for the MIREX 2005 edition of the task, described in [16]. These metrics are framewise (as opposed to note-wise) measures: in this setting, the onsets and offsets of the different notes are not considered, only the fundamental frequency for a given frame is considered. 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 ISMIR04 database is composed of 20 songs and the MIREX05 dataset of 25 songs, both databases are described in [16]. For MIREX 2008, a new dataset (MIREX08) was also proposed, with 8 vocal Indian classical music excerpts[4]. 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. Also note that preliminary results for the IMM were published in [17].

[4]This subset is similar to the examples from http://www.ee.iitb.ac.in/daplab/MelodyExtraction/.

B. Algorithm behaviours: convergence and model

1) 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.

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 proposed model, 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. 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 I) 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, 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.

2) Convergence: In spite of the lack of formal convergence proof for the proposed iterative methods, according to our simulations and tests, the chosen criteria $C_{\text{GSMM}}(\theta, \theta')$ and $C_{\text{IMM}}(\theta)$ and, equivalently, the log-likelihood of the observation $\log p_\theta(\mathbf{X})$ 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. 5, 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.

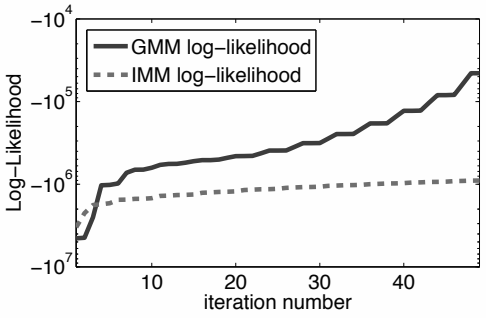


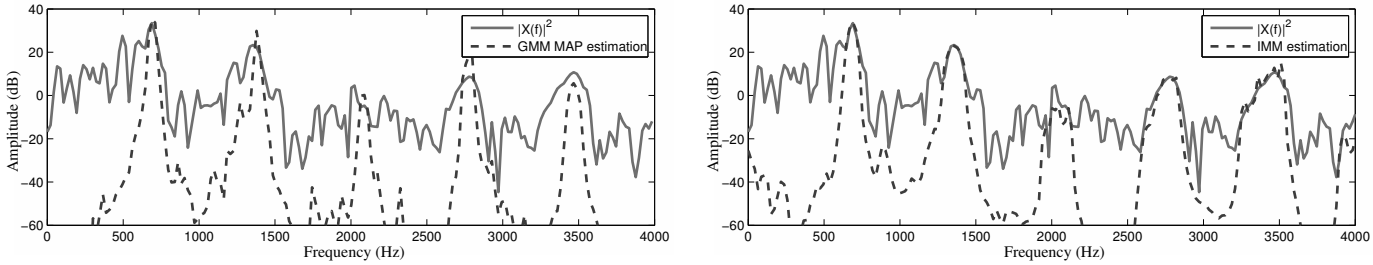
Fig. 5. Evolution of the log-likelihood of the observations for the GSMM and IMM algorithms.

3) Comparison between the proposed models: The IMM and GSMM algorithms lead to parameters that are really different. Theoretically, the main disadvantage of the 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 H_{F_0} reflect the polyphonic content of the music, including the accompaniment, which leads to the need for a melody tracker introduced in section IV.

However, it turns out, in certain circumstances, to be an advantage over the GSMM. Figure 6 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[5], at $t = 9.665$s. On the original spectrum, one can see the main "note", at around $f_0 = 680Hz$, among several other accompaniment notes. This frame actually corresponds to a "chirp", transition between two notes, 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 the GSMM and IMM 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 result does not fit the real data as closely as the IMM estimation does. This illustrates that the IMM can be a better model for vocal parts, especially on frequency transition frames (vibrato): on these segments, the GSMM assumption of having one stable fundamental frequency per frame does not hold.

The IMM could also be used for a polyphonic instrument, but its design as shown on the diagram figure 4 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 GMM diagram figure 3 by an instantaneous mixture. However, such a model leads to many more parameters to be estimated, hence to numerical problems and indeterminacies.

[5] http://www.music-ir.org/mirex/2008/



(a) GMM estimation result

(b) IMM estimation result

Fig. 6. "opera_fem4.wav": spectrum of a frame with a frequency chirp around $f_0 = 690Hz$ of the main melody, and the corresponding estimated spectra by the GSMM and IMM algorithms (derived in section III).

C. Main Melody Estimation Results

Table I provides the main results for the MIREX 2008 evaluation. The results for each of the different databases (ISMIR04, MIREX05 and MIREX08) are separately given. The "Total" column gives the average of these results, weighted by the number of files in the corresponding database.

The bold percentage show the best result for each column. We also provide the results of two other systems that were presented on the previous MIREX campaign in 2006. The proposed GSMM based system is denoted "drd1" and the IMM "drd2". The other systems "clly", "pc", "rk", "vr" are respectively described in [18], [19], [6] and [20].

On average over the 3 databases, the IMM (drd2) obtained the first 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 the IMM algorithm is robust to the variations of the database.

The GSMM, 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. The GSMM algorithm seems to perform quite well in certain favorable cases, such as the 2008 database. For this set, 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 the IMM, but up to a lesser scale than with the GSMM.

Globally, it is interesting to note that, 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 [21], 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 this year's campaign results.

At last, for both the GSMM and the IMM, it also seems that for 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.

D. Other applications of the Proposed Framework

1) Source Separation (De-Soloing) Performances: As in [22] or [23], where the transcription system in [6] is used as pre-processing for de-soloing of music signals, our framework is well designed for audio source separation. We adapted the IMM model in order to better fit the task at hand and also included a second parameter estimation step, which takes advantage of the estimated melody. The details of the implementation are given in [24]. On a database described on http://perso.telecom-paristech.fr/grichard/icassp09/, we obtain results comparable to [22] in terms of SDR [25]: 8.8dB of SDR gain for the separated main voice and 2.6dB of SDR gain for the accompaniment (see details in [24]). We encourage the interested reader to listen to the audio examples available on our website. Early results for the ISMIR 2004 and MIREX 2005 are also available on http://perso.telecom-paristech.fr/durrieu/en/results.html.

	ISMIR04		MIREX05		MIREX08		TOTAL	
System	Raw Pitch Acc.	Overall Acc.	Raw Pitch Acc.	Overall Acc.	Raw Pitch Acc.	Overall Acc.	Raw Pitch Acc.	Overall Acc.
clly1	75.3%	50.2%	68.9%	48.9%	54.7%	51.4%	69.2%	49.8%
clly2	75.3%	68.0%	68.9%	61.4%	54.7%	49.7%	69.2%	62.1%
drd1 (GMM)	65.9%	59.6%	57.4%	52.2%	85.8%	**76.0%**	64.9%	58.6%
drd2 (IMM)	**85.7%**	81.5%	**72.4%**	66.0%	81.8%	75.0%	**78.9%**	73.2%
pc	85.1%	**85.1%**	71.0%	**69.8%**	83.9%	73.3%	78.3%	**76.1%**
rk	82.4%	78.8%	69.7%	64.9%	83.5%	75.3%	77.3%	71.1%
vr	77.1%	70.1%	71.2%	63.5%	**88.2%**	66.7%	75.3%	67.1%
Average	78.1%	70.5%	68.5%	61.0%	76.1%	66.8%	73.3%	65.4%
Dressler	82.9%	82.5%	**77.7%**	**73.2%**				
Poliner	73.2%	71.9%	66.2%	63.0%				

TABLE I

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.

2) Multipitch Tracking: Multipitch tracking is a related task for which one desires to transcribe all the fundamental frequencies within each analysis frame of a polyphonic music signal. We combined the source separation abilities of our IMM model with its melody transcription to provide an iterative scheme for multipitch estimation.

Let J be the number of different sources or "streams" in the polyphonic signal. Let x_0 be the original mixture. For $j = 1 \dots J-1$, we estimate the main melody $\Xi^{(j)}$ on the residual signal x_{j-1} and generate x_j by removing the main voice thanks to the above source separation scheme. At $j = J$, we estimate one last time the melody, adapting the parameter estimation to bass note estimation, which needs better resolutions in the low frequency bins of the STFT.

Such a system was submitted to the MIREX 2008 Multiple Fundamental Frequency Estimation & Tracking task[6]. The results, with 49.5% of accuracy, are promising, achieving the 7th score out of the 15 participating system scores. This shows the potential of systems using source separation in order to reduce the complexity of a task and breaking it into several "easier" tasks, i.e. here transforming a polyphonic music transcription problem into several monophonic transcription ones.

VI. CONCLUSION AND FUTURE WORKS

We have proposed a system that transcribes the main melody from a polyphonic music piece. The method is based on source separation techniques and is closely related to Non-Negative Matrix Factorization (NMF). The main voice is characterized through a source/filter model. The melody sequence is constrained such that it achieves a trade-off between energetic predominance and smoothness, thanks to a Viterbi algorithm. The whole system is completely unsupervised.

The results in terms of accuracy for the framewise detection of the fundamental frequencies of the main melody show that our systems achieve performances at the state of the art. The proposed IMM model proved to be particularly robust to the diversity of the database. The GSMM model achieved top results on the 2008 dataset, which proves the validity of the model under certain circumstances, even if it does not seem robust enough against a strong polyphonic accompaniment.

Detailed analysis of the results for melody transcription as well as source separation results show that the chosen models do not seem able to separate one specific main source. The main part actually is the concatenation of all the sources that at given instants and during a long enough period have a predominant energy in the signal mixture. These mistaken segments are the consequence of the Viterbi algorithm, which sometimes misleads the system, as well as a lack of discrimination between the different instruments. On the other hand, the flexibility of the algorithm has the advantage of enabling separation and estimation of melodies played by a large range of instruments, such as the saxophone or the flute, as the results obtained on the MIREX databases show.

The proposed models can also be adapted to perform source separation, and more specifically main voice de-soloing. The results are promising, even if the main instrument model would need to be further improved to take into account other components of the signal such as unvoiced parts. Using the source separation ability, we could also design a multipitch extraction algorithm that obtained encouraging results and validated the approach consisting in dividing a complex problem into several other "easier" problems.

Future works are essentially related to source separation aspects and aim at modeling the main voice unvoiced parts, and extending the method in order to deal with reverberated signals, e.g. taking into accounts echoes in the main voice and removing it from the mixture during the de-soloing. The techniques introduced in this paper could also be extended to binaural signals, thus improving the results by taking advantage of inter-channel information. At last, a quantization step, both in time and in frequency, giving a more musical representation of the melody sequence should lead to a readable musical score. Such a representation may enable applications such as search by melodic similarities or cover version detection.

APPENDIX I
PARAMETRIC MODELING OF THE SOURCE SPECTRA DICTIONARY $\mathbf{W}^{F_0}$

We initiate 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

[6]http://www.music-ir.org/mirex/2008/index.php/...
...Multiple_Fundamental_Frequency_Estimation_&_Tracking

Hz). In our study, we consider the frequency range $[100, 800]$ Hz. We discretize this frequency axis such that there are 48 elements of the dictionary per octave:

$$\mathcal{F}(u) = 100 * 2^{\frac{u-1}{48}}$$

With these values, we obtain $U = 145$ available fundamental frequencies.

The source spectra are generated following a glottal source model: KLGLOTT88 [26]. 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: same window length, Fourier transform size and weighting window.

The original formula [26] is a continuous time function. To avoid aliasing when sampling that formula, we use the complex amplitude for all the harmonics of the signal up to the Nyquist frequency (about 5kHz in our application). Let c_h be the amplitude of the $h-$ th harmonic, $h \in [1, h_{\max}]$, we have [27]:

$$c_h = \mathcal{F}(u)\frac{27}{4}\left(\exp(-i2\pi hO_q) + 2\frac{1+2\exp(-i2\pi hO_q)}{i2\pi hO_q} - 6\frac{1-\exp(-i2\pi hO_q)}{(i2\pi hO_q)^2}\right)$$

where O_q is the "open quotient" parameter , which we fixed at $O_q = 0.5$. $e_u(t)$ 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 the squared magnitude of this Fourier transform: $w_{fu}^{F_0} = |E_u(f)|^2$, $\forall f \in [1, F]$.

ACKNOWLEDGMENT

The authors would like to thank the audio group of TELECOM ParisTech, especially R. Badeau, for the inspiring environment it provided during the elaboration of this work. The authors would also like to thank A. Ehmann for his help with evaluating our algorithms on the MIREX databases, and the team at IMIRSEL for their effort in preparing the MIREX evaluation campaigns, running all the submissions and gathering all the data to provide the high quality results that were partially presented in this paper. The authors are grateful to the anonymous reviewers whose comments greatly helped to improve the original manuscript.

REFERENCES

[1] M. Ryynänen and A. Klapuri, "Query by humming of midi and audio using locality sensitive hashing," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Las Vegas, Nevada, USA, Apr. 2008, pp. 2249–2252.

[2] G. Peeters, "Sequence representation of music structure using higher-order similarity matrix and maximum-likelihood approach," in *International Conference on Music Information Retrieval*, 2007.

[3] J. Serra, E. Gomez, P. Herrera, and X. Serra, "Chroma Binary Similarity and Local Alignment Applied to Cover Song Identification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 6, pp. 1138–1151, 2008.

[4] M. Goto, "Robust predominant-F 0 estimation method for real-time detection of melody and bass lines in CD recordings," *ICASSP IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, pp. 757–760, 2000.

[5] R. Paiva, "Melody detection in polyphonic audio," Ph.D. dissertation, University of Coimbra, 2007.

[6] M. P. Ryynänen and A. P. Klapuri, "Transcription of the singing melody in polyphonic music," *International Conference on Music Information Retrieval*, 2006.

[7] G. Poliner and D. Ellis, "A classification approach to melody transcription," *International Conference on Music Information Retrieval*, pp. 161–166, 2005.

[8] C. Sutton, E. Vincent, M. Plumbley, and J. Bello, "Transcription of vocal melodies using voice characteristics and algorithm fusion," *Music Information Retrieval Evaluation eXchange*, 2006.

[9] C. Févotte, N. Bertin, and J.-L. Durrieu, "Nonnegative matrix factorization with the itakura-saito divergence: With application to music analysis," *Neural Computation*, vol. 21, no. 3, pp. 793 – 830, March 2009.

[10] L. Benaroya, F. Bimbot, and R. Gribonval, "Audio source separation with a single sensor," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, pp. 191–199, 2006.

[11] 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*, vol. 15, no. 5, pp. 1564–1578, 2007.

[12] G. Fant, *Acoustic Theory of Speech Production*. Mouton De Gruyter, 1970.

[13] 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)*, vol. 39, pp. 1–38, 1977.

[14] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *NIPS*, 2000, pp. 556–562.

[15] T. Virtanen, "Monaural Sound Source Separation by Nonnegative Matrix Factorization With Temporal Continuity and Sparseness Criteria," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 3, pp. 1066–1074, 2007.

[16] 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, and Language Processing*, vol. 14, no. 4, pp. 1247–1256, 2007.

[17] 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*, 2008, pp. 169–172.

[18] C. Cao and M. Li, "Multiple f0 estimation in polyphonic music (mirex 2008)," *Music Information Retrieval Evaluation eXchange*, 2008.

[19] P. Cancela, "Tracking melody in polyphonic audio. mirex 2008," *Music Information Retrieval Evaluation eXchange*, 2008.

[20] V. Rao and P. Rao, "Melody extraction using harmonic matching," *Music Information Retrieval Evaluation eXchange*, 2008.

[21] K. Dressler, "Extraction of the Melody Pitch Contour from Polyphonic Audio," *Music Information Retrieval Evaluation eXchange*, 2005.

[22] 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*, pp. 1417–1420, 2008.

[23] T. Virtanen, A. Mesaros, and M. Ryynänen, "Combining pitch-based inference and non-negative spectrogram factorization in separating vocals from polyphonic music," in *ISCA Tutorial and Research Workshop on Statistical and Perceptual Audition*, Brisbane, Australia, Sept. 2008.

[24] J.-L. Durrieu, G. Richard, and B. David, "An iterative approach to monaural musical mixture de-soloing," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2009, pp. 105–108.

[25] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[26] D. Klatt and L. Klatt, "Analysis, synthesis, and perception of voice quality variations among female and male talkers," *Journal of the Acoustical Society of America*, vol. 87, no. 2, pp. 820–857, 1990.

[27] N. Henrich, "Etude de la source glottique en voix parlée et chantée," Ph.D. dissertation, Université de Paris 6, 2001.