A SOURCE/FILTER APPROACH TO AUDIO MELODY EXTRACTION

Jean-Louis Durrieu, Gaël Richard, Bertrand David
Institut Telecom; Telecom ParisTech; CNRS LTCI
firstname.lastname@telecom-paristech.fr

ABSTRACT

We propose the extension of a previously submitted system to address the Audio Melody Extraction task. The original system is based on the decomposition of the signal into a leading voice and an accompaniment. For this submission, the model for the accompaniment is as in the previous one : it relies on a Non-negative Matrix Factorization (NMF) of the spectrogram. The model for the leading voice is based on the original source/filter model, and includes some refinements such as an explicit smoothing parameterization of the filter parts, in order for the model to be closer to a natural sound production model.

Preliminary tests show that our systems obtain results comparable to the systems provided last year. The GSMM extension proposed in this paper may also benefit from some bug removal from last year's code. Although the proposed models do not lead to great improvements in the results, we believe the underlying semantics of the estimated parameters are easier to interprete and open this model for future indexing applications.

1. INTRODUCTION

Extracting the main melody from a polyphonic music signal can be defined as transcribing the notes that are played by an instrument which has to be somehow "dominating" other instruments from the mixture. This instrument can be in the foreground according to different cues, such as its energy or its frequency range.

During ISMIR 2004 and at MIREX 2005, 2006 and 2008, the evaluations for audio melody extraction showed there were several possible approaches to Audio Melody Extraction (AME) [1]. Most of them are perceptually based, and to a certain extend involving classifiers. However few works have been done that involve generative models for the observed signals. Our submissions at the previous evaluation campaign [2], mostly described in [3], introduced a source/filter model for the leading instrument, inspired by speech processing techniques, with a statistical model for the accompaniment which is equivalent to non-negative matrix factorisation (NMF) of the accompaniment spectrogram matrix. A Viterbi smoothing algorithm allows to control the continuity of the melodic line. Two different models for the main voice part were presented : one which is directly derived from the Gaussian scale mixture models (GSMM) of the source separation litterature, and one which is based on an instantaneous mixture of all the possible basis elements in a dictionary, called the instantaneous mixture model (IMM).

We further developed these models in order to include several important features, especially from a generative point of view : the proposed approach now include explicit smoothness of the filter part of the leading voice. The proposed extension for the GSMM has also been improved : last year's program seemed to have some bugs that should be fixed for this year.

This paper is organized as follows : first we introduce the models we consider for our submissions. The general principles for the estimation of the parameters are then discussed. At last, we give the results obtained on the development files for MIREX 2009 and comment the final results of the evaluation campaign.

2. SIGNAL MODEL

We propose two systems, each of which relying on the same source/filter model, with some differences that we briefly describe in this section.

We model the power spectrogram $\mathbf{S}^X$ of the mixture signal : it is assumed to be the instantaneous sum of 2 contributions, the leading voice $\mathbf{S}^V$ and the accompaniment $\mathbf{S}^M$:

$$\mathbf{S}^X = \mathbf{S}^V + \mathbf{S}^M \tag{1}$$

Under mild assumptions, this is equivalent to the statistical model we introduced in [3] : the (complex) short time Fourier transform (STFT) $\mathbf{X}$ of the mixture is the sum of leading voice $\mathbf{V}$ and the accompaniment $\mathbf{M}$. In particular, with complex Gaussian assumption on the variables $\mathbf{V}$ and $\mathbf{M}$, maximum likelihood estimation of the parameters is equivalent to minizing the Itakura-Saito divergence between the estimated $\mathbf{S}^V + \mathbf{S}^M$ and the observed $\mathbf{S}^X$, where $\mathbf{S}^X = |\mathbf{X}|^2$, $\mathbf{S}^V = |\mathbf{V}|^2$ and $\mathbf{S}^M = |\mathbf{M}|^2$, [4].

We first describe the instantaneous mixture model, and then the GSMM. For both we give the different parameterizations of $\mathbf{S}^V$ which allow us, after estimation, to retrieve the desired main melody. The model for $\mathbf{S}^M$, common both for the IMM and the GSMM, is then described.

2.1 Smooth-Instantaneous Mixture Model (SIMM) for the leading voice

The Instantaneous Mixture Model (IMM) presented in [3] has been improved by adding smoothness to the filter part of the leading voice. This new model was presented in [5].

We assume that the solo part is played by a monophonic and harmonic instrument, e.g. a human singer. We use the source/filter model proposed in [3], which is well adapted to this type of signal and we integrate an additional smoothness constraint on the filter frequency responses inspired by [6].

For $\mathbf{V}$, the speech-processing inspired source/filter model already obtained good results [7]. It allows the algorithm to seek for harmonic signals, thanks to a glottal source model on the source part, while still able to adapt the amplitudes of the different harmonics in the spectral comb through the filter shape estimation. It is parameterized as follows :

$$s_{fn}^{V} = s_{fn}^{\Phi} s_{fn}^{F_0}$$

with s_{fn}^{Φ} and $s_{fn}^{F_0}$ respectively the filter and the source contributions to the power sectrum. We denote the $F \times N$ matrices $\mathbf{S}^{\Phi}$ and $\mathbf{S}^{F_0}$ the matrices whose entries respectively are s_{fn}^{Φ} and $s_{fn}^{F_0}$.

The source part is modelled as a non-negative linear combination of the spectral combs of all the possible (allowed) fundamental frequencies. These spectra form a $F \times U$ matrix $\mathbf{W}^{F_0}$, where U is the number of possible notes. The associated amplitude coefficients form a $U \times N$ matrix $\mathbf{H}^{F_0}$ such that :

$$\mathbf{S}^{F_0} = \mathbf{W}^{F_0}\mathbf{H}^{F_0} \tag{2}$$

Similarly, we define the $F \times K$ filter dictionary matrix $\mathbf{W}^{\Phi}$, where K is the number of different filter shapes that are allowed. The activation coefficient for the resulting filters in $\mathbf{S}^{\Phi}$ form the $K \times N$ matrix $\mathbf{H}^{\Phi}$ such that $\mathbf{S}^{\Phi} = \mathbf{W}^{\Phi}\mathbf{H}^{\Phi}$. To model the smoothness of these filter frequency responses we introduce a $F \times P$ dictionary of smooth "atomic" elements $\mathbf{W}^{\Gamma}$. Each filter, i.e. each column vector from $\mathbf{W}^{\Phi}$, is then decomposed on this basis, as being a non-negative linear combination of the column vectors of $\mathbf{W}^{\Gamma}$. We then define the $P \times K$ matrix $\mathbf{H}^{\Gamma}$ such that : $\mathbf{W}^{\Phi} = \mathbf{W}^{\Gamma}\mathbf{H}^{\Gamma}$. By construction, each filter in $\mathbf{W}^{\Phi}$ is the sum of smooth functions, and is therefore also smooth.

$$\mathbf{S}^{\Phi} = \mathbf{W}^{\Phi}\mathbf{H}^{\Phi} = \mathbf{W}^{\Gamma}\mathbf{H}^{\Gamma}\mathbf{H}^{\Phi} \tag{3}$$

At last, $\mathbf{S}^{V}$ for the solo part is parameterized as follows, where '$\bullet$' represents element-wise product between the matrices :

$$\mathbf{S}^{V} = \mathbf{S}^{\Phi} \bullet \mathbf{S}^{F_0} = (\mathbf{W}^{\Gamma}\mathbf{H}^{\Gamma}\mathbf{H}^{\Phi}) \bullet (\mathbf{W}^{F_0}\mathbf{H}^{F_0}) \tag{4}$$

We will refer to this solo voice model (and by extension to the complete solo/accompaniment model) as the "Smooth-Instantaneous Mixture Model" (SIMM), in contrast with the "Gaussian Scaled Mixture Model" (GSMM) as used in [8]. Indeed, the source part can be seen, in its temporal counter-part, as the instantaneous mixture of all the possible notes, with amplitudes corresponding to the activation coefficients $\mathbf{H}^{F_0}$.

2.2 Smooth-Gaussian Scaled Mixture Model (SGSMM) for the leading voice

The above model is somewhat irrealistic, since it suggests that the source part is composed of all the possible sources active all at the same time. A more realistic model consists in allowing only one source and filter couple at a time. In order to do so, we need to use the underlying statistical model we introduced in [3].

We therefore define a framework with hidden variables representing the states of the filter and the source, namely a Gaussian Scaled Mixture Model (GSMM). The possible states are all the state souples $(k, u) \in [1, K] \times [1, U]$. Let Z_n be the state pair at frame n. Conditionally to $Z_n = (k, u)$, the likelihood of the leading voice STFT vector $\mathbf{v}_n$ is defined as :

$$\mathbf{v}_n | Z_n \sim \mathcal{N}_c(0, b_{kun}\text{diag}(\mathbf{w}_k^{\Phi} \bullet \mathbf{w}_u^{F_0})) \tag{5}$$

where $b_{kun} > 0$ is a scaling amplitude at frame n for the couple (k, u), $\mathbf{w}_k^{\Phi}$ and $\mathbf{w}_u^{F_0}$ respectively are the k^{th} column of $\mathbf{W}^{\Phi}$ and the u^{th} column of $\mathbf{W}^{F_0}$. Both these matrices are parameterized the same way as for the previous SIMM model.

The observation likelihood verifies :

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

$$\Leftrightarrow \qquad \mathbf{v}_n \sim \sum_{k,u} \pi_{ku} \mathcal{N}_c(0, b_{kun}\text{diag}(\mathbf{w}_k^{\Phi} \bullet \mathbf{w}_u^{F_0})) \tag{6}$$

where the prior probability of state $Z = (k, u)$ is denoted π_{ku}. For convenience, the conditional likelihoods $p\left(.|Z_n = (k, u)\right)$ are from here on abbreviated to $p(.|k, u)$.

The proposed SIMM was originally meant as an extension of the SGSMM. To estimate the parameters of the SGSMM, an EM algorithm is needed. The algorithm we developed is computationally quite intensive, and the SIMM was found to be a good trade-off between the computation load and the results, as shown by our experiment described in Section 4.

2.3 Accompaniment model

$\mathbf{S}^{M}$ is modelled as the sum of R elementary contributions (or sources), with distinct spectral shapes as in [4] :

$$\mathbf{S}^{M} = \mathbf{W}^{M}\mathbf{H}^{M} \tag{7}$$

where, as for the leading voice, $\mathbf{W}^{M}$ is the spectral shape matrix and $\mathbf{H}^{M}$ the corresponding amplitude matrix.

This generic model allows to fit a wide range of background sounds such as drums, guitars, bass as well as other classic music instrument.

3. PARAMETER ESTIMATION

In this section, we give the cost functions to be minimized for each model and some hints to derive the updating rules. The principles that are recalled here have been already discussed in [3–5] for the SIMM and [8, 9] for the SGSMM or related matters.

3.1 SIMM updating rules

In our maximum likelihood (ML) estimation framework, taking $-\log$ the complex Gaussian distribution for the mixture, we obtain a following cost function, thanks to Eq. (1), (4) and (7) :

$$C_{\text{SIMM}}(\Theta_{\text{SIMM}}) = \sum_{fn} \log(\hat{s}^X_{fn}) + \frac{s^X_{fn}}{\hat{s}^X_{fn}} \qquad (8)$$

where $\Theta_{\text{SIMM}} = \{\mathbf{H}^\Gamma, \mathbf{H}^\Phi, \mathbf{H}^{F_0}, \mathbf{W}^M, \mathbf{H}^M\}$ and $\hat{s}^X_{fn}$ is such that $\hat{\mathbf{S}}^X = (\mathbf{W}^\Gamma \mathbf{H}^\Gamma \mathbf{H}^\Phi) \bullet (\mathbf{W}^{F_0} \mathbf{H}^{F_0}) + \mathbf{W}^M \mathbf{H}^M$.

In order to find the updating rules, we calculate the partial derivatives of $C_{\text{SIMM}}(\Theta_{\text{SIMM}})$ with respect to each of the parameters. For a given parameter $\theta \in \Theta_{\text{SIMM}}$, the partial derivative have the interesting form $(\boldsymbol{\nabla}^+_\theta C_{\text{SIMM}} - \boldsymbol{\nabla}^-_\theta C_{\text{SIMM}})$, where $\boldsymbol{\nabla}^+_\theta C_{\text{SIMM}} > 0$ and $\boldsymbol{\nabla}^-_\theta C_{\text{SIMM}} > 0$. Using the following updating rule for θ makes θ to "move" towards some (non necessarily global) minimum of $C_{\text{SIMM}}(\Theta_{\text{SIMM}})$:

$$\theta \leftarrow \theta \times \frac{\boldsymbol{\nabla}^-_\theta C_{\text{SIMM}}}{\boldsymbol{\nabla}^+_\theta C_{\text{SIMM}}} \qquad (9)$$

Using such an approach for each of the parameters, and using matrix notations, we obtain the updating rules transcribed in Algorithm 1. The parameters are estimated matrix after matrix, in the following order : first $\mathbf{H}^{F_0}$, then $\mathbf{H}^\Phi$, $\mathbf{H}^M$, $\mathbf{H}^\Gamma$ and $\mathbf{W}^M$. This order may prevent the signal of interest, i.e. the leading voice, to be estimated within the accompaniment model.

3.2 SGSMM updating rules

Let $\Theta_{\text{SGSMM}} = \{\mathbf{H}^\Gamma, \mathbf{B}, \mathbf{W}^M, \mathbf{H}^M\}$ be the parameter set for the SGSMM, where $\mathbf{B}$ is the $K \times U \times N$ tensor whose entries are the b_{kun}. For the ML estimation of the parameters in the SGSMM framework, we iteratively minimize the expectation, under the current estimated parameters, $\Theta^{(i-1)}_{\text{SGSMM}}$ at iteration i, of $-\log$ the joint likelihood of the observation $\mathbf{X}$ and the hidden states $\{Z_n, n \in [1, N]\}$. The expression of the criterion is given in equation (10). The term "CST" is a constant independent from the parameter set Θ_{SGSMM}.

The parameter estimation for the SGSMM is given in Algorithm 2. The updating rule $\mathbf{B}$ does not depend on the *posterior* probabilities of the states, and $\mathbf{B}$ can therefore be computed at the beginning of every EM iteration. Then the *posterior* probabilities are computed (E-step). At each iteration, we update only one of the parameters set (M-step), in the following order : $\mathbf{H}^\Gamma$, $\mathbf{H}^M$ and $\mathbf{W}^M$.

Algorithm 1 Updating rules for the SIMM :
Estimating $\Theta_{\text{SIMM}} = \{\mathbf{H}^\Gamma, \mathbf{H}^\Phi, \mathbf{H}^{F_0}, \mathbf{W}^M, \mathbf{H}^M\}$

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

• Vocal source parameters :

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

$$\text{where} \begin{cases} \mathbf{P}^{F_0} &= \mathbf{S}^X \bullet (\mathbf{W}^\Phi \mathbf{H}^\Phi)/(\hat{\mathbf{S}}^X)^2 \\ \mathbf{Q}^{F_0} &= (\mathbf{W}^\Phi \mathbf{H}^\Phi)/\hat{\mathbf{S}}^X \end{cases}$$

• Vocal filter parameters :

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

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

$$\text{where} \begin{cases} \mathbf{P}^\Phi &= \mathbf{S}^X \bullet (\mathbf{W}^{F_0} \mathbf{H}^{F_0})/(\hat{\mathbf{S}}^X)^2 \\ \mathbf{Q}^\Phi &= (\mathbf{W}^{F_0} \mathbf{H}^{F_0})/\hat{\mathbf{S}}^X \end{cases}$$

• Background music parameters :

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

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

end for

3.3 Viterbi smoothing to estimate the melody line

Once the parameters are estimated, we use the same Viterbi smoothing algorithm as proposed in [3]. The transition $q(u_1, u_2)$ from the fundamental frequency number u_1 to frequency number u_2 is given by :

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

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

4. RESULTS

We report in this section some results we obtained on the development sets and we comment the official results from the MIREX 2009 evaluation campaign [10].

4.1 Development set

We have used all three available database in order to develop our algorithms. These sets are the ADC04 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 1. The fundamental frequency range was set for both algorithms to $[80, 800]$, with 4 pitches per semi-tone for the SGSMM and 8 for the SIMM. The reported SIMM results correspond to a system with $K = 2$ and $R = 100$, after 50 iterations. For the SGSMM algorithm, $K = 2$, $R = 20$, after 15 iterations. The IMM and

$$C_{\text{SGSMM}}(\Theta_{\text{SGSMM}}, \Theta_{\text{SGSMM}}^{(i-1)}) = \sum_{n,k,u} \left[\sum_f \left(\log \frac{|x_{fn}|}{\pi \hat{s}^X_{fn|ku}} - \frac{s^X_{fn}}{\hat{s}^X_{fn|ku}} \right) + \log \pi_{ku} \right] p_{\Theta^{(i-1)}_{\text{SGSMM}}}(k, u|\mathbf{x}_n) - \lambda \left(\sum_{k,u} \pi_{ku} - 1 \right) + \text{CST} \tag{10}$$

$$\text{where } \hat{s}^X_{fn|ku} = b_{kun} w^{\Phi}_{fk} w^{F_0}_{fu} + [\mathbf{W}^M \mathbf{H}^M]_{fn} \tag{11}$$

Algorithm	ADC04	MIREX05	MIR-1K
IMM (08)	0.86/0.82	0.72/0.66	
GSMM (08)	0.66/0.60	0.57/0.52	
SIMM	0.82/0.78	0.79/0.68	0.58/0.55
SGSMM	0.84/0.78	0.79/0.67	0.55/0.51

Table 1. Results of the tested algorithms, given for each development dataset, reported as "Raw pitch/Total Accuracy".

GSMM results 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 the SIMM and SGSMM.

The results of Table 1 show that both systems have quite similar results, which is what we woud have expected, since the SIMM system is an approximation of the ideal model provided by the SGSMM. The results on ADC04 and MIREX05 are of the same order as last year's performances [2]. The results for the SGSMM are much higher than those of the GSMM (drd1 at MIREX 2008 evaluation campaign), but it seems that the program that was provided last year had some numerical problems and returned aberrant results for some songs. Apart from this potential caveat, the results are slightly lower than the previous ones : the 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.

4.2 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 MIREX 2008 set (8 files), excerpts from the MIR-1K (374 files at different SNR conditions -5dB, 0dB, +5dB). For MIREX 2009, our submitted algorithms were denoted as "drd1" for the SGSMM and "drd2" for the SIMM.

The results obtained by this year's submissions are slightly under our last year's best submission, the 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 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 model also seems in general less adapted to the MIR-1K dataset. Further studies on the publicly available dataset may help to decipher the problem. Compared with the first proposed dataset, ADC04, as discussed on the evaluation campaign's wiki, the task corresponding more accurately to the MIR-1K dataset could be a more specific "singing melody extraction", 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 [8].

At last, our submissions seem to have better results on the vocal subsets, especially on the MIREX 2008 subset, for which the SGSMM obtained top results, and on the vocal pieces of the ADC04 and MIREX05 subsets. Non-vocal pieces across the datasets seem to 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 cases be identified with any of the instruments of the mixture.

5. CONCLUSION

We have proposed an extension to the systems submitted at the MIREX 2008 evaluation campaign, Audio Melody Extraction task. An explicit smoothness scheme has been added to the IMM and GSMM models. The results on the development and test sets show that our models are in general able to retrieve the desired melody line.

We still have to make some more tests in order to identify the key features of our models. Our goal is to obtain a generative model that is as realistic as possible, with which we can infer high level information such as notes or tonal key, and which at the same time tightly fits to the signal and enables low level applications such as source separation. We have for instance shown in [5] how our framework can be successfully used in leading instrument separation.

Algorithm 2 EM algorithm for the SGSMM : Estimating $\Theta = \Theta_{\text{SGSMM}} = \{\mathbf{B}, \mathbf{H}^{\Gamma}, \mathbf{H}^{M}, \mathbf{W}^{M}\}$

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

- $\forall k, u, n,\ \ b_{kun} \leftarrow b_{kun} \frac{P^B_{kun}}{Q^B_{kun}}$, where $\left\{ \begin{array}{rcl} P^B_{kun} & = & \sum\limits_f \frac{w^{\Phi}_{fk} w^{F_0}_{fu} s^X_{fn}}{(\hat{s}^X_{fn|ku})^2} \\ Q^B_{kun} & = & \sum\limits_f \frac{w^{\Phi}_{fk} w^{F_0}_{fu}}{\hat{s}^X_{fn|ku}} \end{array} \right.$

E step : compute $\gamma_n^{(i-1)}(k,u) = p_{\Theta^{(i-1)}}(k,u|\mathbf{x}_n)$

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

M step : update the parameters :

- $\forall p, k,\ \ h^{\Gamma}_{pk} \leftarrow h^{\Gamma}_{pk} \frac{w^{\Gamma}_{fp} P^{\Gamma}_{pk}}{w^{\Gamma}_{fp} Q^{\Gamma}_{pk}}$, where $\left\{ \begin{array}{rcl} P^{\Gamma}_{fk} & = & \sum\limits_{u,n} \gamma_n^{(i-1)}(k,u) \times \frac{b_{kun} w^{F_0}_{fu} s^X_{fn}}{(\hat{s}^X_{fn|ku})^2} \\ Q^{\Gamma}_{fk} & = & \sum\limits_{u,n} \gamma_n^{(i-1)}(k,u) \frac{b_{kun} w^{F_0}_{fu}}{\hat{s}^X_{fn|ku}} \end{array} \right.$
- $\forall r, n,\ \ h^{M}_{rn} \leftarrow h^{M}_{rn} \frac{P^{H}_{rn}}{Q^{H}_{rn}}$, where $\left\{ \begin{array}{rcl} P^{H}_{rn} & = & \sum\limits_{k,u,f} \gamma_n^{(i-1)}(k,u) \frac{w^{M}_{fr} s^X_{fn}}{(\hat{s}^X_{fn|ku})^2} \\ Q^{H}_{rn} & = & \sum\limits_{k,u,f} \gamma_n^{(i-1)}(k,u) \frac{w^{M}_{fr}}{\hat{s}^X_{fn|ku}} \end{array} \right.$
- $\forall f, r,\ \ w^{M}_{fr} \leftarrow w^{M}_{fr} \frac{P^{W}_{fr}}{Q^{W}_{fr}}$, where $\left\{ \begin{array}{rcl} P^{W}_{fr} & = & \sum\limits_{k,u,n} \gamma_n^{(i-1)}(k,u) \frac{h^{M}_{rn} s^X_{fn}}{(\hat{s}^X_{fn|ku})^2} \\ Q^{W}_{fr} & = & \sum\limits_{k,u,n} \gamma_n^{(i-1)}(k,u) \frac{h^{M}_{rn}}{\hat{s}^X_{fn|k,u}} \end{array} \right.$

end for

Ongoing studies focus on a description level that is higher than the physical fundamental frequency proposed so far [11]. Another direction is to integrate the smoothing of the melody line directly into the parameter estimation step, for instance by using priors on the parameters or a hidden Markov model (HMM) on the evolution of the states.

6. REFERENCES

[1] 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*, 14(4) :1247–1256, 2007.

[2] Music information retrieval evaluation exchange. online : http ://www.music-ir.org/mirex/2008/, September 2008.

[3] J.-L. Durrieu, G. Richard, and B. David. Singer melody extraction in polyphonic signals using source separation methods. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 169–172, Las Vegas, Nevada, USA, March 31-April 4 2008.

[4] 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) :793 – 830, March 2009.

[5] J.-L. Durrieu, A. Ozerov, C. Fevotte, G. Richard, and B. David. Main instrument separation from stereophonic audio signals using a source/filter model. In *European Signal Processing Conference*, Glasgow, Scotland, August 24-28 2009.

[6] E. Vincent, N. Bertin, and R. Badeau. Harmonic and inharmonic Nonnegative Matrix Factorization for Polyphonic Pitch transcription. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 109–112, Las Vegas, Nevada, USA, 2008.

[7] 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*, pages 105–108, Taipei, Taiwan, April 19-24 2009.

[8] 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.

[9] J.-L. Durrieu, G. Richard, B. David, and C. Févotte. Source/filter model for unsupervised main melody extraction from polyphonic music signals. *submitted to IEEE Transactions on Audio, Speech and Language Processing, special issue on Signal Models and Representations of Musical and Environmental Sounds*, 2010.

[10] Music information retrieval evaluation exchange. online : http ://www.music-ir.org/mirex/2009/, September 2009.

[11] 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 2009.