AN ITERATIVE APPROACH TO MONAURAL MUSICAL MIXTURE DE-SOLOING

Jean-Louis DURRIEU, Gaël RICHARD and Bertrand DAVID

Institut TELECOM, TELECOM ParisTech, CNRS LTCI
46, rue Barrault - 75634 Paris Cedex 13 - France
durrieu@enst.fr

ABSTRACT

In this article, we introduce a novel approach for monaural source separation with the specific aim to separate a polyphonic musical recording into two main sources: a main instrument (or melody) track and an accompaniment track. To that aim, we propose to model the power spectral densities (PSDs) of both contributions with a source/filter model for the main instrument while retaining a model emphasizing temporal repetitions of the musical background. We show that improved source separation performances can be obtained by a two-step estimation strategy where the model parameters are re-estimated in a second stage by adequately exploiting the main melody line estimated in a first stage. The experiments conducted on several monaural signal databases show that our system achieves state-of-the-art performances compared to other unsupervised source separation algorithms.

Index Terms— Music Information Retrieval (MIR), Blind Audio Source Separation (BASS), Melody extraction, Karaoke, Desoloing

1. INTRODUCTION

Blind audio source separation (BASS) has many useful applications, such as speech enhancement, karaoke or audio remixing. It is also gaining interest for a number of Music Information Retrieval (MIR) applications such as music and drum transcription since it can, as a pre-processing step, ease the indexing of complex polyphonic signals [1], [2]. The aim of BASS is to extract separated contributions (or sources) from a mixture by exploiting their differences in terms of spatial location and/or time-frequency (or timbral) content. It is common to categorize the source separation problem according to the difference between the number of sensors or channels and the number of desired contributions: it is called over-determined, determined and under-determined when the number of sensors respectively is larger than, equal to and smaller than the number of sources. This paper addresses the latter category, in the specific case where there is only one sensor and two desired contributions: the main (or solo) instrument and the accompaniment. This specific task is particularly interesting for applications such as karaoke, cover version discovery or author copyright protection.

A number of source separation approaches (see for example [3] and [4]) rely on supervised techniques to extract the vocal part from any other musical background. They introduce a statistical yet flexible framework. The sources are specified and classified by their spectral characteristics. They are then separated using a Wiener time-frequency mask. On the other hand, approaches like [1], [5] or [6] rely on sinusoidal models and unsupervised techniques to label several groups of sinusoids as belonging to either of the expected sources. The use of a sinusoidal model however significantly impairs the subjective quality of the results.

In this paper, we propose an algorithm that takes advantage from both approaches. We first estimate the melody of the predominant instrument, as in [6], then separate this instrument from the rest of the accompaniment thanks to the signal model adapted from [3]. Our algorithm is based on an improved version of the model proposed in [7], which was originally designed for MIR purposes. We show that state-of-the-art source separation performances are obtained by a two-step estimation strategy: first the pitch contour is estimated, then the parameters of our signal model are re-estimated by constraining the search space to the estimated melody.

This paper is organized as follows. We first recall our signal models which are based on a source/filter model for the main instrument and on a Non-negative Matrix Factorization (NMF) for the accompaniment part. The parameter estimation is also briefly described. In section 3, the complete source separation system based on the two-step strategy is described. The experiments and the results obtained are presented in section 4. We conclude and discuss future improvements for the system in the last section.

2. SIGNAL MODEL

The proposed signal model is adapted from the work of Benaroya [3] and is only briefly reviewed below since it was initially introduced in [7].

2.1. Mixture model

The observed musical signal x is composed of two main contributions: v the main instrument's voice and m the accompaniment. We assume that the mixture is instantaneous: $x = v + m$. Their Short-Time Fourier Transform (STFT) matrices, noted by capital letters, follow:

$$X = V + M$$

Let the centered, circularly wide-sense stationary (cwss), signal $y \in \{x, v, m\}$: its Fourier transform Y is distributed as a proper complex Gaussian, centered, with a diagonal covariance matrix. For a given frequency bin f, the corresponding element on the diagonal of this matrix $S_Y(f)$ is the power spectral density (PSD) of y at f.

We assume that V and M are independent. This implies:

$$S_X(f,n) = S_V(f,n) + S_M(f,n) \tag{1}$$

This work was partly realized as part of the Quaero Programme, funded by OSEO, French State agency for innovation, and of the network of excellence IST - K-SPACE, funded by the European Commission under the 6th Framework programme.

with f the frequency bin index and n the frame index. In order to be able to distinguish between the two sources, the PSDs of V and M are parameterized separately with different models as in [7]. In the following sections, we explain these differences.

2.2. Source/Filter Model of the Main Instrument

A generative source/filter model is used to represent the main instrument. Such a model is particularly well adapted for describing speech or singing voice production phenomena and since we mostly consider songs in this work, this model appears particularly relevant. However, due to its generality it is also, to a certain extent, adequate for other musical instruments.

In this model, the source part is strongly related to the pitch intended by the performer, while the filter part is mostly related to the timbre or the vowel sung. We assume that the desired instrument is harmonic and monophonic, and we only consider its voiced components.

The source spectrum $\sigma_{f_0}(f)$ is a "comb" with peaks at every multiple frequency of the fundamental frequency f_0. The amplitudes of the peaks are determined by the glottal source model KLGLOTT88 [8]. To model the timbral flexibility of the main instrument, the filter frequency response $w(f)$ acts as an envelope that reshapes the source comb in order to fit the analyzed signal.

We assume that the instrument produces only a limited number of notes, with a limited range of timbres: we set the fundamental frequency range to the interval $[F_{\text{min}}, F_{\text{max}}]$, with notes spaced every $\frac{1}{8}^{\text{th}}$ tone. The filters are limited to $K = 9$ possible envelopes $w_k(f)$, $1 \leq k \leq K$. Let these spectra form the normalized source dictionary W_{F_0} such that $W_{F_0}(f, f_0) = \sigma_{f_0}(f)$ and filter dictionary W_K, $W_K(f, k) = w_k(f)$. The smoothness of these filters is also simulated: each filter is modelled as a non-negative linear combination of atomic elements as in [9]. At a given frame n, the frequency domain representation of each part is a non-negative linear combination of the corresponding dictionary:

$$S_{F_0}(f, n) = \sum_{f_0} \sigma_{f_0}(f) H_{F_0}(f_0, n)$$

$$S_K(f, n) = \sum_{k} w_k(f) H_K(k, n)$$

where $S_{F_0}(f, n)$ and $S_K(f, n)$ respectively are the source and filter parts of the predominant instrument PSD at frame n and bin f, while $H_{F_0} > 0$ and $H_K > 0$ are the amplitude coefficient matrices. The columns of H_K are normalized, to avoid any ambiguity between H_{F_0} and H_K. For a monophonic instrument, there should be only one non-null coefficient per frame in H_{F_0}. However, the estimation in the first step is unconstrained and the monophonic main melody is tracked in a post-processing step as explained in section 3.1. We see in section 3.2 how the monophonic assumption is integrated in the second estimation. One can write the above equations as matrix products:

$$S_{F_0} = W_{F_0} H_{F_0} \qquad S_K = W_K H_K \tag{2}$$

At last, as proposed in [7], the PSD of the main instrument is identified with the Hadamard product (noted ".*") of S_{F_0} and S_K:

$$S_V = S_{F_0} . * S_K \tag{3}$$

2.3. Accompaniment Model

The accompaniment is modelled by an instantaneous mixture of R contributions. This model is similar to the one in [7] and leads to a NMF problem [10]. It emphasizes the repetition in time of the spectra of notes played by most musical instruments. Let W_R and H_R respectively be the accompaniment dictionary and the associated non-negative amplitude coefficient matrices. The music PSD S_M matrix is [10]:

$$S_M = W_R H_R \tag{4}$$

2.4. Parameter Estimation

Maximum Likelihood (ML) estimation of the proposed model parameters, under the Gaussian assumption, is equivalent to minimizing the Itakura-Saito divergence between the power spectrum $|X|^2$ of the observed STFT and the parameterized PSD S_X [10]. In addition, with equations (1),(2),(3) and (4), we have:

$$S_X = (W_{F_0} H_{F_0}) . * (W_K H_K) + W_R H_R \tag{5}$$

The five parameters in $\Theta = \{H_{F_0}, W_K, H_K, W_R, H_R\}$ are estimated by ML. We use multiplicative updates for all the parameters [7]. The source dictionary W_{F_0} is fixed and generated with the KLGLOTT88 model. The initial set of parameters Θ_0 is randomly drawn, except in the second step estimation, where we choose a specific initial H_{F_0} which is constrained to the estimated melody. Although the convergence is not proved, the resulting decomposition is satisfying as the chosen criterion globally decreases over the iterations.

2.5. Source Separation using Wiener Filters

The minimum mean squared error (MMSE) estimator of v knowing x is given by the Wiener estimator $\hat{v} = \mathrm{E}[v|x]$. In the class of estimators restricted to $\hat{v} = g_v * x$, where g_v is the impulse response of a linear filter, the MMSE estimator is obtained with the Wiener filter for which the frequency response is given by:

$$G_v(f) = \frac{S_{VX}(f)}{S_X(f)}$$

where S_{VX} is the inter-spectrum between v and x. With the independence assumption between v and m, we obtain: $S_{VX} = S_V$ and $S_X = S_V + S_M$, therefore:

$$G_v(f) = \frac{S_V(f)}{S_V(f) + S_M(f)} \tag{6}$$

We apply the filter in equation (6) in the frequency domain, in an adaptive way [3]: for each frame, we estimate the PSDs S_V and S_M with the parameter set Θ obtained by our algorithm and equations (3) and (4). We then compute the corresponding Wiener filters and $\hat{v}$ and $\hat{m}$ are reconstructed with the help of an overlap-add procedure.

3. COMPLETE SOURCE SEPARATION ALGORITHM

Figure 1 shows the outline of the complete blind audio source separation algorithm. It consists of two steps: the first one mainly aims at tracking the pitch contour (or melody) of the solo instrument. The second step estimates the parameters using the sequence of fundamental pitches estimated in the first step.

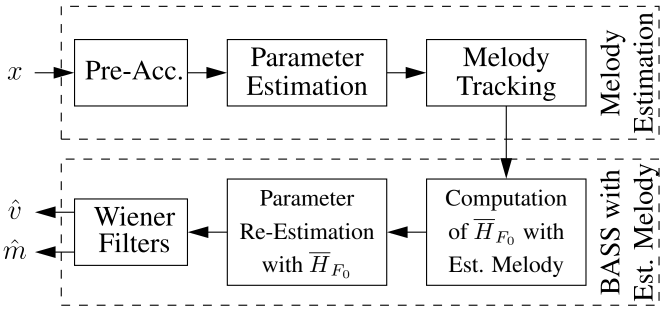

Fig. 1. Solo/Accompaniment Separation: algorithm outline.

3.1. Melody Estimation Step

Pre-accentuation: The mixture signal x is pre-accentuated with a conventional first order moving-average filter, with parameter $a = 0.95$.

Model parameter estimation: A set of parameters Θ_0 is randomly generated. At iteration $i \leq I$, Θ_{i-1} is updated to Θ_i thanks to the multiplicative updates in [7], with $I = 300$ the number of iterations.

Melody tracking: We use the Viterbi smoothing algorithm in [7] to retrieve the melody ϕ_0: $\hat{\phi}_0(n)$ is the estimated fundamental frequency of the solo instrument at frame n. It is important to note that our approach induces some ambiguities and this especially for the source model: the model as such allows the main instrument to be polyphonic, while we are interested in monophonic instruments. The smoothing step therefore has two goals: finding the smooth sequence of predominant pitches and limiting it to one pitch per frame. The algorithm in [7] is further improved by two new contributions: octave error reduction and silence detection.

To circumvent some octave errors, we use a modified “a posteriori probability” matrix G_{F_0} for the Viterbi algorithm: $G_{F_0}(f_0, n) = H_{F_0}(f_0, n) + 0.5H_{F_0}(2f_0, n)$. Our initial algorithm tends to favor the estimation of the pitch as being the upper octave instead of the fundamental frequency on some notes from our database. G_{F_0} is designed to compensate this effect.

To detect silences in the solo track, the separated solo is computed thanks to a Wiener filter masking (see section 3.2). The energy for each frame is computed and the frames with an energy lower than a given threshold are classified as silence frames of the solo. The threshold is chosen such that the energy of all remaining frames is above $(100 - \epsilon)\%$, where $\epsilon = 0.06$ in our system.

3.2. Source Separation Step

Computing $\mathbf{\overline{H}_{F_0}}$: The coefficients of H_{F_0} lying outside a scope of $\frac{1}{2}$ tone around the estimated melody are set to 0:

$$\begin{aligned}\overline{H}_{F_0}(f_0, n) &= H_{F_0}(f_0, n), \text{ if } |f_0 - \hat{\phi}_0(n)| < \frac{1}{4} \text{ tone}, \\ &= 0, \text{ otherwise.}\end{aligned}$$

Given this new matrix and the other parameters in Θ, we could possibly compute the separated signals $\hat{v}_{(1)}$ and $\hat{m}_{(1)}$. However, since we modified H_{F_0}, the estimated parameters are no longer optimal, especially for W_K, and a second estimation taking into account this new parameter matrix is necessary and improves the separation as shown by the results in section 4.

Parameter re-estimation: Again, Θ_0 is randomly drawn, except for the matrix H_{F_0} which is initially set to $\overline{H}_{F_0}$. Since we use multiplicative updates, the null coefficients in $\overline{H}_{F_0}$ do not evolve and stay null. The solo instrument is therefore limited to follow the estimated melody sequence $\hat{\phi}_0$ and the estimated parameters constrained to fit this melody (within $\frac{1}{2}$ tone).

Wiener filters: With the estimated final parameter set Θ, we obtain the separated signals $\hat{v}_{(2)}$ and $\hat{m}_{(2)}$. The pre-accentuation is compensated before comparison with the original sources.

4. EXPERIMENTS

4.1. Dataset description and Evaluation Criteria

Our database is composed of 3 subsets: (A) the SiSEC 2008 development set for the “professionally produced music recordings” separation task[1], (B) some songs from Ozerov and Lagrange’s private database ([4] and [1]) and (C) publicly available songs by S. Hurley, under Creative Commons licence. C is further divided into a pitch contour annotated set C1 and its complementary set C2. These files are described on our webpage[2]. The songs are split into one-minute-long excerpts, discarding the ones that have no solo instrument. The sampling rate is 11025Hz, the analysis window size is 512 samples and the hopsize 64 samples.

We use the objective evaluation criteria proposed for the SASSEC and SiSEC evaluation campaigns [11][1]: the Source to Distortion Ratio (SDR), the Source to Interference Ratio (SIR) and the Sources to Artifacts Ratio (SAR). The SDR/SIR gains are defined as in [4]: e.g. the SDR gain for the estimated source $\hat{v}$ is the difference between the SDR obtained by $\hat{v}$ and the SDR obtained by setting $\hat{v} = x$. These gains give the improvement in SDR/SIR of the estimated $\hat{v}$ compared to an algorithm which directly returns the mixture as the estimated separated source.

4.2. Melody Tracking Performance

In addition to the study done in [7], our submissions to the audio melody extraction task at MIREX 2008[3], scored 1st for the 2008 subset of the database and 2nd for the two other subsets. These results show that our system based on the source/filter parametric model reliably transcribes the desired melody.

On subset C1, the recall is at around 70%, while the precision only scores at 50%. This is explained by the fact that our system essentially focuses on energetic cues to track the melody line, and not on timbral cues. It therefore tracks the solo even if the solo instrument changes during the excerpt.

4.3. Source Separation with the True Pitch Contour

We verify that the model is able to separate the desired signals when the true pitch contour is given. We validate it on the melody-annotated subset C1. We separate the contributions by skipping the “melody estimation” step and use the annotated groundtruth of the melody pitch sequence ϕ_{GT} to initialize $\overline{H}_{F_0}$.

Table 1 summarizes the results of the proposed system given the pitch contour, named “Melody”, for these songs along with two other cases: “Mixture” characterizes the criteria computed by setting $\hat{v} = \hat{m} = x$, “Wiener” gives the results with the optimal Wiener filter computed with the original separated contributions. The first line therefore shows how difficult the task is and the last one gives

[1] Details and software available online at: http://sisec.wiki.irisa.fr/
[2] http://www.tsi.enst.fr/%7Egrichard/icassp09/
[3] http://www.music-ir.org/mirex/2008/

	Main Instrument			Accompaniment		
Method	SDR	SIR	SAR	SDR	SIR	SAR
Mixture	-6.2	-6.0	–	6.2	6.2	–
Melody	8.1	16.1	9.0	14.3	19.4	16.6
Wiener	11.9	21.3	12.8	15.5	26.5	16.6

Table 1. Evaluation criteria (in dB) for the method given the pitch contour on dataset C1.

	Main Instrument			Accompaniment		
Subset	SDR	SIR	SAR	SDR	SIR	SAR
All (1)	1.6	5.4	5.2	7.7	14.6	10.7
A (2)	8.2	17.4	8.9	10.8	15.4	12.9
B (2)	2.4	6.6	5.2	8.5	14.6	11.7
C (2)	2.7	9.2	5.0	9.1	14.1	12.7
C1 (2)	3.5	8.2	4.1	9.7	14.2	12.6
All (2)	2.7	8.1	5.2	8.8	14.4	12.1

Table 2. Evaluation criteria (in dB) for our global system averaged over each subset.

the theoretical performance limit.

For the main instrument as well as the accompaniment, the separation results are satisfying. Most of the interferences and artefacts correspond to the unvoiced part of the vocal part, which is not explicit in our model and therefore estimated as belonging to the accompaniment. Our approach is bounded in average by the result given in table 1, and the good performances there validate the proposed model.

4.4. Source Separation with Estimated Melody

Table 2 shows the results obtained by the proposed algorithm for each set of our database. The number into brackets indicates whether the separation is directly held after the melody extraction step (1) or after the second step (2). The mean "main instrument to accompaniment" ratio is -6.1dB. In average, the SDR/SIR gains obtained by the proposed iterative method on the database respectively are 8.8/13.8 for the solo voice and 2.6/8.0 for the accompaniment.

The figures in table 2 first show the improvement of our iterative approach (2) compared to the direct separation after the melody estimation (1). Informal listening tests confirm that the parameter reestimation really improves the quality and selectivity of the separation. It also seems that most of the interferences are due to estimated fundamental frequencies belonging to instruments of the "accompaniment", especially on the Celtic rock songs from subset B. In those songs, $\hat{v}$ often corresponds to musical instruments performing solos and not to the expected singer voice. It is worth noticing that our algorithm is designed to track the main melody without assuming timbre coherence. It is therefore possible to obtain a main solo track $\hat{v}$ played by different subsequent instruments.

In spite of these drawbacks, our results compare well with the state of the art. In [4], the authors report, for their supervised system, a SDR gain of 10.5dB for the separated voice, while our system obtains an average of 8.5dB SDR gain on the corresponding subset B, without the voice/music automatic segmentation as pre-processing and no learning step, since our approach is unsupervised. In [6], the separated accompaniment obtains a SDR gain average of 0.8dB at a -5dB "main instrument to accompaniment" ratio, while we obtain a SDR gain of 2.6dB. Some separation results are available on our webpage at http://www.tsi.enst.fr/%7Egrichard/icassp09/.

5. CONCLUSIONS AND FUTURE WORKS

We have proposed a single-channel solo-instrument / accompaniment separation algorithm based on the estimation of the main melody. The performance are very promising and are at the state of the art. Informal tests suggest that the artifacts of our system are less objectional than prior approaches. Contrary to source separation based on sinusoidal models, the choice of Wiener filters to separate the signals seems to significantly improve the quality of the results. The system could be further improved by taking into account the unvoiced parts of the main instrument or by extending this framework to multi-channel signals. A post-processing step to discriminate between the different solo instruments that were extracted would also help for source tracking and therefore for specific tasks such as singer/accompaniment separation.

6. ACKNOWLEDGEMENTS

The authors would like to thank R. Badeau and C. Févotte for the interesting discussions about the model, and A. Ozerov for kindly sharing his database with them.

7. REFERENCES

[1] M. Lagrange, L. G. Martins, J. Murdoch, and G. Tzanetakis, "Normalized cuts for predominant melodic source separation," *IEEE Trans. on ASLP*, 2008.

[2] O. Gillet and G. Richard, "Transcription and Separation of Drum Signals From Polyphonic Music," *IEEE Trans. on ASLP*, 2008.

[3] L. Benaroya, F. Bimbot, and R. Gribonval, "Audio source separation with a single sensor," *IEEE Trans. on ASLP*, 2006.

[4] 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 Trans. on ASLP*, 2007.

[5] Y. Li and D.L. Wang, "Separation of Singing Voice From Music Accompaniment for Monaural Recordings," *IEEE Trans. on ASLP*, 2007.

[6] M. Ryynanen, T. Virtanen, J. Paulus, and A. Klapuri, "Accompaniment separation and karaoke application based on automatic melody transcription," *IEEE ICME*, 2008.

[7] J.-L. Durrieu, G. Richard, and B. David, "Singer melody extraction in polyphonic signals using source separation methods," in *IEEE ICASSP*, 2008.

[8] D.H. Klatt and L.C. Klatt, "Analysis, synthesis, and perception of voice quality variations among female and male talkers," *JASA*, 1990.

[9] E. Vincent, N. Bertin, and R. Badeau, "Harmonic and inharmonic Nonnegative Matrix Factorization for Polyphonic Pitch transcription," *ICASSP*, 2008.

[10] C. Févotte, N. Bertin, and J.-L. Durrieu, "Nonnegative matrix factorization with the itakura-saito divergence: With application to music analysis," *to appear in Neural Computation*, 2008.

[11] E. Vincent, H. Sawada, P. Bofill, S. Makino, and J.P. Rosca, "First Stereo Audio Source Separation Evaluation Campaign: Data, Algorithms and Results," *Lecture Notes in Computer Science*.