SCORE INFORMED AUDIO SOURCE SEPARATION USING A PARAMETRIC MODEL OF NON-NEGATIVE SPECTROGRAM

Romain Hennequin, Bertrand David, and Roland Badeau

Institut Telecom, Telecom ParisTech, CNRS LTCI
46, rue Barrault - 75634 Paris Cedex 13 - France
<forename>.<surname>@telecom-paristech.fr

ABSTRACT

In this paper we present a new technique for monaural source separation in musical mixtures, which uses the knowledge of the musical score. This information is used to initialize an algorithm which computes a parametric decomposition of the spectrogram based on non-negative matrix factorization (NMF). This algorithm provides time-frequency masks which are used to separate the sources with Wiener filtering.

Index Terms— audio source separation, music information retrieval, machine learning, non-negative matrix factorization.

1. INTRODUCTION

Underdetermined audio source separation has been a major field of research for the past decades. When applied to musical signals, the goal is usually to obtain tracks corresponding to isolated instruments. Blind source separation was first addressed notably with non-negative matrix factorization (NMF) [1]. One of the main drawbacks of this technique is the difficulty to cluster the factorized elements and associate them with a source. This is certainly a reason why numerous works introduce additional information to improve separation results. Different kinds of information have been considered: in [2], the different spectral shapes of each source are learned on isolated sounds and are then used to decompose the mixture. In [3], source signals are used as a side information in a coder/decoder scheme.

Recently, the use of an aligned MIDI file to guide source separation was addressed in several works. In [4], stereo source separation based on spatial cues is improved by the knowledge of the score in order to accurately separate time-frequency bins with overlap. In [5], the score of the solo helps to separate it from the accompaniment with a classifier approach. In [6], P. Smaragdis proposed a probabilistic latent component analysis (PLCA) for isolating sounds in a mixture from the presentation of a humming query. This query mimics the desired target to be extracted and serves as a prior in the PLCA decomposition of the mixture. This approach is applied in [7], replacing the humming query by an aligned MIDI file and a synthesizer: each track of the MIDI file is synthesized and the provided signals are used as priors in the PLCA decomposition of the mixture. In [8], harmonic filters are generated from the score: in each analysis frame, the fundamental frequency of each active note in the MIDI file is finely assessed from the peaks in the spectrum.

In this paper, we propose a new approach for score informed source separation, based on a parametric decomposition of the power spectrogram of the mixture. The information extracted from the score is used to initialize the algorithm which provides the decomposition. The algorithm then locally optimizes the parameters (notably the fundamental frequency of each atom at each frame). The obtained decomposition provides a time-frequency mask for each source, which permits to separate the sources from the mixture by Wiener filtering.

The paper is organized as follows. An unusual non-negative framework which uses time-dependent frequency templates for decomposing audio spectrograms is first presented in section 2. We then derive an algorithm for score-informed source separation in section 3. A comparative evaluation of the performance of this algorithm is provided in section 4 where PLCA-based algorithms [7] are employed for benchmarking. Finally, we draw some conclusions and outline future work in section 5.

2. SPECTROGRAM MODEL

The model of power spectrogram is inspired by NMF but uses time-dependent parametric atoms (as presented in detail in [9]). These atoms are harmonic and we only consider the separation of a mixture of harmonic (or quasi-harmonic) instruments thus excluding percussive instruments.

2.1. Non-negative Matrix Factorization

Given an $F \times T$ non-negative matrix $\mathbf{V}$ and an integer R such that $FR + RT \ll FT$, NMF approximates $\mathbf{V}$ by the product $\hat{\mathbf{V}}$ of an $F \times R$ non-negative matrix $\mathbf{W}$ and an $R \times T$ non-negative matrix $\mathbf{H}$:

$$\forall f \in \{1, \dots F\}, t \in \{1, \dots T\} \quad [\mathbf{V}]_{ft} \approx [\hat{\mathbf{V}}]_{ft} = \sum_{r=1}^{R} w_{fr} h_{rt}. \tag{1}$$

When $\mathbf{V}$ is the magnitude or power spectrogram of a musical signal, the templates that are redundant in multiple frames are hopefully most of the time harmonic templates corresponding to musical tones. Thus, each column of $\mathbf{W}$ should correspond to a note and each row of $\mathbf{H}$ is the time activation associated with that note. However, this property is not guaranteed and generally, further constraints are added [10].

The approximation in (1) is generally quantified with an element-

The research leading to this paper was supported by the CONTINT program of the French National Research Agency (ANR), as a part of the DReaM project (ANR-09-CORD-006-03), and by the Quaero Programme, funded by OSEO, French State agency for innovation.

wise divergence to be minimized with respect to $\mathbf{H}$ and $\mathbf{W}$:

$$\mathcal{C}(\mathbf{W},\mathbf{H}) = D(\mathbf{V}||\hat{\mathbf{V}}) = \sum_{f=1}^{F}\sum_{t=1}^{T} d([\mathbf{V}]_{ft}, [\hat{\mathbf{V}}]_{ft}). \quad (2)$$

In this paper, the general class of β-divergence (see [9] for its expression) is considered and it is particularized to the Kullback-Leibler divergence (case $\beta = 1$) in our experiments.

2.2. Model of source spectrograms

As stated previously, the model of the power spectrogram (for each source of the mixture) considered in this paper is the parametric model presented in [9]: the spectrogram of a single instrument (a source) indexed by k is decomposed with parametric harmonic atoms (frequency templates). In opposition to NMF, these atoms can vary over time. Thus equation (1) is replaced by:

$$[\hat{\mathbf{V}}_k]_{ft} = \sum_{r=1}^{R} w_{kfr}^{f_0^{krt}} h_{krt}. \quad (3)$$

The time-dependence of a harmonic atom is based on the variation of its fundamental frequency f_0. This is emphasized by the t dependence in f_0^{krt}, which is the fundamental frequency of the rth atom of source k at time t: this parameterization permits to accurately model phenomena such as vibrato. The whole rth atom thus writes $w_{kfr}^{f_0^{krt}}$ where f denotes the frequency bin. These atoms are synthesized in the following way:

$$w_{kfr}^{f_0^{krt}} = \sum_{p=1}^{n_h} a_{kp} g(f - p f_0^{krt}). \quad (4)$$

Figure 1 represents such an harmonic atom. The function g corresponds to a single harmonic. Thus g is the squared modulus of the Fourier transform of the analysis window (used to compute the spectrogram). a_{kp} is the amplitude of the pth harmonic for every atom of source k: in order to limit the number of parameters in the model, we use the same set of harmonic amplitudes for every atom of a source. n_h is the number of harmonics. Each source has its own set of harmonic amplitudes and each atom of each source has its own time-varying fundamental frequency.

In this model, the following assumptions are made:

- one assumes that the harmonic part is stationary within a single frame (then the Fourier transform of an harmonic is the Fourier transform of the analysis window centered around the frequency of the harmonic);
- interferences between harmonics are assumed to be weak (this assumption is valid when the fundamental frequency of the atom is not too low);
- one assumes that interferences between peaks of negative frequencies (not taken into account) and positive frequencies are weak;
- one assumes that the frequency aliasing introduced by the signal sampling is weak (which allows using the analytical expression of the continuous Fourier transform of the analysis window).

As the cost function (2) to be minimized is multimodal with respect to the fundamental frequency of each atom, it is impossible to perform a global minimization of this function. It is thus necessary to introduce numerous atoms: in the proposed system, we chose to have an harmonic atom for each semitone of the chromatic scale (*i.e.* an atom per MIDI note). Atoms and activations can thus be indexed with a MIDI note number and the activations of a source can then be viewed as a piano roll.

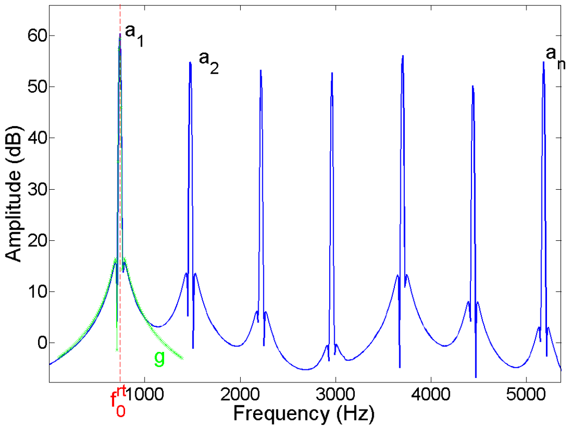

Fig. 1. Parametric atom defined in (4).

2.3. Model of the mixture spectrogram

The power spectrogram of the mixture is supposed to be the sum of the spectrograms of the sources (this is a common assumption with non-negative decompositions such as NMF which is generally explained by the approximate statistical independence of the sources). Thus the model of the mixture spectrogram is: $\mathbf{V}^{\text{mix}} \approx \hat{\mathbf{V}}^{\text{mix}} = \sum_{k=1}^{K} \hat{\mathbf{V}}_k$, where $\hat{\mathbf{V}}_k$ is the parametric spectrogram of source k (there are K sources) given in equation (3).

3. SCORE INFORMED SOURCE SEPARATION

The model of parametric spectrogram introduced in the previous sections is used to decompose the mixture spectrogram by means of a multiplicative descent algorithm which aims at minimizing (2), initialized with the information of the score. The score (in this paper a temporally aligned MIDI file) presumably permits to initialize the decomposition in a neighborhood of an optimal decomposition. In this paper, the MIDI file is supposed perfectly aligned with the mixture signal and thus we do not deal with the problem of MIDI alignment which can be done automatically [11].

3.1. Initialization with the musical score

The MIDI file provides a piano roll for each source. Each of these piano rolls has a direct link with the activations of the atoms of the corresponding sources. Thus, they can be used as an initializing mask for activations: while a note is active in the piano roll, the activation of the corresponding harmonic atom is set to 1. For all other instants, this activation is set to 0. Since we use a multiplicative algorithm (presented in section 3.2), coefficients initialized to 0 will remain 0 over iterations. Thus, it is better to slightly enlarge initialization to 1 before the beginning of a note and after the end of a note: this is necessary to avoid possible alignment errors and to take the possibly slow release of a note into account. An illustration of initializing activation masks is given in figure 2.

3.2. Algorithm

The algorithm resembles those used for standard NMF: the decomposition is obtained by minimizing a β-divergence between $\mathbf{V}^{\text{mix}}$ and

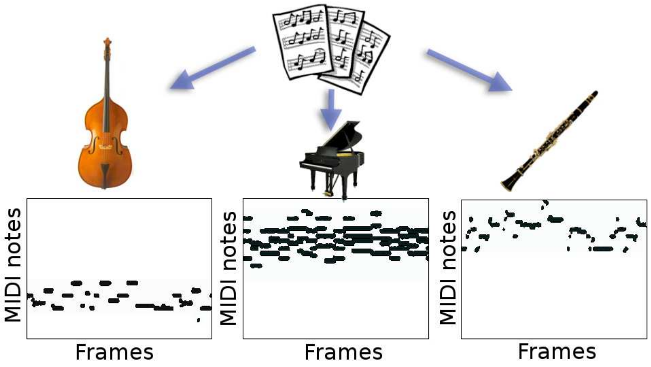

Fig. 2. Activation mask for the 3 instruments of a piece: activations of each source are initialized with a binary piano roll extracted from the corresponding MIDI track.

$\hat{\mathbf{V}}^{\text{mix}}$ with respect to the parameter of the model *i.e.* for each source:

- the fundamental frequency of each atom r at each instant t: f_0^{krt},
- the set of harmonic amplitudes a_{kp},
- the activation of each note at each instant h_{krt}.

The minimization is done with multiplicative update rules which are successively applied to each of the previous parameters. These rules particularly ensure that the parameters remain non-negative and become constant if the partial derivative of the cost function with respect to the considered parameter goes to zero.

These update rules can be straightforwardly derived from the mono-instrument (a single source) case presented in [9]. Thus we give them without providing the whole derivation. The update rules of parameters for each source k are given by:

$$
\begin{aligned}
f_0^{krt} &\leftarrow f_0^{krt}\frac{\mathcal{F}_{krt}}{\mathcal{G}_{krt}},\\
h_{krt} &\leftarrow h_{krt}\frac{\mathcal{M}_{krt}}{\mathcal{P}_{krt}},\\
a_{kp} &\leftarrow a_{kp}\frac{\mathcal{N}_{kp}}{\mathcal{Q}_{kp}},
\end{aligned}
$$

with:

$$
\begin{aligned}
\mathcal{G}_{krt} &= \sum_{f=1}^{F}\sum_{p=1}^{n_h} a_{kp}pP(f-pf_0^{krt})(\hat{V}_{ft}^{\text{mix}})^{\beta-2}(f\hat{V}_{ft}^{\text{mix}}+pf_0^{krt}V_{ft}^{\text{mix}}),\\
\mathcal{F}_{krt} &= \sum_{f=1}^{F}\sum_{p=1}^{n_h} a_{kp}pP(f-pf_0^{krt})(\hat{V}_{ft}^{\text{mix}})^{\beta-2}(pf_0^{krt}\hat{V}_{ft}^{\text{mix}}+fV_{ft}^{\text{mix}}),\\
\mathcal{P}_{krt} &= \sum_{f=1}^{F} w_{kfr}^{f_0^{krt}}(\hat{V}_{ft}^{\text{mix}})^{\beta-1},\\
\mathcal{M}_{krt} &= \sum_{f=1}^{F} w_{kfr}^{f_0^{krt}}(\hat{V}_{ft}^{\text{mix}})^{\beta-2}V_{ft}^{\text{mix}},\\
\mathcal{Q}_{kp} &= \sum_{f=1}^{F}\sum_{t=1}^{T}\sum_{r=1}^{R} g(f-pf_0^{krt})h_{krt}(\hat{V}_{ft}^{\text{mix}})^{\beta-1},\\
\mathcal{N}_{kp} &= \sum_{f=1}^{F}\sum_{t=1}^{T}\sum_{r=1}^{R} g(f-pf_0^{krt})h_{krt}(\hat{V}_{ft}^{\text{mix}})^{\beta-2}V_{ft}^{\text{mix}},
\end{aligned}
$$

where P is a non-negative function (see [9]) defined from the derivative of the Fourier transform of the analysis frame g as $P(f) = -\frac{g'(f)}{f}$.

4. RESULTS

The performance of the algorithm is assessed by means of experiments done on different synthetic databases, aligned on MIDI files. It is compared to that of an algorithm based on PLCA [7].

4.1. Description of the database

To our knowledge, no publicly available database exists which provides real-recorded mixtures of musical signals, real-recorded separated source signals and the corresponding aligned MIDI files. We thus designed an home-made database aiming at realistically rendering important characteristics of real musical streams, as the possibly important overlap of the sources in the time-frequency domain (in opposition to the randomly built database proposed in [7]), while providing separated tracks. For each piece, the tracks are summed to obtain the whole mix. The database is obtained from 12 MIDI files of royalty-free string quartets of Bach, Beethoven and Boccherini. These MIDI files are processed to synthesize wave files with 2 different methods. The first technique relies on isolated sounds recorded from real musical instruments with three levels of velocity. Their onset is synchronized with the `Note On` message and the sounds are then faded out at `Note Off`. The instrument list includes violin, viola and cello. This first technique will be referred to as "M1". The second technique is based on the TiMidity[1] software fed with a common soundfont set (Crisis General Midi 3.0[2]). The latter technique will be referred to as "M2". M2 includes a reverberation effect present in the soundfont whereas M1 does not.

It is worth noting that in this paper, we do not test the degradation of the performance of the algorithm with respect to alignment errors. The wave files of the database are downsampled to 11025Hz. The mixture signals are monophonic. The database (MIDI files, separated tracks and mixtures) is available at http://perso.enst.fr/hennequi/database.zip.

4.2. Experiment

The database provides the monophonic mixture and the aligned MIDI file for each piece. Only the 30 first seconds of each piece are processed. Mixture signals are separated using the information of the MIDI file with two algorithms:

- the algorithm that we presented,
- the algorithm based on PLCA presented in [7].

The PLCA-based algorithm requires a training stage where the MIDI tracks of the different files are synthesized and used as priors in the decomposition of the mixture. In order to make the comparison fair between algorithms, the synthesizer used to generate the training separated tracks should differ from the one we use to generate mixture signals. As sounds in our database are synthesized with two different methods, we use one for test and the other for training and vice versa. We also provide the results for the PLCA-based algorithm with the true sources used as priors (same training signals and test signals): they can be thought as an upper bound for the PLCA-algorithm performance, and also as a high-rank reference.

[1] http://timidity.sourceforge.net/

[2] http://www.bismutnetwork.com/10Music/Crisis/Soundfont3.0.php

For both algorithms, the mixture signals (and the separated signal tracks for the PLCA-based algorithm) were transformed into spectrograms using a short-time Fourier transform with 92ms-long Hann window, and 75% overlap.

The database provides the original separated sources, which are used to evaluate and compare the performance of both algorithms with the BSS_EVAL toolbox [12]: performance is assessed with signal to interference ratio (SIR), signal to artifact ratio (SAR) and signal to distortion ratio (SDR), all defined in [12]. The parameters of the PLCA-based algorithm (number of atoms and prior weights) were optimized (on a single piece) to give the best results possible.

4.3. Results

Results are given in tables 1 and 2. In both tables, the first row corresponds to the ratios obtained with our algorithm, the second row corresponds to the ratios obtained with the PLCA-based algorithm using different methods to synthesize training signals and test signals, and the third row corresponds to the ratios obtained with the PLCA-based algorithm using the true separated signals as training signals. Table 1 corresponds to the experiment using M1 as synthesizing method for the test signals and table 2 corresponds to the experiment using M2 as synthesizing method for the test signals. The given ratios are averaged over all pieces and all sources.

In table 1, the values of the ratios show that our algorithm outperforms the PLCA-based algorithm by more than 1dB in SAR and SDR and by 6dB in SIR, on our experimental database. Moreover, the ratios are not so far from the ones provided by the oracle-like PLCA-based algorithm. In table 2, the performance importantly deteriorates for both algorithms, probably due to the reverberation produced by M2. Our algorithm has slightly weaker results than the PLCA-based one.

It is worth noting that:

- the isolated samples of violin and viola from M1 contain important vibrato although there is only very slight vibrato in the sounds from M2. As our model explicitly takes vibrato into account, this phenomenon is more accurately modeled with our algorithm which probably partially explains the results of table 1;
- our approach is less supervised than the PLCA approach since it does not require to synthesize the tracks. The PLCA performance strongly depends on the synthesizer used to generate the training tracks: for instance the signals synthesized with M2 do not have the same decay times than the ones from M1 and have reverberation, probably resulting in incorrect activation priors in the PLCA.

Thus, our method performs similarly to the PLCA-based method but has the strength of being less supervised.

	SIR	SAR	SDR
Parametric decomposition algorithm	20.2	7.7	7.2
PLCA-based algo (training set: M2)	14.2	6.3	4.8
PLCA-based algo (training set: M1)	22.8	11.5	11.1

Table 1. Results on the signal synthesized with M1.

	SIR	SAR	SDR
Parametric decomposition	12.6	3.3	2.1
PLCA-based algo (training set: M1)	12.4	4.5	3.1
PLCA-based algo (training set: M2)	20.1	10.6	10.1

Table 2. Results on the signal synthesized with M2.

5. CONCLUSION

In this paper, we presented a new method to perform score informed source separation on a musical mixture. This method is based on a parametric model of non-negative spectrogram which uses harmonic atoms. We designed an evaluation database and our method reached a performance comparable to that of a PLCA-based algorithm while not requiring any priors.

As our model is limited to harmonic instruments, future works will include percussive instruments. Moreover, the spectrogram model can be further refined by, for instance, including other timbral parameters such as those proposed in [13]. The supervised learning of harmonic templates should also be explored.

6. REFERENCES

[1] T. Virtanen, “Monaural sound source separation by nonnegative matrix factorization with temporal continuity,” *IEEE TASLP*, vol. 15, no. 3, pp. 1066–1074, March 2007.

[2] P. Smaragdis, B. Raj, and M. Shashanka, “Supervised and semi-supervised separation of sounds from single-channel mixtures,” in *7th International Conference on ICA*, London, UK, September 2007.

[3] M. Parvaix, L. Girin, and J.-M. Brossier, “A watermarking-based method for informed source separation of audio signals with a single sensor,” *IEEE TASLP*, vol. 18, no. 6, pp. 1464–1475, August 2010.

[4] J. Woodruff, B. Pardo, and R. Dannenberg, “Remixing stereo music with score-informed source separation,” in *ISMIR*, Victoria, Canada, October 2006.

[5] C. Raphael, “A classifier-based approach to score-guided source separation of musical audio,” *Computer Music Journal*, vol. 32, no. 1, pp. 51–59, Spring 2008.

[6] P. Smaragdis and G. Mysore, “Separation by humming: User-guided sound extraction from monophonic mixtures,” in *WASPAA*, New Paltz, NY, USA, October 2009, pp. 69 – 72.

[7] J. Ganseman, P. Scheunders, G. Mysore, and J. Abel, “Evaluation of a score-informed source separation system,” in *ISMIR*, Utrecht, Netherlands, August 2010.

[8] M. Every and J. Szymanski, “A spectral-filtering approach to music signal separation,” in *DAFx*, Naples, Italy, October 2004, pp. 197–200.

[9] R. Hennequin, R. Badeau, and B. David, “Time-dependent parametric and harmonic templates in non-negative matrix factorization,” in *DAFx*, Graz, Austria, September 2010, pp. 246–253.

[10] N. Bertin, R. Badeau, and E. Vincent, “Enforcing harmonicity and smoothness in Bayesian non-negative matrix factorization applied to polyphonic music transcription,” *IEEE TASLP*, vol. 18, no. 3, pp. 538–549, February 2010.

[11] R. Dannenberg and N. Hu, “Polyphonic audio matching for score following and intelligent audio editors,” in *ICMC*, San Francisco, CA, USA, 2003, pp. 27–34.

[12] E. Vincent, R. Gribonval, and C. Févotte, “Performance measurement in blind audio source separation,” *IEEE TASLP*, vol. 14, no. 4, pp. 1462–1469, July 2006.

[13] R. Hennequin, R. Badeau, and B. David, “NMF with time-frequency activations to model nonstationary audio events,” *IEEE TASLP*, vol. 19, no. 4, pp. 744 – 753, May 2011.