SCORE INFORMED AUDIO SOURCE SEPARATION USING CONSTRAINED NONNEGATIVE MATRIX FACTORIZATION AND SCORE SYNTHESIS

Joachim Fritsch

Master ATIAM
University Pierre and Marie Curie
4 place Jussieu, 75005 Paris, France
joachim.fritsch@atiam.fr

Mark D. Plumbley

Centre for Digital Music
Queen Mary University of London
Mile End Road, London E1 4NS, UK
mark.plumbley@eecs.qmul.ac.uk

ABSTRACT

In this paper we present a new method for musical audio source separation, using the information from the musical score to supervise the decomposition process. An original framework using nonnegative matrix factorization (NMF) is presented, where the components are initially learnt on synthetic signals with temporal and harmonic constraints. A new dataset of multitrack recordings with manually aligned MIDI scores is created (TRIOS), and we compare our separation results with other methods from the literature using the BSS_EVAL and PEASS evaluation toolboxes. The results show a general improvement of the BSS_EVAL metrics for the various instrumental configurations used.

Index Terms— Audio source separation, musical score, nonnegative matrix factorization, multitrack dataset

1. INTRODUCTION

Musical audio source separation seeks to separate the signal of each instrument or musical source in a polyphonic mixture. Once separated, the sources can be processed separately and reassembled eventually, and so musical audio source separation can be used for music remastering, desoloing, denoising, etc. Many approaches have been addressed in the last two decades in order to achieve this separation. The most commonly used consists of decomposing a time-frequency representation of the mixture signal, with methods such as Nonnegative Matrix Factorization (NMF), Independent Component Analysis (ICA), Probabilistic Latent Component Analysis (PLCA), etc. Among these factorization techniques, NMF is probably the most popular for musical audio, as it describes the musical signal as a nonsubstractive combination of sound objects (or 'atoms') over time.

The first musical application of NMF has been automatic music transcription [1], followed by musical audio source separation [2]. In this latter work, the authors present a framework for Blind Source Separation (BSS) using NMF and time-frequency masking, where no prior musical information is provided to the separation algorithm. The extracted sounds thus need to be identified and grouped after the decomposition process, in order to recreate the estimated signals of the separated instruments. In [3], temporal continuity and sparseness criteria are incorporated within the NMF algorithm, in order to improve the detection and the isolation of musical sounds. But once again, the extracted components need to be associated with a specific source manually, which decreases dramatically the interest of the method in terms of automation and general performance.

In the last few years, the use of a symbolic representation of the musical signal (such as an aligned MIDI score) has been addressed to supervise the decomposition process. A musical score contains indeed a wide range of information, such as the pitch, the onset time and the duration of each note played by each instrument. This data can therefore be used to provide temporal and spectral information to the separation algorithm, and help improve its degree of performance as all the components are automatically assigned to an instrument. In [4], the MIDI score of each instrument is synthesized separately, and the components of a PLCA decomposition are learnt on these synthetic signals in a preliminary phase. Each instrument has a fixed number of atoms initialized randomly, and the data learnt is then used to initialize a second PLCA routine on the the actual mixture. In [5], the information from the score is used to initialize an algorithm based on a parametric decomposition of the spectrogram, using an original NMF framework. In [6], the separation is performed in real-time, with a score-follower using a hidden Markov model approach and a source separator extracting the harmonics of each instrument. In [7], the NMF decomposition is constrained by the information extracted from the score, and the basis function and the gain of each note are initialized with an harmonic comb and a binary function, respectively. In [8] finally, the solo voice from a song is represented by a source-filter model, and is extracted through a NMF algorithm with similar harmonic and temporal constraints generated from the score.

Work partly supported by EPSRC Leadership Fellowship EP/G007144/1 and EU FET-Open Project FP7-ICT-225913 "SMALL".

In this paper, we propose a new method for score informed source separation, combining ideas from the various approaches mentioned above. We learn the components of each instruments on synthetic signals in a preliminary phase, and we also use the temporal and harmonic information provided from the score to constrain a classic NMF algorithm.

The paper is organized as follows. In section 2 we describe our general framework for source separation, using constrained NMF and score synthesis. Then, in section 3, we assess our method with the usual evaluation toolboxes and we compare its separation results with two previous methods. Conclusion and perspectives for future work are presented in section 4. Finally, we present in an appendix a new dataset of multitrack recordings with aligned MIDI scores, created for the evaluation of score informed source separation methods or other applications.

2. DESCRIPTION OF THE METHOD

Our score informed source separation method is composed of two different phases, consisting of consecutive NMF routines. In the preliminary learning phase, the components of each instruments are learnt separately on signals synthesized from the score, and in the unmixing phase these components are then adapted to fit the actual instrumental mixture.

For each NMF routine, we use the β-divergence as a cost function and the Maximization-Minimization (MM) descent algorithm described in [9]. The important part of this algorithm is the initialization of the basis function and the activation coefficients of each component, as demonstrated bellow. A fully detailed and illustrated description of our score informed source separation method can be found in [10].

2.1. Score synthesis and preliminary learning phase

We initially synthesize the MIDI score of each instrument separately, in order to create a model for the signals of the different sources we intend to extract. This method has been introduced in [4], where a Dynamic Time Wrapping (DTW) technique was used to align the synthesized signals on the mix. In this paper we only consider perfectly aligned MIDI files, and so we do not deal with the problem of score-to-audio alignment or audio realignment by time wrapping.

The synthetic signals are then decomposed individually by a NMF routine using one component per note, which generates relevant models for the spectral basis and the amplitude of each note. The spectral bases of pitched instruments are initialized by harmonic combs, and those of percussive instruments are initialized by uniform distributions. We also add some extra-components to collect the residuals sounds from the synthetic signals of the pitched instruments.

This 'score synthesis' method has the advantage of being very easy to use, compared to what could be the implementation of analytic models for each register of each instrument.

2.2. Temporal and harmonic constraints

The information from the musical score is not only used to provide signal models by score synthesis as in [4], but also to incorporate temporal and harmonic constraints in the decomposition process which aims to factorize one note per atom.

After extracting the onset and offset times from the MIDI score of an instrument, we initialize the activation coefficients of each note by a simple binary function, equal to 1 if the note is being played and to 0 if not. This creates a 'pianoroll' representation of the score, used to initialize the gain matrix as in [5] and [7]. The advantage of this constraint is to enforce the coefficients initialized to 0 to remain to 0. The moments of silence will therefore remain silent, and only the coefficients initialized to 1 will fit the actual temporal envelopes of the corresponding notes. In practice, we enlarge slightly the initializations to 1 at the beginning and the end of each note, to avoid possible alignment errors or slow releases of notes.

As mentioned above, we initialize the spectral basis $w_{f,k}$ of each component k with a harmonic comb. This constraint is inspired from [7], and helps the NMF algorithm to segregate the notes being played simultaneously by differentiating them according to their harmonic structure. If we call N_h the number of harmonics of the model, g the magnitude spectrum of the analysis window and f_0^k the fundamental frequency of the corresponding note, we define the initialization of $w_{f,k}$ as

$$w_{0f,k} = \sum_{n=1}^{N_h} g(f - n f_0^k). \qquad (1)$$

2.3. Unmixing phase and general framework

Once the preliminary learning phase is over, we aggregate the spectral bases and the activation coefficients learnt in order to initialize a single NMF routine on the actual instrumental mixture. During this unmixing phase, the synthetic models are then adapted to fit the real world data. Again, we add some extra components with random initializations to collect the residuals sounds from the musical mixture in an additional source, such as impacts, blowing, clapping, plucking, etc.

Eventually, we use the factorization obtained from the unmixing phase to extract the signal of each estimated source, with the Wiener-filtering technique described in [9]. The general framework of our method is presented in Figure 1.

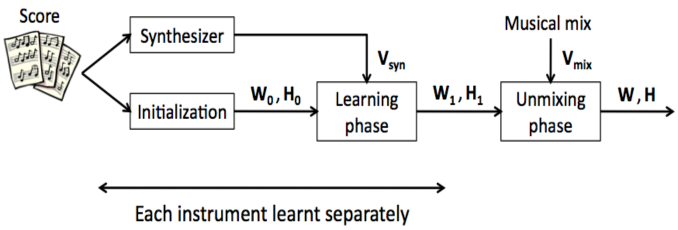

Fig. 1. General framework for the proposed method

		BSS_EVAL 3.0			PEASS 2.0			
Extract (dataset)	Method	SDR (dB)	SIR (dB)	SAR (dB)	OPS (%)	TPS (%)	IPS (%)	APS (%)
Quintet (MIREX)	Ganseman et al., 2010 [4]	6.64	12.44	10.34	23.73	**44.36**	29.51	**42.10**
	Hennequin et al., 2011 [5]	6.10	12.78	7.87	25.05	29.73	**48.94**	29.99
	Proposed	**10.07**	**16.62**	**11.39**	**27.08**	37.90	36.02	36.80
Quartet (Bach10)	Ganseman et al., 2010 [4]	5.82	9.92	9.06	9.61	21.69	10.47	43.32
	Hennequin et al., 2011 [5]	3.60	7.35	7.08	**22.83**	**31.39**	**30.19**	**46.38**
	Proposed	**7.34**	**11.88**	**10.04**	8.94	21.34	10.45	40.51
Mozart (TRIOS)	Ganseman et al., 2010 [4]	7.58	11.79	10.39	18.71	37.46	21.04	**49.50**
	Hennequin et al., 2011 [5]	8.94	15.42	10.23	**36.43**	**52.89**	**53.92**	48.22
	Proposed	**10.27**	**15.44**	**12.17**	18.36	34.53	22.57	26.04
Schubert (TRIOS)	Ganseman et al., 2010 [4]	4.46	9.76	7.24	18.90	38.89	18.17	46.19
	Hennequin et al., 2011 [5]	9.70	**15.36**	11.56	**34.91**	**47.93**	**49.20**	**47.81**
	Proposed	**10.20**	13.80	**12.89**	22.26	42.08	32.27	42.87
Brahms (TRIOS)	Ganseman et al., 2010 [4]	6.14	10.98	8.91	25.19	**57.89**	29.03	39.25
	Hennequin et al., 2011 [5]	5.48	11.60	8.37	**27.27**	36.04	**48.54**	40.06
	Proposed	**9.80**	**15.59**	**11.50**	22.58	37.67	30.00	**50.02**
Lussier (TRIOS)	Ganseman et al., 2010 [4]	7.70	12.09	9.99	26.62	44.30	31.74	**54.41**
	Hennequin et al., 2011 [5]	2.47	8.55	6.52	**28.32**	**45.46**	**53.87**	30.39
	Proposed	**9.10**	**15.13**	**10.87**	28.06	35.95	35.26	26.36
Take Five (TRIOS)	Ganseman et al., 2010 [4]	−2.65	2.48	5.22	24.05	**37.73**	28.31	24.29
	Hennequin et al., 2011 [5]	−3.87	1.51	4.87	19.57	27.56	**35.17**	10.94
	Proposed	**5.81**	**13.70**	**7.71**	**34.06**	20.90	32.99	**24.65**

Table 1. Separation results of our method compared with two previous methods from the literature. The three methods are applied to the same 10 seconds of each extract from the datasets used. We display the mean BSS_EVAL and PEASS metrics obtained for all the extracted sources of each musical extract. Higher is better for all scores, best scores are shown boldfaced.

3. EVALUATION OF THE METHOD

As already done in [11] on a smaller scale, we assess our source separation method with popular evaluation toolboxes, and we compare it with previous methods from the literature.

3.1. Evaluation metrics

We use the BSS_EVAL toolbox [12], where the performance measures are defined as energy ratios between the original sources, the extracted sources and the various estimated error terms. These energy ratios are the Signal to Distortion Ratio (SDR), the Signal to Interference Ratio (SIR) and the Signal to Artifacts Ratio (SAR), all expressed in dB.

We also use the PEASS toolbox [13], which provides similar performance measures but with also additional metrics in the form of perceptually-motivated scores rather than energy ratios. These scores are the Overall, Target-related, Interference-related and Artifacts-related Perceptual Scores (OPS, TPS, IPS and APS, respectively), all expressed in %.

3.2. Experimental setup

We apply our method to various musical extracts: a quintet from the MIREX 2007 dataset, a quartet from the Bach10 dataset [6], and trios from the TRIOS dataset (see appendix).

For the NMF routines, we use the Kullback-Leibler divergence (β=1) on the magnitude spectrogram with 15 iterations for the learning phase and 10 iterations for the unmixing phase, as this gives better results [10]. The spectrogram is computed with a 4096-point (93ms) Hann window and 87.5% overlap. For the pianorolls used for initialization, we add 0.1s before and 0.2s after each note, and for the unmixing phase we add 30 extra-components to collect the residual sounds. We evaluate our method on 10 seconds of each recording, on selected sections where all the instruments are playing.

3.3. Separation results

We compare our separation results with an updated version of [4] and with [5], both adapted to suit the datasets used[1]. The results are summarized in table 1, where the metrics of the extracted sources are averaged for each different extract.

From these results, we notice a general improvement of the BSS_EVAL metrics with our proposed method. This is probably due to the fact that the information from the score is exploited to its full potential, with the score synthesis and the constraints mentioned above. The PEASS metrics give better results with [5] though, so this enhancement still depends on

[1] The authors would like to thank Joachim Ganseman and Romain Hennequin for providing the code of their own source separation methods.

the various evaluation criteria and the way they are calculated.

We also compare the computation time of each method with the experimental setup described above (calculated on a macbook pro 2.26 GHz Intel Core 2 Duo, with 8GB memory). The results are presented in table 2. We notice that our method is slower than [4] but much faster than [5], certainly due to the relative simplicity of its NMF algorithm used in both phases.

	Computation time (s)		
Method	Trios	Quartet	Quintet
Ganseman et al., 2010 [4]	11	13	17
Hennequin et al., 2011 [5]	13066	6234	10296
Proposed	41	37	69

Table 2. Average computation time for the three methods depending on the number of instruments in the mixture.

The different datasets used for this experiment, the resulting extracted sounds and the code of our proposed method are all available through the C4DM Research Data Repository at `http://c4dm.eecs.qmul.ac.uk/rdr/`.

4. CONCLUSION

In this paper, we have presented an efficient method for score informed source separation. It factorizes one note per component in a NMF framework, by initially informing the features of these components with a constrained learning phase on separated signals synthesized from the score.

This method appeared to give good audio results and good performance measures in comparison with other methods from the literature. Future work could include the incorporation of a smoothness criteria for better perceptual results.

5. APPENDIX: THE TRIOS DATASET

The TRIOS dataset is a new score-aligned multitrack dataset that can be used for the evaluation of various tasks, such as score informed source separation, automatic music transcription, etc. This dataset consists of the separated tracks from five recordings of chamber music trio pieces, with their aligned MIDI scores. The five recordings are extracted from:

- a trio for clarinet, viola and piano by Mozart
- a trio for violin, cello and piano by Schubert
- a trio for violin, French horn and piano by Brahms
- a trio for trumpet, bassoon and piano by Lussier
- a version of ”Take Five” for alto sax, piano and drums

The separated tracks and the aligned MIDI scores are created as following. First, the original MIDI files are created or downloaded, and imported into a MIDI sequencer. The different tracks are then recorded separately, whilst the musicians listen to the other synthesized parts through headphones. The recordings are afterwards edited and mixed, and the MIDI scores are eventually manually aligned with Sonic Visualizer.

6. REFERENCES

[1] P. Smaragdis and J. C. Brown, “Non-negative matrix factorization for polyphonic music transcription,” *Signal Processing*, vol. 57, no. 3, pp. 177–180, 2003.

[2] B. Wang and M. D. Plumbley, “Musical audio stream separation by non-negative matrix factorization,” *Electronic Engineering*, vol. 38, no. 3, pp. 206–210, 2005.

[3] T. Virtanen, “Monaural sound source separation by non-negative matrix factorization with temporal continuity and sparseness criteria,” *IEEE Trans. Audio, Speech and Language Proc.*, vol. 15, no. 3, pp. 1066–1074, 2007.

[4] J. Ganseman, G. Mysore, P. Scheunders, and J. Abel, “Source separation by score synthesis,” in *Proc. ICMC*, New York, NY, USA, 2010.

[5] R. Hennequin, B. David, and R. Badeau, “Score informed audio source separation using a parametric model of non-negative spectrogram,” in *Proc. ICASSP*, Prague, Czech Republic, 2011, pp. 45–48.

[6] Z. Duan and B. Pardo, “Soundprism: An online system for score-informed source separation of music audio,” *J. Sel. Topics Signal Processing*, vol. 5, no. 6, pp. 1205–1215, 2011.

[7] S. Ewert and M. Müller, “Using score-informed constraints for NMF-based source separation,” in *Proc. ICASSP*, Kyoto, Japan, 2012, pp. 129–132.

[8] C. Joder and B. Schuller, “Score-informed leading voice separation from monaural audio,” in *Proc. ISMIR*, Porto, Portugal, 2012, pp. 277–282.

[9] C. Févotte and J. Idier, “Algorithms for nonnegative matrix factorization with the beta-divergence,” *Neural Computation*, vol. 23, no. 9, pp. 2421–2456, 2011.

[10] J. Fritsch, “High quality musical audio source separation,” M.S. thesis, UPMC/IRCAM/TELECOM ParisTech, Paris, France, 2012.

[11] J. Fritsch, J. Ganseman, and M. D. Plumbley, “A comparison of two different methods for score-informed source separation,” in *Proc. MML*, Edinburgh, Scotland, UK, 2012, pp. 11–12.

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

[13] V. Emiya, E. Vincent, N. Harlander, and V. Hohmann, “Subjective and objective quality assessment of audio source separation,” *IEEE Trans. Audio, Speech, and Language Proc.*, vol. 19, no. 7, pp. 2046–2057, 2011.