A Comparison of Two Different Methods for Score-Informed Source Separation

Joachim Fritsch JOACHIM.FRITSCH@ATIAM.FR
Master ATIAM, University Pierre and Marie Curie, 4 place Jussieu, 75005 Paris, France

Joachim Ganseman JOACHIM.GANSEMAN@UA.AC.BE
IBBT-Visionlab, University of Antwerp, Universiteitsplein 1, 2610 Antwerp, Belgium

Mark D. Plumbley MARK.PLUMBLEY@EECS.QMUL.AC.UK
Centre for Digital Music, Queen Mary University of London, Mile End Road, London E1 4NS, UK

Abstract

We present a new method for score-informed source separation, combining ideas from two previous approaches: one based on parametric modeling of the score which constrains the NMF updating process, the other based on PLCA that uses synthesized scores as prior probability distributions. We experimentally show improved separation results using the BSS_EVAL and PEASS toolkits, and discuss strengths and weaknesses compared with the previous PLCA-based approach.

1. Introduction

Musical audio source separation seeks to isolate the different instruments in a musical mixture. Many approaches have been proposed in order to conduct this separation, of which those using Nonnegative Matrix Factorization (NMF) and Probabilistic Latent Component Analysis (PLCA) have been shown to be effective.

More recently, the use of information from musical scores has been addressed to guide these algorithms and to improve the quality of separation. (Ganseman et al., 2010) aligned the synthesized score to the original audio, providing priors to the PLCA decomposition of the mixture. (Hennequin et al., 2011) used the score to initialize an algorithm based on a parametric decomposition of the spectrogram with NMF.

Work partly supported by an IWT Flanders Specialization Grant, EPSRC Leadership Fellowship EP/G007144/1, and EU FET-Open Project FP7-ICT-225913 "SMALL".
In *5th International Workshop on Machine Learning and Music*, Edinburgh, Scotland, UK, 2012.

In this paper we adapt and combine both methods, by synthesizing the score and learning the components of the different instruments separately, and then using the information learnt to initialize the decomposition process with NMF. Our proposed method is presented as 'Method A' and compared with an updated version of (Ganseman et al., 2010), presented as 'Method B'.

2. Score-Informed Source Separation

A musical score provides a wide range of information, such as the pitch, the onset time and the duration of each note played by each instrument. This information can therefore be used to supply spectral and temporal information to the separation algorithm. In this paper we use a perfectly aligned MIDI file and we do not consider the problem of score-to-audio alignment.

2.1. Description of Method A

As in (Ganseman et al., 2010), we initially learn the dictionaries and the activation coefficients of each instrument separately with the synthesized scores.

We use the Itakura-Saito (IS) NMF of the power spectrogram with multiplicative updates (Févotte, 2010), and we initialize the activation coefficients with a 'pianoroll' representation of the score, with one component per note. We also use a harmonic model to initialize the dictionaries, with a harmonic comb adapted to the fundamental frequency of each note (Hennequin et al., 2011). We add about 10 extra components with random initializations, to collect the residual sounds.

We use the information learnt to initialize a second IS-NMF routine on the actual musical mix, adding again about 30 extra components to collect the residual sounds. Finally, we separate the different instru-

		BSS_EVAL 3.0			PEASS 2.0			
	method	SDR (dB)	SIR (dB)	SAR (dB)	OPS (%)	TPS (%)	IPS (%)	APS (%)
bassoon	A-10	**11.97**$_{\pm 0.01}$	**20.43**$_{\pm 0.01}$	**12.67**$_{\pm 0.01}$	27.39$_{\pm 0.09}$	32.27$_{\pm 0.58}$	**62.16**$_{\pm 0.37}$	27.42$_{\pm 0.31}$
	B-20	10.68$_{\pm 0.32}$	18.82$_{\pm 0.75}$	11.47$_{\pm 0.30}$	**33.08**$_{\pm 0.95}$	**35.06**$_{\pm 2.10}$	57.98$_{\pm 1.75}$	**37.76**$_{\pm 1.47}$
clarinet	A-10	**14.45**$_{\pm 0.01}$	**23.46**$_{\pm 0.02}$	**15.06**$_{\pm 0.01}$	**26.33**$_{\pm 0.49}$	**41.61**$_{\pm 0.35}$	**33.04**$_{\pm 0.59}$	**30.58**$_{\pm 0.42}$
	B-20	11.92$_{\pm 0.48}$	17.42$_{\pm 0.74}$	13.45$_{\pm 0.48}$	14.81$_{\pm 1.85}$	25.74$_{\pm 2.95}$	14.93$_{\pm 2.29}$	25.67$_{\pm 2.89}$
flute	A-10	**16.51**$_{\pm 0.00}$	**22.11**$_{\pm 0.01}$	**17.94**$_{\pm 0.01}$	**37.41**$_{\pm 0.22}$	**35.68**$_{\pm 0.71}$	**60.41**$_{\pm 0.72}$	30.77$_{\pm 0.66}$
	B-20	12.49$_{\pm 0.57}$	21.86$_{\pm 0.55}$	13.05$_{\pm 0.59}$	32.18$_{\pm 1.11}$	31.56$_{\pm 2.25}$	51.91$_{\pm 2.92}$	**33.94**$_{\pm 2.19}$
horn	A-10	**11.10**$_{\pm 0.01}$	**20.96**$_{\pm 0.02}$	**11.61**$_{\pm 0.02}$	**37.76**$_{\pm 0.29}$	49.30$_{\pm 0.33}$	**47.84**$_{\pm 0.52}$	49.10$_{\pm 0.24}$
	B-20	5.03$_{\pm 0.45}$	8.27$_{\pm 0.65}$	8.44$_{\pm 0.23}$	11.41$_{\pm 0.79}$	**62.71**$_{\pm 4.90}$	2.87$_{\pm 0.81}$	**66.85**$_{\pm 4.10}$
oboe	A-10	**7.93**$_{\pm 0.01}$	**17.60**$_{\pm 0.02}$	**8.50**$_{\pm 0.01}$	**26.58**$_{\pm 0.30}$	40.50$_{\pm 0.27}$	**33.95**$_{\pm 0.41}$	17.83$_{\pm 0.31}$
	B-20	−0.52$_{\pm 1.07}$	1.92$_{\pm 1.54}$	5.46$_{\pm 0.61}$	25.03$_{\pm 1.18}$	**42.01**$_{\pm 2.63}$	16.37$_{\pm 1.71}$	**57.50**$_{\pm 1.01}$

Table 1. Quality of source separation results of a woodwind quintet. We display mean BSS_EVAL and PEASS metrics calculated over 100 runs, with standard deviation shown in subscript. Method A was run with 10 iterations and method B with 20 iterations in the mixture factorization phase. Higher is better for all scores, best scores are shown boldfaced.

ments with a Wiener masking method (Févotte, 2010).

2.2. Description of Method B

This method (Ganseman et al., 2010) only uses synthesized score parts to learn the dictionary and activation matrices that serve as prior distributions to PLCA. PLCA has been shown to be numerically equivalent to NMF with a Kullback-Leibler divergence. The method does not rely on any MIDI representation, so in the following experiment we apply it with a fixed number of 20 components per source on the magnitude spectrogram. Not anticipating a 6th source, we also do not provide additional components. To allow a fairer comparison, we altered the reconstruction phase to also use the normalized source estimates as a mask on the mixture spectrogram, i.e. Wiener filtering.

3. Results and Conclusion

We apply both methods to the first 15 seconds of the woodwind quintet recording from the MIREX 2007 F0-tracking competition. A 4096-point STFT with 87.5% overlap was used. Scores were synthesized using the EIC2 synthesizer integrated in Ableton Live, and the matrices for initialization (Method A) or prior distributions (Method B) were learnt from those in 30 iterations. Afterwards Method A was run for 10 iterations and Method B for 20 iterations, as this gave good results for each.

We use the BSS_EVAL (Vincent et al., 2006) and PEASS (Emiya et al., 2011) toolboxes for evaluation. The results of our experiment are summarized in table 1. We find that in this example, Method A gives overall better results, due to the harmonic and temporal constraints that Method A incorporates in the update process. The lack of those is likely the cause of Method B to have worse interference-related metrics (SIR, IPS), having more leakage from other sources into the extracted sounds. From the standard deviation measurement, we also notice that Method A has a more stable behavior than Method B. The dataset used for the experiment, the code and the resulting sound files are available through the C4DM Research Data Repository at `http://c4dm.eecs.qmul.ac.uk/rdr/`.

In the future, Method B could be improved by incorporating harmonic and temporal constraints similar to those from Method A. The parametric model of this latter would also need adjustments in the case of inharmonic or percussive sounds.

References

Emiya, V., Vincent, E., Harlander, N., and Hohmann, V. Subjective and objective quality assessment of audio source separation. *IEEE Trans. Audio Speech Lang. Proc.*, 19(7):2046–2057, 2011.

Févotte, C. Itakura-Saito nonnegative factorizations of the power spectrogram for music signal decomposition. In Wang, Wenwu (ed.), *Machine Audition*, chapter 11. IGI Global Press, 2010.

Ganseman, J., Mysore, G., Scheunders, P., and Abel, J. Source separation by score synthesis. In *Proc. ICMC*, pp. 462–465, New York, USA, 2010.

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

Vincent, E., Gribonval, R., and Févotte, C. Performance measurement in blind audio source separation. *IEEE Trans. Audio Speech Lang. Proc.*, 14(4): 1462–1469, 2006.