MUSICAL AUDIO STREAM SEPARATION BY NON-NEGATIVE MATRIX FACTORIZATION

Beiming Wang
Queen Mary, University of London
Department of Electronic Engineering

Mark D. Plumbley
Queen Mary, University of London
Department of Electronic Engineering

ABSTRACT

Our research is to develop a methodology for separating musical audio into streams of individual sound sources, such as instruments or voice. In this paper, we show the current progress of our research and a system built on the Non-negative Matrix Factorization (NMF) algorithm. The system was tested on both artificially mixed audio and real musical recording. This work is closely related to the task of blind source separation, computational auditory scene analysis and automatic music transcription. It will contribute to the areas such as music information retrieval, digital audio effects, and musical audio coding.
Keywords – Blind Source Separation, Non-negative Matrix Factorization, Automatic Music Transcription

1. INTRODUCTION

Humans have the ability to perceive multiple numbers of separate signals in certain environments consisting of different sources. Sometimes those individual signals can also be located, denoised or recognised successfully. This ability helps us to retrieve the information even it is not in its original form. For instance, we can recognize the singer's voice from the accompanied music, or read words embedded in a picture. For an audio signal, Bregman [4] called this *auditory scene analysis*. The practical realization of this problem by building a certain computer model is known as *computational auditory scene analysis* (CASA) [8]. Typically, we have no prior knowledge about the sources and how they are mixed. We extract or separate the required information only based on the given mixtures, so this is also a Blind Source Separation (BSS) problem.

The BSS problem exists widely in many fields such as audio and video signal processing, biomedical engineering, econometrics, and data mining. Focusing on audio processing, many applications that can benefit from the solution of BSS include automatic speech recognition, speech enhancement, automatic music transcription, music information retrieval, audio stream re-mixing and so on. Therefore, it has attracted a great deal of attention in the recent decade and plenty of methods have emerged [7], [13], [15]. There are mainly two crucial features in methods developed so far. One is the relation between the number of the given observed mixtures (which are also known as sensors) and the number of sources, and the other one is the way how source signals are mixed. In terms of the relation between the number of source signals, mixtures can be overdetermined, determined or underdetermined [9]. Certain methods require a determined mixture, such as most Independent Component Analysis algorithms [7], [3], [5], [2]. Since it is not always possible to know how many sources are present in our observation and acquire an equal number of sensors, separation of underdetermined mixtures seems more useful (but difficult). An extreme situation of this problem is the separation using only one single observation [6], [16]. In most cases we assume that the mixtures were linearly mixed from the source signals, corresponding to certain artificial mixtures and music produced in the studio. But in the real world, a non-linear mixing model will be more suitable considering the reflection, attenuation and delay.

The aim of our research is to develop a methodology for separating musical audio into streams of individual instruments. In musical audio, it is very common that many instruments are played at the same time. Therefore, the task is particularly challenging since the mixture is highly under-determined and the recording condition may vary from the studio to the real world. Furthermore, there is a high possibility that some instruments play the same note at the same time. Since they will share many common frequencies and this makes the task difficult to solve in the frequency domain.

The approach presented in this paper was inspired by automatic music transcription [14]. Intuitively, if the results of transcription can provide us all the individual notes of each instrument, then the separation can be achieved by classifying these notes into channels of individual instruments. Sometimes we call these notes *bases* and call the whole set of notes a *dictionary* because they are the basic elements that make up musical audio. However, in our task, the bases in the dictionary are not necessarily as explicit as notes in a transcription problem, because all the bases will be grouped together eventually. So as long as each inferred basis comes from one particular instrument, the separation can be realized after correct classification of the bases. Thus, a looser criterion than transcription can be adopted.

In our current system, we use the non-negative matrix factorization (NMF) algorithm to decompose the input

Email: beiming.wang@elec.qmul.ac.uk

signal in time-frequency domain. Then we generate time-frequency masks by comparing the energies of decomposed bases and apply those masks to the spectrogram. Finally, grouping of bases is made in the time domain to produce separated audio streams. In following sections, We explain the NMF algorithm and the masking method in section 2 and 3. The experimental results is given in section 4. More discussion about improving the performance and future work will be addressed in the end.

2. NON-NEGATIVE MATRIX FACTORIZATION

Non-negative Matrix Factorization, first proposed by Lee and Seung [11], is a data-adaptive linear representation method. The goal of this algorithm is to decompose a matrix $V \in \Re^{\geq 0, M\times N}$ (where $\Re^{\geq 0, M\times N}$ is an M by N non-negative real value matrix) into the product of two non-negative matrices: a basis matrix $W \in \Re^{\geq 0, M\times R}$ and a coefficient matrix $H \in \Re^{\geq 0, R\times N}$. Lee and Seung [12] also developed two fast multiplicative algorithms for NMF.

2.1. Definitions and learning rule

Given an M by N non-negative matrix V, we wish to approximate V by the product of two non-negative matrices W and H:

$$V \approx WH \tag{1}$$

where W is an M by R matrix and H is R by N, respectively. The goal of NMF is to find such a pair of W and H which minimizes the error of reconstruction. In terms of different measure of error, there are two cost functions defined. The first cost function is the Euclidean distance between V and WH:

$$C = \|V - WH\| = \sum_{ij} (V_{ij} - (WH)_{ij})^2 \tag{2}$$

An alternative measure is the divergence between V and WH:

$$D = \sum_{ij} \left(V_{ij} \log \frac{V_{ij}}{(WH)_{ij}} - V_{ij} + (WH)_{ij} \right) \tag{3}$$

Both of these measures are lower bounded by zero and optimized if and only if $V = WH$.

Setting vector w_r to be the rth column of W and vector h_r to be the rth row of H, that is, $W = \{w_1, w_2, ..., w_R\}$, $H = \{h_1, h_2, ..., h_R\}^T$, then equation 1 can be re-written as:

$$V \approx \sum_{r=1}^{R} w_r h_r \tag{4}$$

Regarding each vector w_r as a basis feature in V, the corresponding h_r is the vector of coefficients or encoding of this feature.

To minimize equation 2 or 3, standard gradient descent rules can be adopted, although Lee and Seung also derived fast multiplicative update rules for both of them. More details including proof of convergence can be found in their paper [12].

2.2. NMF on music

In section 2.1, we can see that NMF method can decompose the original signal into the sum of different basis. Depending on what these bases are, various applications benefit from the result. Dealing with music, Smaragdis [18] showed that the components in basis matrix W can be individual notes and proposed an approach of polyphonic music transcription using NMF. Therefore, if there are multiple instruments presented, we can separate them by doing separation among those notes.

The NMF method was initially developed for image signal processing since a 2-D image can be regarded as a non-negative matrix. Time domain audio signals are not suited for this method since they include both positive and negative values. However, the magnitude of the spectrogram meets the non-negative requirement perfectly. A spectrogram $S(t, \omega)$ is calculated by dividing the time domain signal $s(\tau)$ into small frames and performing Discrete Fourier Transform (DFT) on each frame:

$$S(t, \omega) = DFT\Big\{ s(\tau)\gamma(\tau - t) \Big\} \tag{5}$$

where $\gamma(\tau - t)$ is a small time window. The magnitude of spectrogram is $\|S(t, \omega)\|$.

Figure 1 shows the result of decomposition of the spectrogram of an artificial signal using NMF. The bases in W matrix are features in frequency domain which can be notes in a certain situation, and the H matrix records their locations along time.

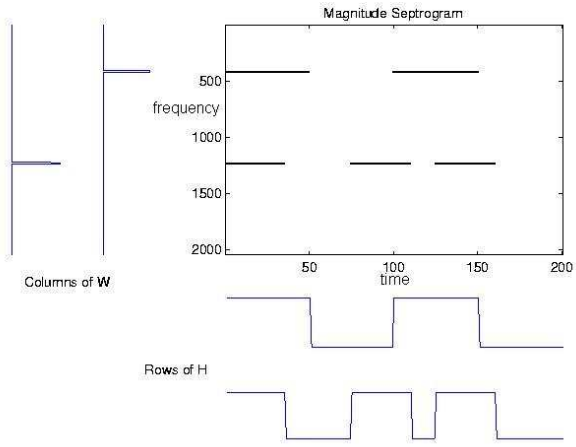

Figure 1. Decomposition of a simple spectrogram using NMF

3. SEPARATION AND RECONSTRUCTION

Having achieved the decomposition result in section 2, we can easily separate the two different frequencies by multiplying the rows of W with the corresponding columns of H, that is, $w_r h_r$. As shown in figure 2, the separated objects are the magnitudes of spectrograms as well, since the whole processing are constrained to be non-negative. If we want to recover the original audio, we also need phase information.

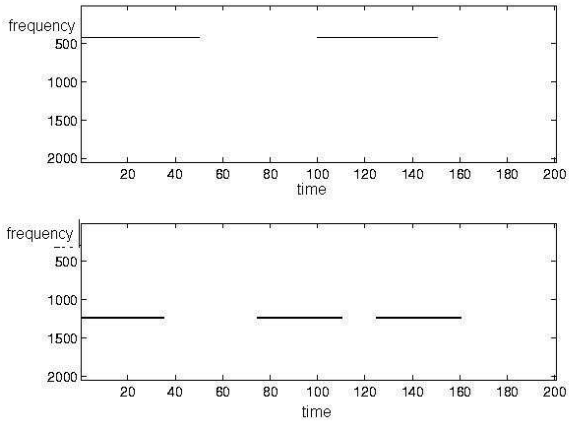

Figure 2. Separation of the spectrogram

To overcome this problem, we can estimate the phases for the outputs by probabilistic inference, but this method is computationally expensive [1]. In this paper, we obtain the phases from the original mixed signal.

From equation 4, we can see that $w_r h_r$ is a part of V and it is contributed only by basis r. We compare each time-frequency point among all the bases, and generate masks for each bases. If one point were the maximum one in all the basis' spectrograms at the same position, it will be marked as 1 in its related masker. Otherwise, zero will be assigned. This is illustrated in formula 6:

$$M_{ij}^{(k)} = \begin{cases} 1 & if \quad k = \arg\max_r (w_r h_r)_{ij} \\ 0 & others \end{cases} \tag{6}$$

where M^k is the mask for basis k. By putting these binary masks on the original spectrogram, we have both the magnitude and the phase.

This idea based on the assumption that over a small time-frequency region, one instrument (or sound source) dominates. In other words, the maximum of the individual spectrograms over each region is approximately equal to the sum of all the spectrograms. This is also the core assumption of DUET [15] method.

4. EXPERIMENTAL RESULTS

In this section, we show the performance of our system on both a manually made mixture and real music audio.

4.1. Artificial musical audio

In experiment one, we tested on a manually mixed audio input so we have the knowledge of the original signals. The input signal is sampled at 22,050Hz and lasts for 10 seconds. 32 bases were learned by NMF, which we grouped into two streams. Figure 3 shows the comparison between the original signal and the recovered ones. A small difference between the original and the recovered signals is slightly visible from their waveforms. Listening to the results, we found the recovered flute sound had better quality while the piano sound suffered some interference towards the end.

A simple evaluation was done by comparing the recovered signals with the original ones. The signal-to-noise ratio (SNR) expressed in decibels (dB) is used and the noise is calculated by least square error:

$$SNR(i) = 10 \log_{10} \frac{\sum_t [x_i(t)]^2}{\sum_t [x_i(t) - r_i(t)]^2} \tag{7}$$

where $x_i(t)$ is the ith input signal and $r_i(t)$ is the separated counterpart. We ran the algorithm several times with different initial value and bases number so the SNR varied. The experiment results are shown in table 1, from where we can see that the flute sound has higher SNR than the piano sound. The result matches what we expected from our listening tests.

SNR (dB)	flute	piano
32 bases NMF	15.2 ± 3.0	11.6 ± 3.0
56 bases NMF	21.7 ± 1.9	18.0 ± 1.9

Table 1. Evaluation of results

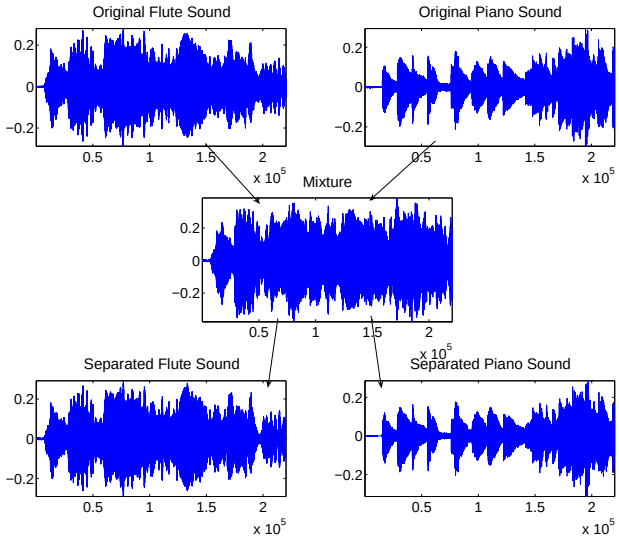

Figure 3. Separation of artificial mixture

The sound sources above are completely independent since there is no relation between the content. However, in real music, instruments often play the same note at the same time or share common harmonics, which is likely to make instrument separation more difficult.

4.2. Real musical audio

In experiment two, the test recording is a music clip played by flute and guitar. The flute is the main instrument and the guitar is played as an accompaniment. The flute

sound ends before the guitar. The input data is real musical audio where the original individual sources are not available, so only subjective analysis can be given. From figure 4, we can see visually that the flute sound is not completely zero at the end. When we listen to it, the noise is actually some guitar sound leaking into the flute channel. Since the flute dominates most of the time, it is more likely that energy will be allocated to the flute channel rather than the guitar channel. The result sounds acceptable on the whole and the rhythm is clearly extracted as confirmed visually in figure 4.

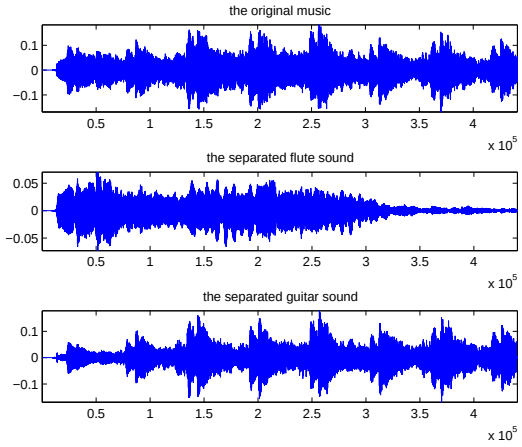

Figure 4. Separation of real music

5. DISCUSSION

One problem of performing NMF in practice is how to choose a proper value of the number of bases R, because most of the time we will not know this beforehand. Even if we do know the total amount of the notes appearing in the mixture, this is not enough to tell us the size of dictionary since we may need several bases to present one note.

It was known that an over-complete basis dictionary with a sparse coefficients matrix has greater flexibility in matching structure in the provided data [13]. So one feasible way is setting R with a value greater than the real number of the notes appearing in the data. Although we can not have a dictionary of notes, as long as each basis comes from one single instrument, the separation task is achievable. This can be benefit from over-complete representation solutions and leads Hoyer [10] to propose a sparse constrained NMF version which enhances the performance when the sparseness is set appropriately.

Smaragdis [17] developed a convolutive NMF model to take advantage of the temporal structure in the spectrogram (see also related work by Virtanen [19]). In his system, repeated features are well extracted and it has been shown to be suitable for percussive music.

In our current system, the recovered bases are till manually classified. An automatic grouping method will make it more practical when the number of bases increases. Other work to be done in the near future includes quantitative evaluation of our system on real music input which means that the source signals are required. Some digital music using Creative Commons lincense are distributed along with the individual tracks before being mixed, using these, a more objective evaluation can be performed.

6. CONCLUSION

As shown in our current progress, non-negative matrix factorization is a possible way to find individual components in mixed data. Although it may not be robust enough to achieve a basis set corresponding directly to notes, NMF and binary masking works well for our separation task. Since sparse representation has been proved very useful in polyphonic music transcription [14], in future work we will investigate combining these two methods.

7. ACKNOWLEDGEMENTS

BW is supported by EPSRC Grant GR/S75802/01 and a Research Studentship from the Department of Electronic Engineering at Queen Mary University of London. This work is also partially supported by EPSRC Grant GR/S82213/01 and EU-FP6-IST-507142 project SIMAC (Semantic Interaction with Music Audio Contents): more information can be found at the SIMAC website http://www.semanticaudio.org.

8. REFERENCES

[1] K. Achan, S. T. Roweis, and B. J. Frey. Probabilistic inference of speech signals from phaseless spectrograms. In S. Thrun, L. Saul, and B. Schölkopf, editors, *Advances in Neural Information Processing Systems 16*, pages 1393–1400. MIT Press, Cambridge, MA, 2004.

[2] S.-I. Amari, A. Cichocki, and H. Yang. A new learning algorithm for blind source separation. In D. S. Touretzky, M. C. Mozer, and M. E. Hasselmo, editors, *Advances in Neural Information Processing Systems 8*, pages 757–763. MIT Press, Cambridge, MA, 1996.

[3] A. J. Bell and T. J. Sejnowski. An information-maximization approach to blind separation and blind deconvolution. *Neural Computation*, 7(6):1129–1159, 1995.

[4] A. S. Bregman. *Auditory Scene Analysis, The perceptual Organization of Sound*. MIT Press, fourth edition, 2001.

[5] J. F. Cardoso and P. Comon. Independent component analysis, a survey of some algebraic methods. In *IEEE International Symposium on Circuits and Systems (ISCAS'96)*, volume 2, pages 93–96, 1996.

[6] M. A. Casey and A. Westner. Separation of mixed audio sources by independent subspace analysis. In *Proceedings of the International Computer Music Conference*, pages 154–161, Berlin, Germany, Auguest 2000.

[7] P. Comon. Independent component analysis — a new concept? *Signal Processing*, 36(3):287–314, 1994.

[8] D. P. W. Ellis. *Prediction-driven Computational Auditory Scene Annlysis*. PhD thesis, Department of Electronic Engineering and Computer Science, MIT, 1996.

[9] I. F. Gorodnitsky and B. D. Rao. Sparse signal reconstruction from limited data using FOCUSS: a reweighted minimum norm algorithm. *IEEE transaction on Signal Processing*, 45(3):600–616, 1997.

[10] P. O. Hoyer. Non-negative matrix factorization with sparseness constraints. *Journal of Machine Learning Research*, 5:1457–1469, Nov 2004.

[11] D. D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401:788–791, 1999.

[12] D. D. Lee and H. S. Seung. Algorithms for non-negative matrix factorization. In V. T. Todd Leen, Tom Dietterich, editor, *Advances in Neural Information Processing Systems 13*, pages 556–562. MIT Press, Cambridge, MA, 2001.

[13] B. A. Olshausen and D. J. Field. Sparse coding with an overcomplete basis set: A strategy employed by V1? *Vision Research*, 37(23):3311–3325, 1997.

[14] M. D. Plumbley, S. A. Abdallah, T. Blumensath, and M. E. Davies. *Sparse Representations of Polyphonic Music*. To appear in *Signal Processing*, 2005.

[15] S. Rickard and F. Dietrich. DOA estimation of many W-disjoint orthogonal sources from two mixtures using DUET. In *Proceedings of the 10th IEEE Workshop on Statistical Signal and Array Processing (SSAP2000)*, pages 311–314, Pocono Manor, August 2000.

[16] S. T. Roweis. One microphone source separation. In V. T. Todd Leen, Tom Dietterich, editor, *Advances in Neural Information Processing Systems 13*, pages 793–799. MIT Press, Cambridge, MA, 2001.

[17] P. Smaragdis. Non-negative matrix factor deconvolution; extraction of multiple sound sources from monophonic inputs. In *5th International Conference on Independent Component Analysis and Blind Source Separation (ICA'04)*, page 494, Granada, Spain, 2004.

[18] P. Smaragdis and J. C. Brown. Non-negative matrix factorization for polyphonic music transcription. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA'03)*, pages 177–180, October 2003.

[19] T. Virtanen. Separation of sound sources by convolutive sparse coding. In *Proceedings of the ISCA Tutorial and Research Workshop on Statistical and Perceptual Audio Processing (SAPA 2004)*, Jeju, Korea, 3 October 2004.