INVESTIGATING SINGLE-CHANNEL AUDIO SOURCE SEPARATION METHODS BASED ON NON-NEGATIVE MATRIX FACTORIZATION

Beiming Wang **Mark D. Plumbley**

Centre for Digital Music, Department of Electronic Engineering
Queen Mary, University of London
beiming.wang, mark.plumbley@elec.qmul.ac.uk

ABSTRACT

Our research aims to separate multiple sound sources from a single-channel audio mixture, and in this paper, we present a framework featured by Non-negative Matrix Factorization (NMF). Within this framework, we proposed two approaches which are referred as *Un-directed* and *Directed NMF model*. The *Un-directed NMF model* decomposes the mixing data in an unsupervised manner but requires human interaction for clustering. We have developed a simple graphical user interface for this task. Provided with isolated training data, the Directed NMF is performed under the direction of pre-trained models and therefore, does not need user interaction. Experimental results show this framework is a feasible way to achieve high quality separation. Successful separation of individual sound sources could assist with other tasks such as automatic music transcription, object coding, special sound effects and so on.

Keywords: Single-channel separation, Non-negative Matrix Factorization, Blind Source Separtion

1 INTRODUCTION

Real world audio signals, especially music, is often a combination of a number of independent sound sources such as various instrumental sounds, human voice, natural sound and so on. Blind Audio Source Separation (BASS) is the task of recovering those sound sources from the given mixtures, and can contribute to speech/instrument identification, automatic music transcription, special sound effects and many other applications[14].

Generally speaking, the difficulty of the BASS problem increases as the number of mixtures is reduced. In musical recordings, its very common to have multiple instruments playing at the same time and there is a high possibility that some instruments play a same note at the same time. Furthermore, musical recordings usually require good fidelity, which means high perceptual quality is needed after processing. Therefore, separating different sound sources from single-channel musical audio is a very challenging topic.

A general instantaneous Blind Source Separation problem can be formulated as:

$$X = AS \tag{1}$$

where X is the observed I mixtures: $X(t) = [x_1(t), x_2(t), \cdots, x_I(t)]^T$ and S is the source signals to be estimated: $S(t) = [s_1(t), s_2(t), \cdots, s_J(t)]^T$. A denotes a mixing matrix which is corresponding to the mixing conditions. In the single-channel separation case, the number of mixtures I drops to 1 and the formula can be simplified as:

$$x(t) = \sum_{j=1}^{J} s_j \tag{2}$$

and our task is to extract s_j from the give mixture x.

During the last decade, many BASS algorithms have emerged, such as ICA[3], DUET[8] and sparse coding[7]. However, most of these work only when there are at least two mixtures. Many researchers have developed approaches specificly for the issue of single-channel mixture by utilising additional training data, such as the refiltering approach based on Factorial Hidden Markov Model (FHMM) proposed by Roweis[9], Jang and Lee's Maximum Likelihood approach[4], and the approach based on an extension of classical Wiener filtering proposed by Benaroya et al[2].

This paper is organised as follows: in section 2, we introduce Non-negative Matrix Factorization (NMF), the core algorithm in our approach, followed by a detailed analysis of our proposed framework and methods in section 3. The experiments and results comparisons are provided in section 4.

2 NON-NEGATIVE MATRIX FACTORIZATION

Non-negative Matrix Factorization, first proposed by Lee and Seung [5], is a data-adaptive linear representation method for 2-D matrices. Given an 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), the algorithm represents V as the product of two non-negative matrices:

$$V \approx \widehat{V} = WH \tag{3}$$

where $W \in \Re^{\geq 0, M \times R}$ is called the *basis matrix* and $H \in \Re^{\geq 0, R \times N}$ is the *coefficient matrix*. The parameter R indicates the number of basis used to represent the original matrix. To find such a pair of W and H which minimises the error of the approximation in (3), two alternative cost functions are defined, Euclidean distance, C, and Divergence, D:

$$C = \| V - WH \| = \sum_{ij} (V_{ij} - (WH)_{ij})^2 \tag{4}$$

$$D = \sum_{ij} (V_{ij} \log \frac{V_{ij}}{(WH)_{ij}} - V_{ij} + (WH)_{ij}). \tag{5}$$

Both of these two measures are lower bounded by zero and optimised if and only if $V = WH$. To minimise (4) or (5), standard gradient descent rules can be adopted. Lee and Seung[6] have derived fast multiplicative update rules for both of them. For more details see [6].

3 METHODS

Since a 2-D image can be regarded as a non-negative matrix, NMF was first applied to parts-based representation in image processing. However, by transforming a sound wave into time-frequency domain, the magnitude of the spectrogram meets all the requirements of NMF perfectly. Based on spectrograms, Smaragdis [10] adapted NMF to the polyphonic music transcription task. We use NMF in a similar way to solve the single-channel musical audio source separation problem.

3.1 Overview of the framework

Our method can be generalised to a framework consisting of three steps: decomposition, separation and reconstruction. It is based on the idea that music signal may be represented by a set of basic components, which can be notes or other general harmonic structures. We call these components basis vectors and decompose the single mixture into a number of basis vectors. If different sound objects use different basis sets, the separation task is equivalent to clustering those basis vectors into their corresponding sources. By combining the clustered basis set, we obtain estimated source signals. In this section, we introduce these three steps in details.

3.2 Decomposition using NMF

According to basic matrix algebra, equation (3) can be rewritten as:

$$V \approx \widehat{V} = \sum_{r=1}^{R} w_r h_r \tag{6}$$

where V is the magnitude spectrogram of the mixture. Vector w_r is the rth column of $W = \{w_1, w_2, ..., w_R\}$, which is the set of basis functions we are looking for. Correspondingly, vector h_r is the rth row of $H = \{h_1, h_2, ..., h_R\}^T$ and are the coefficients or weights of each basis. This operation is illustrated in Figure 1 for a 'toy' spectrogram containing only two different frequencies.

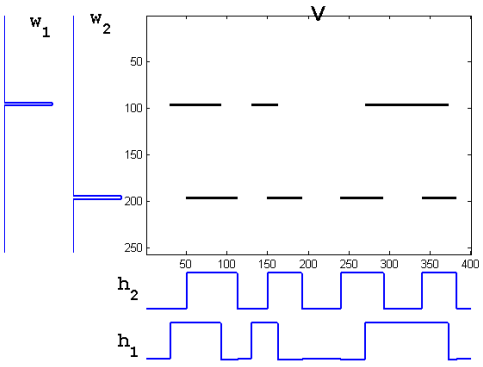

Figure 1: Decomposition of a simple spectrogram using NMF

When NMF is performed on the mixture without additional constraints, we call it an *Un-directed NMF model*. The basis functions learnt from an un-directed model are learnt from mixture instead of the sources, so, the learnt basis can either be a component shared by multiple sources, or a feature composed of them.

If we provide training data for each source, the un-directed NMF can be guided with the information acquired from training. We call this a *Directed Decomposition NMF model*. A training step is performed before decomposing the mixture with the same algorithm, and echoes some other related methods [2, 4, 9]. At the instance of two sources, $x = s_1 + s_2$, basis sets W_1 and W_2 are learnt from training data $\tilde{s}_1$, $\tilde{s}_2$ according to (3):

$$\begin{aligned} V_1 &\approx W_1 H_1 \\ V_2 &\approx W_2 H_2 \end{aligned} \tag{7}$$

where V_1 and V_2 are the magnitude spectrograms of $\tilde{s}_1$ and $\tilde{s}_2$ respectively. W_1 and W_2 are then used to represent the mixture by setting matrix W to be the combination of the two basis sets:

$$W = [W_1, W_2] = [w_{11}, w_{12} \ldots w_{1p}, w_{21}, \ldots w_{2q}] \tag{8}$$

where p and q are the size of the two sets respectively. This combined basis set forces the mixture to be represented by W, which is not further updated during decomposition.

3.3 Separation

In order to reconstruct the original sources, the basis vectors w_r learnt during decomposition need to be grouped. For the *Directed Decomposition NMF model*, this step is very straightforward since the pre-trained basis W is already grouped and does not change during processing:

$$\widehat{V}_1 = \sum_{r=1}^{p} w_r h_r \tag{9}$$

$$\widehat{V}_2 = \sum_{r=p+1}^{p+q} w_r h_r \tag{10}$$

where $\widehat{V}_1$, $\widehat{V}_2$ are the magnitude spectrograms of the estimated sources.

Without the hint of additional data, the grouping problem is more difficult for *Un-directed NMF model*. The user may have to manually cluster all the basis vectors by viewing or listening to them. We developed a simple graphical user interface to make this job easier (Figure 2). Even with the help of GUI, manual grouping can be

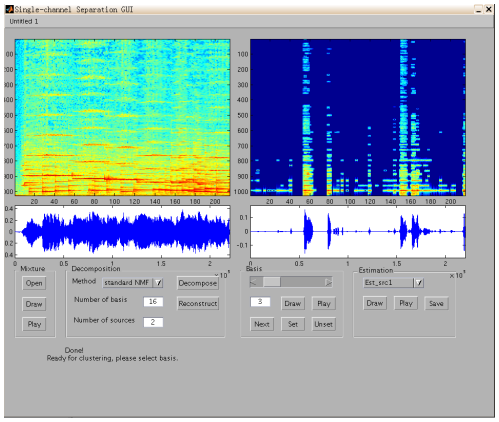

Figure 2: Graphical user interface for single channel separation

time consuming and inaccurate. If additional training data is available, we can take advantage of it to help grouping even if we did not use this during decomposition. Here we suppose that all the bases from a certain source belong to a subspace spanned by its corresponding trained basis set, $w_1 \dots w_p$ for instance. With this hypothesis, the bases in W learnt from the mixture as in *Un-directed NMF model* can be clustered by measuring their distance to those subspaces, and classified into the closest subspace. We call his the *Directed Clustering NMF model*.

3.4 Reconstruction

By multiplying the rows of W with the corresponding columns of H, we obtain the magnitude spectrograms of each basis as shown in Figure 3. Therefore, by adding up the basis spectrograms for a particular source, we can reconstruct the estimated source signals.

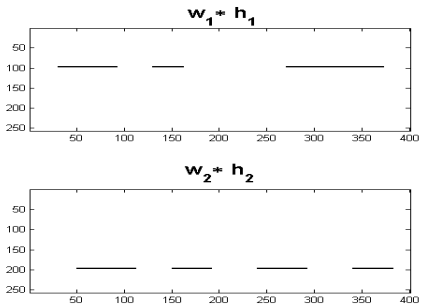

Figure 3: Separation of the spectrogram

As in many other time-frequency methods, our framework operates on the magnitude spectrogram. It does not involve any estimation about phase information which can affect the perceptual quality of the estimated audio[1]. To estimate the lost phases, we perform a binary masking in time-frequency domain.

This idea is based on the assumption that over a small time-frequency region, one instrument (or sound source) dominates. In other words, the energies of different sources are not heavily overlapped at a particular time-frequency point. This is also the basic assumption of the DUET [8] method. According to this, for each point in a spectrogram, we allocate both the magnitude and phase to the source which has higher energy by binary masking.

4 RESULTS

We evaluated the proposed methods on three datasets. The first and second datasets are selected from the Blind Audio Source Separation evaluation database[1] maintained by E. Vincent et al[11], in which examples are all extracted from multi-track recorded real music samples. *Dataset1* includes guitar and tin whistling sounds with similar melody. *Dataset2* has a guitar as the leading instrument and repeated drums rhythm in the background. The two sources in *Dataset3* are flute and cello sound samples from different music and manually mixed together, so they are not significantly harmonically related to each other.

4.1 Evaluation methods

In our experiments, we use the SDR, SIR and SAR values calculated by BSS_EVAL toolbox[13] to measure the performance. Because of the assumption that time-frequency masking methods are based on, there is always a minimum error depending on the mixture data. This error indicates the best performance achievable by the approach involving time-frequency masking and can be calculated by the BSS_ORACLE toolbox [12]. Therefore, we also include the "oracle estimation" of each dataset.

4.2 Experimental results

Table 1 shows the results averaged over 10 trials in which 32-basis NMF is chosen for decomposition. The oracle result indicates that this approach can be further improved if a better algorithm can be found, and when we listen to one output, short sounds from another source can be heard from time to time. Since the changing of SAR and

Table 1: Evaluation of *Un-directed NMF model*

(dB)	SAR	SIR	SDR	SDR_Oracle
Dataset1	13.62	33.53	13.51	31.07
Dataset2	14.40	34.15	14.33	22.16
Dataset3	8.06	22.24	7.65	18.88

SIR value has the same trend as SDR, we only use SDR in the following comparisons. In Table 2, the Directed NMF methods are evaluated with an additional 30 seconds of

[1]Data available at http://bass-db.gforge.inria.fr/BASS-dB/.

training data from the corresponding music file provided for each dataset. The *Directed Decomposition* (DD) approach brings a noticeable improvement (10dB in SDR) for *Dataset1* but only a small increase for *Dataset3* and even a decrease for *Dataset2*. The difference corresponds to the similarity between the source signal s_i in mixture and its training data $\tilde{s}_i$. As the assumption of Directed Decomposition implies, the more s_i and $\tilde{s}_i$ are similar, the more bases they have in common. Since the similarity affects the performance, this approach should achieve its best result when the training data is identical to the original source. Therefore, we tested this using the original source signals as training data (Table 3). While there is still a gap between Directed Decomposition method and the oracle score, both Direct Decomposition and Clustering methods are greatly improved.

Table 2: Evaluation of Directed NMF model (UD denotes the Un-directed NMF model, DD denotes the Directed Decomposition NMF model and DC denotes the Directed Clustering NMF model)

SDR (dB)	UD	DD	DC	Oracle
dataset1	13.51	23.75	13.60	31.07
dataset2	14.33	10.18	9.98	22.16
dataset3	7.65	8.26	6.13	18.88

Table 3: Evaluation of Directed NMF model with original sources

SDR (dB)	UD	DD	DC	Oracle
dataset1	13.51	24.68	13.67	31.07
dataset2	14.33	18.14	14.69	22.16
dataset3	7.65	13.30	7.48	18.88

5 CONCLUSION

We have proposed a framework for single channel audio separation and realized it using the Non-negative Matrix Factorization algorithm. Three approaches are developed and compared with different datasets. As expected, the *Directed NMF* approach performs well when a proper training set is provided. The *Directed* approach can either be used to improve the overall performance by pre-training the basis vectors, or providing a reliable clustering criterion. The results achieved by training with original source shows that there are basis sets that can make sound source separation distinct. Nevertheless, the problem of how to learn such a basis set from limited information remains to be solved.

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 http://www.semanticaudio.org.)

References

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

[2] L. Benaroya, F. Bimbot, and R. Gribonval. Audio source separation with a single sensor. *IEEE Transactions on Speech and Audio Processing*, 14(1):191–199, Jan 2006.

[3] A. Hyvärinen, J. Karhunen, and E. Oja. *Independant Component Analysis*. John Wiley & Sons, 2001.

[4] G-J. Jang and T-W. Lee. A maximum likelihood approach to single-channel source separation. *Journal of Machine Learning Research*, 4:1365–1392, Dec 2003.

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

[6] D. D. Lee and H. S. Seung. Algorithms for non-negative matrix factorization. In *Advances in Neural Information Processing Systems 13*, pages 556–562. MIT Press, Cambridge, MA, 2001.

[7] 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.

[8] 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 (SSAP'00)*, pages 311–314, Pocono Manor, August 2000.

[9] S. T. Roweis. One microphone source separation. In *Advances in Neural Information Processing Systems 13*, pages 793–799. MIT Press, Cambridge, MA, 2001.

[10] P. Smaragdis and J. C. Brown. Non-negative matrix factorization for polyphonic music transcription. In *Proceedings of the IEEE WASPAA'03*, pages 177–180, October 2003.

[11] E. Vincent, R. Gribonval, and C. Féotte. BASS-dB: the blind audio source separation evaluation database, . URL http://www.irisa.fr/metiss/BASS-dB/.

[12] E. Vincent, R. Gribonval, and M. D. Plumbley. Oracle estimators for the benchmarking of source separation algorithms. Submitted, .

[13] E. Vincent, C. Févotte, L. Benaroya, and R. Gribonval. A tentative typology of audio source separation tasks. In *Proceedings of ICA'03*, pages 715–720, Nara, Japan, April 2003.

[14] E. Vincent, S. A. Abdallah, M. D. Plumbley, M. G. Jafari, and M. E. Davies. Blind audio source separation. Technical Report C4DM-TR-05-01, Centre for Digital Music, Queen Mary University of London, November 2005.