MONOPHONIC INSTRUMENT SOUND SEGREGATION BY CLUSTERING NMF COMPONENTS BASED ON BASIS SIMILARITY AND GAIN DISJOINTNESS

Kazuma Murao Masahiro Nakano Yu Kitano Nobutaka Ono Shigeki Sagayama
Graduate School of Information Science and Technology
The University of Tokyo, Japan
{ k-murao, mnakano, kitano, onono, sagayama } @hil.t.u-tokyo.ac.jp

ABSTRACT

This paper discusses a method for monophonic instrument sound separation based on nonnegative matrix factorization (NMF). In general, it is not easy to classify NMF components into each instrument. By contrast, monophonic instrument sound gives us an important clue to classify them, because no more than one sound would be activated simultaneously. Our approach is to classify NMF components into each instrument based on basis spectrum vector similarity and temporal activity disjointness. Our clustering employs a hierarchical clustering algorithm: group average method (GAM). The efficiency of our approach is evaluated by some experiments.

1. INTRODUCTION

In music signals, there are usually multiple sound sources such as a human singing voice and instruments sound. The task to separate mixed signals into individual sources is called sound source separation for music signals. It has several applications such as music equalizer, music information searching, automatic transcription, and structured coding of music. This paper discusses a method to separate monaural musical audio into individual musical instruments.

Sound source separation for music signal has been widely investigated recently. Some methods are based on supervised learning of individual source models [1–3]. They need solo excerpts beforehand. Other unsupervised approaches have also been studied [4–6]. Because any prior information for instrumental sound sources cannot be used, some unsupervised methods make assumption about common harmonic structure [4, 5] or employ the excitation-filter model of sound production [6]. We propose an efficient unsupervised method focusing on monophonic instrument sound.

Our method have two stages. At the first stage, we factorize the observed spectrogram into some components based on nonnegative matrix factorization (NMF) [9, 10]. In the case of music signals, each component usually represents a musically meaningful element, so that different elements are expected to correspond to different components.

However, considering music instrumental source separation, methods based on NMF generally encounter difficulties in the components clustering step. And most of the algorithm count on manual clustering [7]. Some clustering methods separate percussive instrument sources [8,12], but are rarely used with harmonic instruments sources.

This paper proposes a method for clustering components that employs not only spectral information but also temporal information. The outline of this paper is as follows. Section 2 gives a overview of NMF algorithm and component-clustering problem. The proposed clustering method is explained in Section 3, and experimental evaluation of proposed method are presented in Section 4. Section 5 covers the conclusions and future works.

2. NONNEGATIVE MATRIX FACTORIZATION

Nonnegative matrix factorization and some unsupervised sound source separation algorithms are based on a signal model where the spectrum vector $\mathbf{x}_t$ ($t = 1, ..., T$) in frame is modeled as a linear combination of *basis vectors* $\mathbf{b}_j$ ($j = 1, ..., J$). This can be written as

$$\mathbf{x}_t = \sum_{j=1}^{J} g_{j,t}\mathbf{b}_j, \tag{1}$$

where J is the number of basis vectors, and its time-varying gain (amplitude) $g_{j,t}$, T being the number of frames.

This model can be written using a matrix notation as

$$\mathbf{X} = \mathbf{BG}, \tag{2}$$

where $\mathbf{X} = [\mathbf{x}_1, ..., \mathbf{x}_T]$, $\mathbf{B} = [\mathbf{b}_1, ..., \mathbf{b}_J]$, and $[\mathbf{G}]_{j,t} = g_{j,t}$.

Here, $\mathbf{g}_j = [g_{j,1}, ..., g_{j,t}]^{\mathbf{T}}$ is defined as *gain vector* corresponding to the basis vector, then the term *component* refers to one basis vector $\mathbf{b}_j$ and one corresponding gain vector $\mathbf{g}_j$. Each source is modeled as a sum of the components. The separation is done by first factorizing the spectrogram of the input signal into components and second grouping these to sound sources.

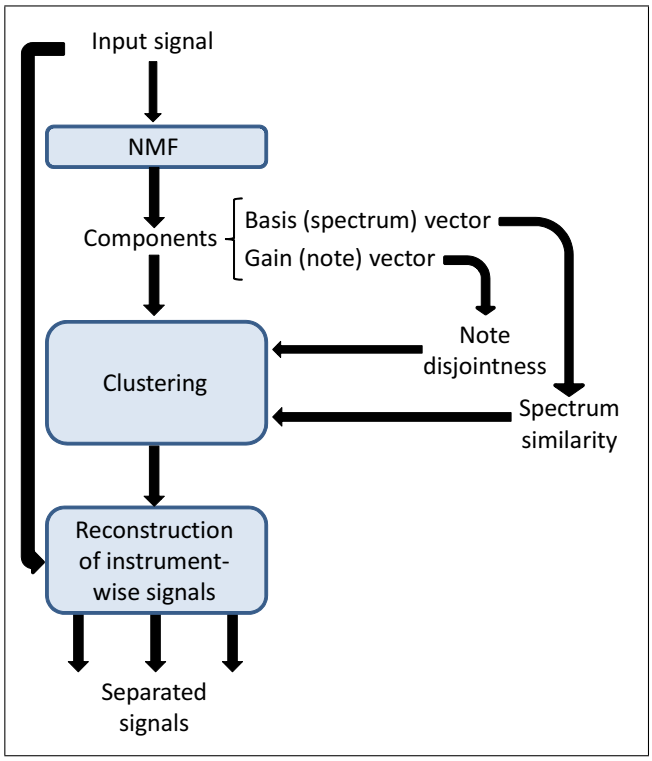

Figure 1. Flow diagram of the method.

The NMF algorithms proposed by Lee and Seung [9] do the decomposition by minimizing the reconstruction error between the observation and the model while constraining the matrices to be entry-wise nonnegative as follows:

$$D(\mathbf{X}\|\mathbf{BG}) = \sum_{f,t} d([\mathbf{X}]_{f,t}\|[\mathbf{BG}]_{f,t}), \qquad (3)$$

here $d(y\|z)$ is a function of two scalar variables. The various measures for reconstruction error are proposed. The Euclidean distance, the generalized Kullback-Leibler divergence [9], or the Itakura-Saito divergence [11] are mostly used. We choose here the generalized Kullback-Leibler divergence, which has produced good results in earlier sound source separation studies [14].

In standard NMF, the only constraints is the elementwise non-negativity of all matrices. Then, several constraints have been proposed in order to achieve expected solutions. The most famous constrains are sparsity [13] and temporal continuity [11, 14]. We use the sparsity and temporal continuity proposed in [14].

We wish to use NMF to decompose the observed signal into the components. However, it is not easy to know which source each component is assigned to. In the next section an automatic clustering method is proposed.

3. CLUSTERING OF NMF COMPONENTS

3.1 Outline

As a result of NMF, basis vectors $\mathbf{b}_j$ and gain vectors $\mathbf{g}_j$ are obtained, each of which could ideally represent spectrum and temporal activity of each note, respectively. The problem here is how to classify obtained components $(\mathbf{b}_j, \mathbf{g}_j)$ into each instrument. The contribution of this paper is to exploit both information of $\mathbf{b}_j$ and $\mathbf{g}_j$) for mixture of monophonic instrumental tracks without any prior about each instrument. Our approach consists of 1) measuring the basis spectrum similarity $C_1(i,j)$ for any pairs of $\mathbf{b}_j$ and $\mathbf{b}_j$, 2) measuring the temporal activity disjointness $\tilde{C}_2(i,j)$ for any pairs of $\mathbf{g}_i$ and $\mathbf{g}_j$, 3) calculating a closeness measure $C(i,j)$ for any pairs of $(\mathbf{b}_i, \mathbf{g}_i)$ and $(\mathbf{b}_j, \mathbf{g}_j)$ by product of $C_1(i,j)$ and $\tilde{C}_2(i,j)$, and 4) applying a kind of hierarchic clustering method.

3.2 Similarity of Basis Spectra

Monophonic source signal is represented by a sinusoidal model [15] as

$$s(t) = \sum_{r=1}^{R} A_r(t)\cos[\theta_r(t)] + e(t) \qquad (4)$$

where $e(t)$ is the noise term, $A_r(t)$ and $\theta_r(t) = \int_0^t 2\pi r f_0(\tau)d\tau$ are the instantaneous amplitude and phase of the rth harmonic, respectively, $f_0(\tau)$ is the fundamental frequency at time τ, and R is number of the harmonic overtone. Harmonic structure is an approximately invariant feature for a harmonic instrument when it is played in a narrow pitch range. [16]

In logarithmic frequency (log-frequency) scale, the harmonic frequencies are located $\log 2$, $\log 3$, ..., away from the log-fundamental frequency, and the relative-location relation remains constant no matter how fundamental frequency fluctuates and is an overall parallel shift depending on the fluctuation degree. Thus among the harmonics between the two spectrums of the same instruments are similar; even in case spectrums fundamental frequencies are different, shapes of the spectrums are same when shifted.

The basis vector $\mathbf{b}_j$, which NMF factorize into, represents average spectrum in logarithmic frequency scale. Therefore the correlation-like criterion between two basis vectors are defined as

$$C_1(i,j) = \max_q \frac{\sum_p b_{p+q,i} b_{p,j}}{|\mathbf{b}_i||\mathbf{b}_j|}, \qquad (5)$$

where $b_{p,j}$ is pth value of the basis vector $\mathbf{b}_j$. Put another way, criterion $C_1(i,j)$ means maximum cross-correlation between normalized $\mathbf{b}_i$ and $\mathbf{b}_j$. In intuitive explanation, two spectra are compared, moving along the frequency axis, and are measured largest overlap. For the spectra by harmonic instrument, two spectrums overlap most when two fundamental pitches nearly go over. As a side-effect, two spectrums by inharmonic instruments mark higher value than value between harmonic and inharmonic instrumental spectrums.

Table 1 shows an example of this correlation-like criterions that is calculated by real instrumental signals: RWC music database [17] RWC-MDB-I-2001 No.31-1 and No.33-1, down-sampled to 16 kHz single-channel files. Each spectrum is taken by Wavelet transform of single tone signal. Two spectra of same instrument almost mark higher

	Clarinet			Flute		
	A4	H4	C5	A4	H4	C5
Clarinet A4	**1.00**	**0.96**	**0.81**	0.58	0.67	**0.81**
Clarinet H4		**1.00**	0.74	0.63	0.66	0.72
Clarinet C5			**1.00**	**0.82**	0.73	**0.92**
Flute A4				**1.00**	**0.95**	**0.80**
Flute H4					**1.00**	**0.80**
Flute C5						**1.00**

Table 1. The simirarity measure of basis spectra calculated by indivisual instrumental signals. The higher values than 0.8 are shown in bold style.

value than two spectrums of other instrument mark. However in some cases two spectra which belonging to other instrument mark high numerical number: for example, Clarinet C5 and Flute C5. This result presents that criterion as basis spectrum similarity (5) indicates measure to some extent, but is not enough for the grouping.

3.3 Disjointness of Temporal Activities

Not only basis spectrum $\mathbf{b}_j$, but also temporal activity $\mathbf{g}_j$ should also include cues for clustering components into instrumental tracks. As a simple case to exploit such information, we suppose that all instrumental tracks are *monophonic*, which means each instrumental track consists of a single note sequence.

Figure 2 shows an example of piano-roll representation of three monophonic instrumental tracks. Obviously, any different note activities are disjoint in the same track. Note that there are also many pairs of disjoint note activities over different tracks. Hence, we can't assert that two different note activities belong to the same track even if they are disjoint. However, if two different note activities are NOT disjoint, they should belong to different instrumental tracks.

The disjointness of two different temporal activities represented by gain vectors $\mathbf{g}_i$ and $\mathbf{g}_j$ can be simply calculated by

$$C_2(i,j) = 1 - \frac{\mathbf{g}_i \cdot \mathbf{g}_j}{|\mathbf{g}_i||\mathbf{g}_j|}. \qquad (6)$$

If $\mathbf{g}_i$ and $\mathbf{g}_j$ are disjoint, $C_2(i,j) = 1$. While if they have co-occurrence, $C_2(i,j)$ should take a small value. Therefore, it can be exploited as a closeness measure. Figure 3 shows an expected result, which was calculated by (6) with using temporal activities in piano roll representation shown in Figure 2 as $\mathbf{g}_j$.

The magnitude of $C_2(i,j)$ itself is not significant because it depends on the frequency of the co-occurence. It is only important for clustering whether it is almost zero or not. Furthermore, because of imperfect decomposition by NMF, spectral leakage, reverbration, etc, $C_2(i,j)$ is actualy not equal to zero even if ith component and jth component belong to the same instrumental track. Therefore, we 1) neglect tiny values of $g_{t,j}$ and set them to be zero, 2) calculate $C_2(i,j)$ by (6), and 3) binarize it with a small

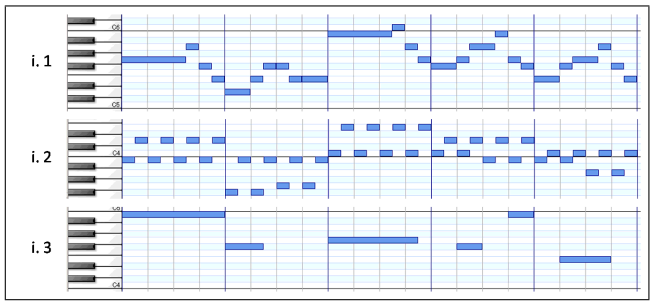

Figure 2. The piano roll representation of three monophonic instrumental track. Any different note activities are disjoint in the same track.

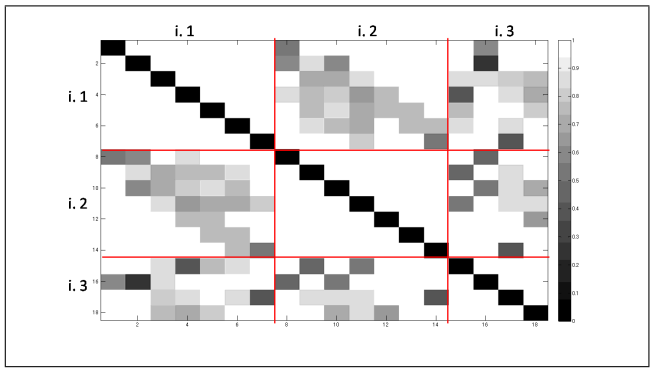

Figure 3. Criterions between two gain vectors according to the equation (6), corresponding to figure 2. The two vertical line and two horizontal line show the borderlines of the instruments. Values on diagram position are ignorable for the clustering.

threshold ϵ such as

$$\tilde{C}_2(i,j) = \begin{cases} 1 & (C_2(i,j) \geq \epsilon) \\ 0 & (C_2(i,j) < \epsilon) \end{cases}. \qquad (7)$$

3.4 Combining Two Different Criterions

Previous criterions are both scales running from zero to one. In both criterions, higher value means two components' sameness. This paper examines the measure of two different components' closeness as

$$C(i,j) = C_1(i,j) \cdot \tilde{C}_2(i,j). \qquad (8)$$

3.5 Clustering by Group Average Method

To find an optimal partitioning of the components into N classes, the following clustering algorithm called *group average method* (GAM) is employed.

1. At the beginning, all components are considered as different clusters.
2. Two components that have the highest criterion value are connected into the same (new) cluster.
3. Criterions between new cluster and other cluster are updated under the update rule:

$$d(K_1, K_2) = \frac{1}{n_1 n_2} \sum_{i \in K_1} \sum_{j \in K_2} d\,(i,j)\,, \qquad (9)$$

input data	sampling rate	16 kHz
	length	10 sec
	number of instruments	3
frequency analysis	frame shift	16 ms
	frequency resolution	12.0 cent
	frequency range	50–7961 Hz
NMF [9]	iteration	200
	number of components	10–40
Clustering	ϵ	0.05
	number of clusters	4

Table 2. Experimental conditions

where $d(A, B)$ is the criterion between cluster A and B, $d(i, j) = C(i, j)$ is the criterion between components i and j, n_1 and n_2 are the number of components that K_1 and K_2 contain.

4. Iteration: repeat step 2 and 3 until total number of clusters reaches L.

Criterion-update avoids chaining effect where wrong components connects into a chain reaction.

3.6 Reconstruction of Instrument-wise Spectrograms

Spectrograms corresponding to a certain instrument K_l ($l = 1, ..., L$) , $\hat{\mathbf{X}}_l$, can be reconstructed by the equation:

$$\hat{\mathbf{X}}_l = \sum_{j \in K_l} \hat{\mathbf{X}}^{(j)} = \sum_{j \in K_l} \mathbf{b}_j \mathbf{g}_j. \tag{10}$$

Spectrogram of instrument *l* is reconstructed as

$$[\hat{\mathbf{Y}}_l]_{f,t} = \frac{[\hat{\mathbf{X}}_l]_{f,t}}{[\hat{\mathbf{X}}]_{f,t}} [\mathbf{X}]_{f,t}. \tag{11}$$

where $\hat{\mathbf{X}} = \sum_{l=1}^{L} \hat{\mathbf{X}}_l$.

4. EXPERIMENTAL EVALUATION

4.1 Source Conditions

To verify the potential performance of the proposed method as sound source separation, the proposed method was tested on a real performance music data from *MIREX 2007 Evaluation Tasks* [18] : transcription of *String Quartet No.5 3rd Movement Var.5* composed by L. V. Beethoven (see table 2 for the list of the experimental data). We used the data composed of three woodwind instruments (flute, oboe and bassoon). Mixed signal was the result of summing the source signals in time domain, and 9 input signals (10 seconds) were clipped from the mixed signal every 5 seconds.

Time series of amplitude spectrum was analyzed using Gabor wavelet transform with a frame shift of $16\mathrm{ms}$ for input digital signals of $16\mathrm{kHz}$ sampling rate. The lower bound of the frequency range and the frequency resolution were $50\mathrm{Hz}$ and $12\mathrm{cent}$, respectively.

4.2 Evaluated Algorithms and Conditions

The following algorithms were tested.

- Proposed method 1: Components clustering employed both basis vector similarity and gain vector disjointness.

  Since there is no reliable method for the estimation of the number of the components, proposed method was tested by factorizing the input signal into 10–40 components and we decided it to earn the best result.

  In the clustering step, the number of the clusters was chosen as 4 because in the real performance music other than pure instrumental sound (e.g. sounds of breath) were contained.

- Proposed method 2: Components clustering employed only basis vector similarity. Compared with Proposed method 1, the contribution of the time activity disjointness can be evaluated.

- Correct clustering: Components clustering to be assigned each component to a source which leads to the highest signal-to-noise (SNR) as

  $$\mathrm{SNR}(m, j) = 10 \log_{10} \frac{\sum_{f,t} [\mathbf{Y}_m]_{f,t}^2}{\sum_{f,t} ([\mathbf{Y}_m]_{f,t} - [\hat{\mathbf{X}}^{(j)}]_{f,t})^2}. \tag{12}$$

  where $\mathbf{Y}_m$ and $\hat{\mathbf{X}}^{(j)}$ are the *m*th reference and *j*th separated component. A component *j* is assigned to a source *m* which leads to the highest SNR.

- NMF2D [4]: Factorization is doned by NMF2D instead of NMF. When analyzing real music signals, the NMF2D was considered to give good results.

4.3 Evaluation Criterion

The quality of the separated sources was measured by calculating the SNR improvement between the original spectrogram $\mathbf{Y}$ and corresponding separated magnitude spectrogram $\hat{\mathbf{Y}}$ according to the equation

$$\mathrm{SNR[dB]} = \frac{1}{M} \sum_{m=1}^{M} 10 \log_{10} \left(\frac{\sum_{f,t} [\mathbf{Y}_m]_{f,t}^2}{\sum_{f,t} ([\mathbf{Y}_m]_{f,t} - [\hat{\mathbf{Y}}_l]_{f,t})^2} - \frac{\sum_{f,t} [\mathbf{Y}_m]_{f,t}^2}{\sum_{f,t} ([\mathbf{Y}_m]_{f,t} - [\mathbf{X}]_{f,t})^2} \right). \tag{13}$$

For each original spectrogram, the SNR improvement that employs baseline using mixed signal are measured. The SNR has been used in several source separation studies to measure the separation quality.

4.4 Results

The SNR improvement for each data and algorithms are shown in table 3. Average values are means among all of data.

Proposed method 1 marks an average improvement 2.75 dB. For all data, proposed method 1 sets positive values.

	SNR [dB]			
	proposed 1	proposed 2	NMF2D	correct clustering
data (1): 0–10 sec	5.62	-9.05	-0.16	5.94
data (2): 5–15 sec	4.87	-0.71	-0.88	4.88
data (3): 10–20 sec	4.41	-8.88	-0.30	4.45
data (4): 15–25 sec	0.25	-6.23	-2.82	2.52
data (5): 20–30 sec	2.08	-2.86	-1.29	3.34
data (6): 25–35 sec	3.00	-7.82	-0.48	3.66
data (7): 30–40 sec	0.72	-3.17	-1.12	1.48
data (8): 35–45 sec	1.11	-5.71	-3.15	1.42
data (9): 40–50 sec	2.70	-13.13	-1.65	3.93
average	2.75	-6.40	-1.32	3.51

Table 3. SNR results of the evaluated algorithm in dB.

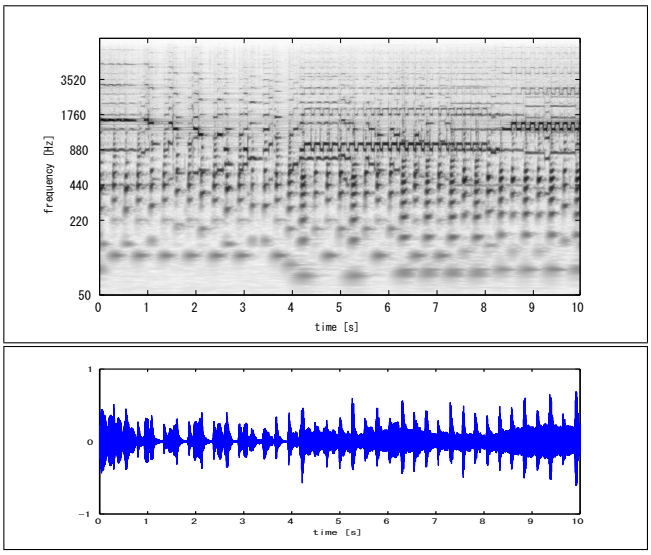

Figure 4. An input signal with three instrumental tracks (flute, oboe, and bassoon). Spectrogram (upper) and corresponding waveform (lower).

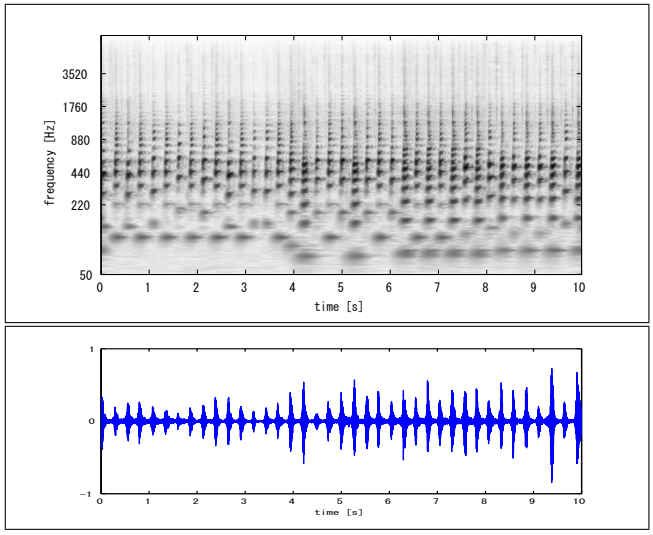

Figure 5. A source signal (bassoon track) of the mixture shown in figure 4.

The average improvement value of correct clustering is 3.54 dB. For two data (data (2) and data (3)) proposed method 1 and correct clustering mark almost same values. It shows that clustering step is maximally effective. In some other data proposed method sets close values to correct clustering.

Comparing SNR values between proposed method 1 and 2, it shows that in clustering step the contribution of the gain vector disjointness is effective.

The SNR values of NMF2D method are lower than that of proposed method 1. The reason is considered to be that, in these real music data, the NMF2D assumption that all notes for an instrument is an identical pitch shifted time-frequency signature does not hold.

Figures 4 , 5 and 6 show an example of experimental results: figure 4 is an input signal in which three instrumental signals (flute, oboe and bassoon) are mixed, figure 5 is a source signal with bassoon sounds, figure 6 is a separated signal which is corresponded to the bassoon's source signal. Even in other two instrumental sounds the results

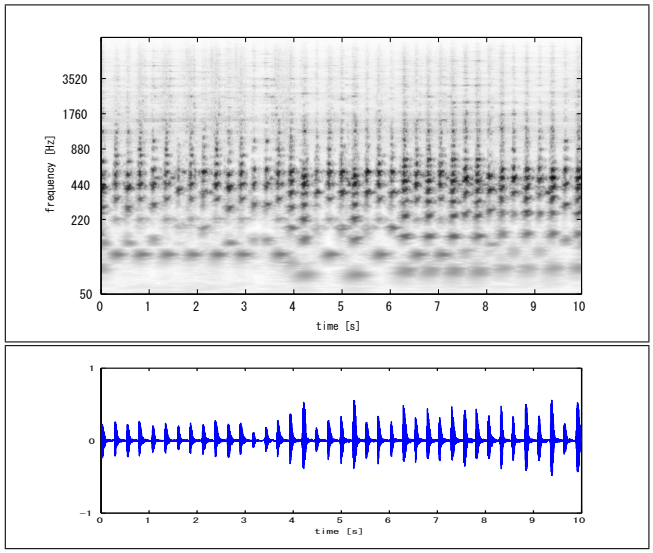

Figure 6. A separated signal (bassoon track) from the mixture shown in figure 4.

equaled to it.

5. CONCLUSION

This paper discussed a method for monophonic instrument sound separation. The method used nonnegative matrix factorization to factorize the spectrogram of the input signal into components. Then we introduced an criterion that measured two distinguish components: basis spectrum similarity and temporal activity disjointness. The grouping was done by clustering components under this measure. The experiment results showed that in some data the proposed method marked values equal to the correct clustering which employed source signals.

Future work includes the improvement of nonnegative matrix factorization by including the proposed criterion, that aims at accuracy enhancement of the decomposition.

6. ACKNOWLEDGEMENT

This research was supported by CrestMuse Project under JST and Grant-in-Aid for Scientific Research (KAKENHI) (A) 20240017.

7. REFERENCES

[1] M. Helén and T. Virtanen: "Separation of Drums from Polyphonic Music using Non-negative Matrix Factorization and Support Vector Machine," in *Proc. ISMIR*, pp. 337–344, 2005.

[2] F. R. Bach and M. I. Jordan: "Blind One-microphone Speech Separation: A Spectral Learning Approach," in *Proc. NIPS*, pp. 65–72, 2005.

[3] P. Leveau, E. Vincent, G. Richard, and L. Daudet: "Instrumentspecific Harmonic Atoms for Mid-level Music Representation," *IEEE Trans. Audio, Speech, and Language Processing*, Vol. 16, No. 1, pp. 116–128, 2008.

[4] M. N. Schmidt and M. Mørup: "Nonnegative matrix Factor 2-D Deconvolution for Blind Single Channel Source Separation," in *Proc. ICA*, pp. 700–707, 2006.

[5] K. Miyamoto, H. Kameoka, T. Nishimoto, N. Ono and S. Sagayama: "Harmonic-Temporal-Timbral Clustering (HTTC) For the Analysis of Multi-instrument Polyphonic Music Signals," in *Proc. ICASSP*, pp. 113–116, 2008.

[6] A. Klapuri, T. Virtanen and T. Heittola: "Sound Source Separation in Monaural Music Signals using Exitation-filter Model and EM Algorithm," in *Proc. ICASSP*, pp. 5510–5513, 2010.

[7] B. Wang and M. D. Plumbley:" Investigating single-channel audio source separation methods based on non-negative matrix factorization, "in *Proc. ICA*, pp. 17–20, 2006.

[8] M. A. Casey and A. Westner: "Separation of Mixed Audio Sources by Independent Subspace Analysis," in *Proc. ICMC*, pp. 154–161, 2000.

[9] D. D. Lee and H. S. Seung: "Algorithms for Non-negative Matrix Factorization," Advances in *Proc. NIPS*, Vol. 13, pp. 556–562, 2000.

[10] P. Smaragdis and J. C. Brown: "Non-negative Matrix Factorization for Polyphonic Music Transcription," in *IEEE Workshop on Applications of Signal Process. Audio Acoust.*, New Paltz, NY, pp. 177–180, 2003.

[11] C. Févotte, N. Bertin, and J.-L. Durrieu: "nonnegative matrix factorization with the Itakura-Saito divergence. With application to music analysis," *Neural Computation*, Vol. 21, No. 3, pp. 793–830, 2009.

[12] C.Uhle, C.Dittmar and T. Sporer: "Extraction of Drum Tracks from Polyphonic Music using Independent Subspace Analysis," in *Proc. ICA*, pp. 843–848, 2003.

[13] P. O. Hoyer: "Non-negative Matrix Factorization with Sparseness Constraints," *J. Mach. Learning Res.*, Vol. 5, pp. 1457–1469, 2004.

[14] T. Virtanen,: "Monaural Sound Source Separation by Non-Negative Matrix Factorization with Temporal Continuity and Sparseness Criteria," *IEEE Transactions on Audio, Speech and Language Process.*, Vol. 15, No. 3, pp. 1066–1074, 2007.

[15] X. Serra: "Musical Sound Modeling with Sinusoids Plus Noise," in *Musical Signal Processing*, C. Roads, S. Popea, A. Picialli and G. D. Poli, Eds. London, U.K.:Swets & Zeitlinger, 1997.

[16] M. Kim and S. Choi: "Monaural Music Source Separation: Nonnegativity, Sparseness, and Shift-invaliance," in *Proc. ICA*, pp. 617–624, 2006.

[17] M. Goto, H. Hashiguchi, T. Nishimura and R. Oka, "RWC music database: music genre database and musical instrument sound database," *Proc. ISMIR*, pp. 229–230, 2003.

[18] MIREX 2007 Evaluation Tasks: "Multiple Fundamental Frequency Estimation & Tracking," http://www.music-ir.org/mirex/2007/index.php/Multiple_Fundamental_Frequency_Estimation_&_Tracking, 2007.