Proceedings of 20th International Congress on Acoustics, ICA 2010

23-27 August 2010, Sydney, Australia

A Study on the Musical Source Separation Method using NMF and Musical Cues

Seokjin Lee, Sang Ha Park, and Koeng-Mo Sung

Applied Acoustics Lab., Institute of New Media and Communications, Seoul National University, Seoul, Republic of Korea.

PACS: 43.60.Dh, 43.60.Hj.

ABSTRACT

Our research is to develop a signal processing process for separating musical audio signals into streams of individual sound sources. In this paper, we present a method that uses NMF method and some musical cues for musical source separation. The conventional separation method based on NMF method classify the separated note events into each stream 'manually', so the conventional methods are difficult to use in the real engineering. However, our method performs 'automatically' the classification process different from the conventional methods. The proposed process consists of the separation step and the reconstruction step. In the separation step, the audio stream is divided into "musical events," groups of the notes which have same frequency structures. The separation method is based on the Wang's method, which used Non-negative Matrix Factorization (NMF) method. In the reconstruction step, the divided note events are automatically grouped into streams of the individual sound sources using the musical cues. The proposed musical cues consist of the timbre features, the temporal features, and the pitch components. The proposed separation system is evaluated with some musical signals which contain the multi musical sources. The evaluation shows the proposed method can perform 'automatically' the separation using the proposed cues and have the same performance as manual classification.

INTRODUCTION

The Blind Source Separation (BSS) problem exists widely in many fields in the acoustical engineering such as speech recognition, automatic music transcription, music informaiont retrival, 3D audio upmixing and so on. The source separation method in the acoustical engineering focused on human perception ability, so the separation process was achieved through computer model known as Computational Auditory Scene Analysis (CASA) [1].

The BSS problem has attracted a great deal of attention and plenty of methods have emerged. There are some crucial features, and the one of them is relation between the number of the given observed mixtures and the number of the sources. The mixtures can be overdetermined, determined or underdetermined with the relation between the numbers. Certain methods require a determined mixture such as most Independent Component Analysis algorithms (ICA) [2]. In the real world, however, there are often underdetermined mixtures, and extremely, one or two obersevations. Therefore, the separation of the underdetermined mixtures seems more useful.

To achieve the separation of the mono-channel musical audio stream, the separation method using Non-negative Matrix Factorization (NMF) was developed [3]. In the separation method, the audio stream separated into basis using the NMF in the time-frequency domain. And using the basis energy, the time-frequency masks are generated and the masks are applied to the spectrogram of the mixtures. After that, the decomposed component is grouped to each source. The method in [3] shows relatively good performance, but the component grouping method is not developed.

In this paper, the mono-channel audio separation method using the NMF is presented, and the musical cues of the decomposed component are studied. To verify the usefulness of the cues to grouping the components into the source streams, we calculate and present the musical cues of the instrumental signals.

THE NMF METHOD

The NMF method is decomposition method solving the following factorization problem:

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

where V is a given M by N non-negative matrix to be decomposed, W and H are non-negative matrix factors. Let the number of the basis is R, and W is a M by R matrix and H is a R by N matrix. The purpose of the NMF algorithm is to find W and H which are minimize the distance between V and WH as

$$D(V,WH) \tag{2}$$

where $D(a,b)$ is a distance function between a and b. If we use the Euclidean distance as the distance measure, the cost function can be defined as

$$C = \|V - WH\| = \sum_{i,j} \left(V_{ij} - (WH)_{ij}\right)^2 . \quad (3)$$

To minimize the cost function, modified gradient algorithm as known as multiplicative update rule is derived by Lee and Seung [4]. The update process by the multiplicative update rule is defined as

$$H_{\alpha\mu} \leftarrow H_{\alpha\mu} \frac{\left(W^T V\right)_{\alpha\mu}}{\left(W^T WH\right)_{\alpha\mu}}, \quad (4)$$

$$W_{i\alpha} \leftarrow W_{i\alpha} \frac{(VH^T)_{i\alpha}}{\left(WHH^T\right)_{i\alpha}}. \quad (5)$$

Setting vector w_r to be the r-th column of W and vector h_r to be the r-th row of the H as

$$W = \{w_1 \quad w_2 \quad \cdots \quad w_R\}, \quad (6)$$

$$H = \{h_1 \quad h_2 \quad \cdots \quad h_R\}^T, \quad (7)$$

then decomposed component v_r can be presented as

$$v_r = w_r h_r, \quad (8)$$

and the v_r is the M by N non-negative matrix. The relation between given matrix V and component v_r is given as

$$V \approx \sum_{r=1}^{R} v_r . \quad (9)$$

THE DECOMPOSITION OF MUSICAL STREAMS USING THE NMF

The NMF method works only for non-negative matrix, but the time-domain musical stream is not non-negative data. However, the magnitude of the spectrogram is non-negative matrix, so the NMF method can be applied to the magnitude of the spectrogram. That is,

$$V(t,\omega) = \left|STFT\{s(t)\}\right|, \quad (10)$$

where $s(t)$ means the mixed signal, and $STFT\{\ \}$ means short-time fourier transform of the signals.

When the NMF method is applied to the magnitude of spectrum as above, the bases in the W matrix, w_r, are features in the frequency domain which can be notes, and the bases in the H matrix, h_r, means the locations in time-domain. In addition, the component v_r, which is multiplication of w_r and h_r, means the note events of mixed signals. The example diagram of the NMF is shown in the Figure 1.

After the factorization as shown above, we can decompose the magnitude spectrogram into the components v_r. However, the components have the magnitude information only. To recover the original audio, the phase information is also needed.

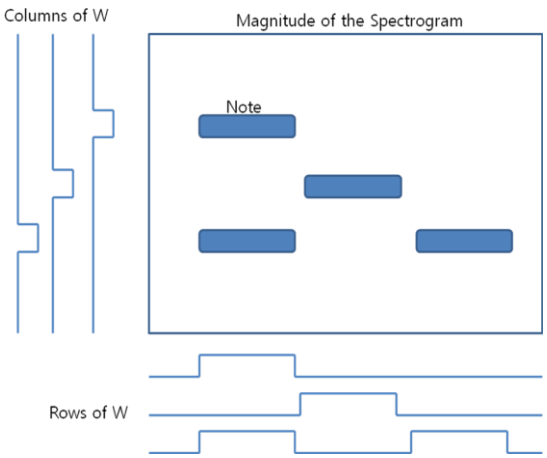

Figure 1. A diagram of the NMF example.

To overcome this problem, Wang and Plumbley [3] suggest a separation method using masks applied to the original spectrogram (not a magnitude spectrogram). The main idea of the suggested method is based on the assumption that over a small time-frequency region, one source dominates.

According to [3], each time-frequency point among all the bases is compared, and then masks for each component are generated. A point of the mask will be marked as 1 if it has maximum value among all the bases at the same position, otherwise it will be marked as 0. The mask generation process can be illustrated as:

$$M_{ij}^{(k)} = \begin{cases} 1 & if \quad k = \arg\max_r \left(v_r\right)_{ij} \\ 0 & others \end{cases}, \quad (11)$$

where $M^{(k)}$ is the mask for the basis k. By applying the masks to the original spectrogram, the component corresponding to each basis is decomposed.

MUSICAL CUES FOR AUTOMATIC CLUSTERING OF THE DECOMPOSED COMPONENTS

The decomposed component has information about note events, but the component does not equal to a source signal. In general cases, the number of the components (and it is equal to the number of the bases) is larger than the number of sound sources. To reconstruction each source signal, the components–which are correspond to the note events–must be clustered into source groups. In the previous method [3], the clustering process performed manually, however, the manual clustering is nearly impossible in the real systems. Therefore, a study on the automatic clustering process is needed.

We consider a clustering method using some musical cues to the automatic clustering process. To develop the clustering method, we perform a study on following musical cues which can be used to clustering process.

The spectral flatness and pitch-normalized flatness

The spectral flatness is defined as the ratio between the geometric and arithmetic mean of the estimated power spectrum [5]. While the spectral flatness is generally calculated with the fourier transform of signals, we calculate with w_r because it has spectral informations of the decomposed components.

$$F_r = \frac{\sqrt[N]{\prod_{k=1}^{N}|w_r(k)|^2}}{\frac{1}{N}\sum_{k=1}^{N}|w_r(k)|^2} \quad (12)$$

The spectral flatness is between zero and one, because the arithmetic mean of a non-negative values is always greater than its geometric mean. The pitch-normalized spectral flatness can be obtained by dividing the spectral flatness with its pitch.

The spectral centroid and pitch-normalized centroid

The spectral centroid is defined as the 'center of gravity' of the magnitude spectrum, reflect in its 'brightness' [5]. We calculate the spectral centroid with w_r because of the same reason as the spectral flatness.

$$C_r = \frac{\sum_{k=1}^{N/2} k|w_r(k)|}{\left(\frac{N}{2}+1\right)\sum_{k=1}^{N/2}|w_r(k)|} \quad (13)$$

The pitch-normalized spectral centroid is obtained by dividing the spectral centroid with corresponding pitch.

The spectral kurtosis

The kurtosis is well known as a measure from higher-order statistics that is based on the fourth and second-order moments of the signal [5].

$$K_r = \frac{\frac{1}{N}\sum_{k=1}^{N}\left(w_r(k)\right)^4}{\left(\frac{1}{N}\sum_{k=1}^{N}\left(w_r(k)\right)^2\right)^2} \quad (14)$$

The audio noise-likeness index

The audio noise-likeness index is derived from the frequency vector of the components [6]. The noise-likeness index is defined as ratio of spectral pitch tone power to the average power of the rest.

$$ANLI_r = \frac{\left(\sum_{k=1}^{N} w_r(k) - \max\left(w_r(k)\right)\right)\Big/\left(N-1\right)}{\max\left(w_r(k)\right)} \quad (15)$$

The mel-frequency cepstral coefficient

The mel-frequency cepstrum is a representation of the short-term power spectrum of a sound, based on a linear cosine transform of a log power spectrum on a nonlinear mel scale of frequency. And mel-frequency cepstral coefficient (MFCC) is coefficients that collectively make up a mel-frequency cepstral (MFC) [5].

In our analysis, the MFCCs are calculated with w_r. First, map the powers of the w_r onto the mel scale, using triangular overlapping windows. And then, take the logs of the powers at each of the mel frequencies. Finally, take the discrete cosine transform of the list of mel log powers.

The coefficients of the auto-correlation function

The auto-correlation coefficient is a sort of normalized auto-correlation [5]. In our research, the auto-correlation coefficients are calculated with time envelope, h_r.

$$ACF_r(a) = \frac{\sum_{k=a}^{N-1} h_r(k-a)h_r(k)}{\sum_{k=0}^{N-1} h_r(k)^2} \quad (16)$$

The percussiveness

The percussiveness feature is extracted from the time-domain amplitude envelope. To calculate the percussiveness, the percussive impulse template is modelled first. This model template is convoluted with local maxima of the amplitude envelope. And then, the correlation coefficient of the convolved model and the original envelope is calculated. The calculated correlation coefficient represents the degree of the percussiveness [7].

The peak time and the peak fluctuation

The peak time means an average length in the time envelope and peak fluctuation is a deviation between the length of these peaks. The peak is defined as an area where the time envelope is above a threshold of 0.8*maximum [8: drum NMF].

The perceptual linear prediction coefficients

The perceptual linear prediction (PLP) coefficient is a kind of linear prediction coeffiecient (LPC) which consider a human perception property. The PLP coefficient calculation process consists of the four steps: critical-band analysis, equal-loudness preemphasis, intensity-loudness conversion, and calculation of the LPC.

To calculate the PLP coefficient, the short-time power spectrum $P(\omega)$ is warped along its frequency axis ω into Bark frequency Ω by [9 : PLP]

$$\Omega(\omega) = 6\ln\left\{\omega/1200\pi + \left[\left(\omega/1200\pi\right)^2 + 1\right]^{0.5}\right\} \quad (17)$$

After that, the power spectrum in the Bark frequency domain $P(\Omega)$ is convoluted with critical-band masking curves $\Psi(\Omega)$ into $\Theta(\Omega)$. Next, The calculated $\Theta(\Omega)$ is preemphasized by the simulated equal-loudness curve $E(\omega)$ as

$$\Xi\left(\Omega(\omega)\right) = E(\omega)\Theta\left(\Omega(\omega)\right). \quad (18)$$

The last operation prior to the all-pole modelling (which means linear prediction) is the cubic-root amplitude compression

$$\Phi(\Omega) = \Xi(\Omega)^{0.33}. \quad (19)$$

Finally, the PLP coefficients can be obtained by all-pole modelling of $\Phi(\Omega)$. In our research, the frequency component w_r is used instead of the power spectrum $P(\omega)$.

Table 1. The values of the musical cues of each source.

Musical cues	Piano		Cello	
	Average	Std.	Average	Std.
Flatness	0	0	0	0.0002
Flatness_n	0	0	0	0
Centroid	0.0698	0.0167	0.1028	0.0374
Centroid_n	0.0002	0.0001	0.0005	0.0002
Kurtosis	461.26	119.65	383.92	128.67
ANLI	0.0032	0.0012	0.0050	0.0026
MFCC_{max}	17.640	2.0404	20.597	2.9373
MFCC_{avg}	2.0460	0.6078	2.5595	0.4945
ACF	1.0918	1.5677	0.0470	0.0920
Percuss.	0.5068	0.1359	0.5544	0.1036
Peak time	1.8120	1.1786	2.0103	1.8957
Peak fluc.	0.3878	0.7148	0.3099	0.4712
MFCC_{norm}	503.27	66.040	533.26	88.487
DPLP	0.0462	0.0235	0.0234	0.0118

EXPERIMENTS

To verify the usefulness of the musical cues, we performed experiments with some piano and cello music. The 15 piano music pieces including partita and fugue are used, and the 10 cello music pieces are used for our experiments. The total 153 piano notes and 101 cello notes are analysed.

Most of the musical cues applied as shown above, but the ACF and PLP are slightly modified. In the case of the ACF, we analysed the time lag which maximize the ACF, and in the case of the PLP, we analysed the norm of the derivative of the PLP as following:

$$DPLP_r = \frac{1}{N_a - 1}\sum_{k=1}^{N_a}\left|PLP_r(k) - PLP_r(k-1)\right|^2, \quad (20)$$

where N_a is an order of the PLP model.

The experimental result is shown in the Table 1 and 2. As shown in the Table 1, the pitch-normalized centroid, the maximum MFCC, the time-lag which maximizes the ACF, and the squared derivative of the PLP (DPLP) can be used to separate the piano and cello signal. Table 2 shows the result divided into the low-frequency band and the high-frequency band. As shown in the Table 2, some cues such as pitch-normalized centroid are more useful if the frequency-band is considered.

From Figure 2 to Figure 5 show distribution of each feature value. The x-axis means the sample number to discriminate each note event, and the y-axis means the value of each cue. Therefore, the distribution along the y-axis is important imformation. The blue circle means the piano data, and the red square means the cello data. And the x-marked value inside the circle or the square means that the data is in a high frequency band. The blue dashed line means the average value of the piano data, and the red solid line means the average value of the cello data. As shown in the distribution diagrams, the pitch-normalized centroid and DPLP show good properties, but it is hard that any feature is used solely. Therefore, it is expected that the performance is good when three or four features are used together.

In the Figure 6, an example of the separation result by hard threshoud is shown. (a) is a waveform of the original piano source, (b) is a waveform of the original cello source. (c) is a artificial mixture of (a) and (b), (d) and (e) is a separation result. As shown in Figure 6, the separation and categorization process works properly, but there are some errors with separation result.

Table 2. The musical cues correspond to frequency band.

Freq.-band	Musical cues	Piano		Cello	
		Avg.	Std.	Avg.	Std.
Low Freq.	Flatness	0	0	0.0001	0.0002
	Flatness_n	0	0	0	0
	Centroid	0.0617	0.0131	0.0976	0.0363
	Centroid_n	0.0002	0.0001	0.0006	0.0002
	Kurtosis	432.22	134.47	353.62	131.23
	ANLI	0.0035	0.0014	0.0054	0.0028
	MFCC_{max}	17.441	1.6607	20.542	3.1906
	MFCC_{avg}	2.2325	0.6742	2.8101	0.3800
	ACF	1.0367	1.3642	0.0456	0.0936
	Percuss.	0.5284	0.1283	0.5602	0.0983
	Peak time	2.2147	1.4074	2.0837	1.2685
	Peak fluc.	0.6261	0.9084	0.3555	0.4944
	MFCC_{norm}	505.88	59.762	536.04	96.600
	DPLP	0.0494	0.0269	0.0237	0.0124
High Freq.	Flatness	0	0	0	0
	Flatness_n	0	0	0	0
	Centroid	0.0778	0.0159	0.1092	0.0377
	Centroid_n	0.0001	0.0000	0.0003	0.0001
	Kurtosis	490.30	94.139	421.44	114.83
	ANLI	0.0029	0.0010	0.0046	0.0024
	MFCC_{max}	17.838	2.3430	20.666	2.5879
	MFCC_{avg}	1.8596	0.4635	2.2493	0.4411
	ACF	1.1496	1.7459	0.0488	0.0899
	Percuss.	0.4852	0.1399	0.5451	0.1090
	Peak time	1.4093	0.6879	1.9194	2.4568
	Peak fluc.	0.1495	0.2882	0.2533	0.4343
	MFCC_{norm}	500.66	71.676	529.81	77.124
	DPLP	0.0431	0.0190	0.0231	0.0110

The main cause of the errors is as following: If there is some region of the mixture spectrogram where the power of each source is about the same – it means that any source do not dominant, the decomposed component corresponding to the region has information about both sources – piano and cello here. And the values of the musical cues are dubious. To solve this problem, further research about the better decompsotion method is needed.

CONCLUSION

In this paper, a study on the the musical source separation system including the automatic categorization is performed. To compose the automatic categorization process, several musical cues are applied and evaluated. The piano music samples and the cello music samples are used to evaluate the musical cues. As a result of the study, the fact that some of the musical cues may be useful to the musical source separation system is revealed.

There are some future works to enhance the musical source separation system. One of that is enhancement of the component decomposition system, and another task is developement of the categorization method which use the several musical cues together.

REFERENCE

1 D.P.W.Ellis, *Prediction-driven Computational Auditory Scene Analysis*. Ph.D. thesis, Department of Electronic Engineering and Computer Science, MIT, 1996.

2 P. Comon, "Independent component analysis – a new comcept?, " *Signal Processing*, 36(3), 287-314 (1994).

3 Beiming Wang and Mark D. Plumbley, "Musical Audio Stram Separation by Non-negative Matrix Factorization," *Proc. DMRN summer conf.* (2005).

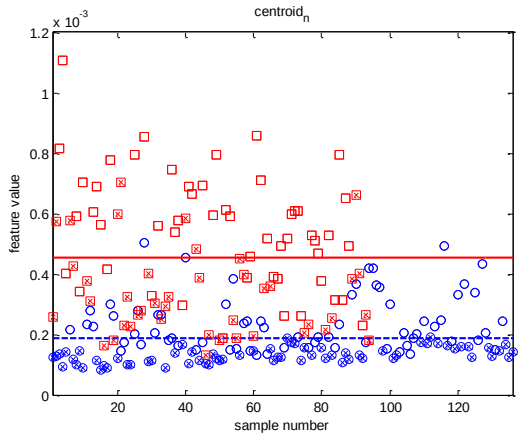

Figure 2. The result of the pitch-normalized centroid.

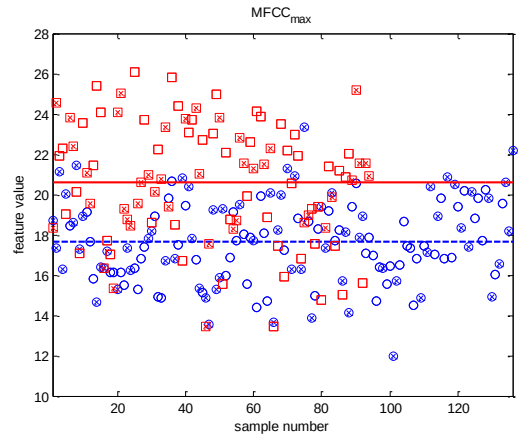

Figure 3. The result of the maximum value of the MFCC.

4 D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization" in *Advances in Neural Information Processing 13 (Proc. NIPS*2000)*, MIT Press (2001).

5 Erik Larsen and Ronald M. Aarts, Audio Bandwidth Extension (Wiley, Chichester, 2004) pp. 211-214.

6 Vladimir Soutchiline, "On Audio Noise Likeness-Index Evaluation," *Proc. the 99th AES Convention* (1995).

7 Chistian Uhle, Christian Dittmar, and Thomas Sporer, "Extraction of Drum Traks from Polyphonic Music Using Independent Subspace Analysis," *4th International Symposium on Independent Component Analysis and Blind Signal Separation (ICA 2003)* (2003).

8 Marko Helen and Tuomas Virtanen, "Separation of Drums from Polyphonic Music using Non-negative Matrix Factorization and Support Vector Machine," *Proc. EUSIPCO* (2005).

9 Hynek Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *J. Acoust. Soc. Am.* (1989).

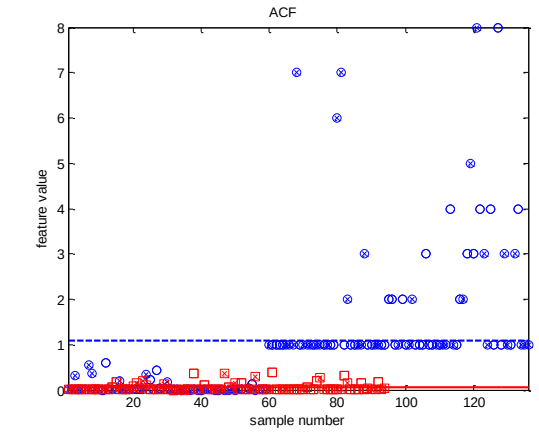

Figure 4. The result of the ACF.

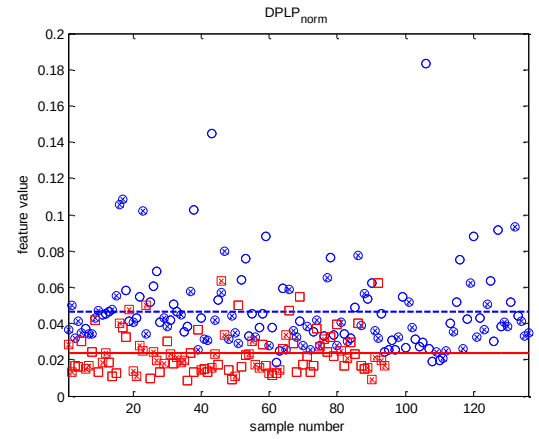

Figure 5. The result of the DPLP.

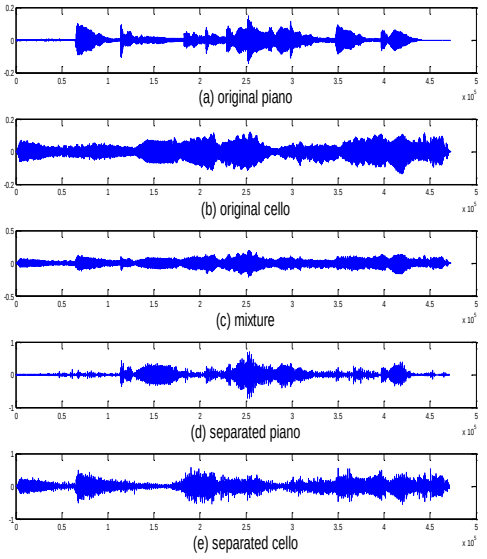

Figure 6. The result of the separation process with automatic categorization system.