Harmonic and Percussive Separation Based on NMF and Tonality Mask

Keunwoo Choi, Sang Bae Chon, and Kyeongok Kang

In this letter, we present a new algorithm for the harmonic and percussive separation of jazz music. Using a short-time Fourier transform and nonnegative matrix factorization, the signal is decomposed into rank components. Each component is then split into harmonic and percussive parts using masks calculated based on their tonalities. Finally, the harmonic and percussive parts are separated after applying the masks and a summation. We evaluate the algorithm based on real audio examples using both objective and subjective assessments. The proposed algorithm performs well for the separation of harmonic and percussive parts of jazz excerpts.

Keywords: Music information retrieval, source separation.

I. Introduction

There has been a significant amount of research on the problem of source separation of music signals, which leads to a diverse range of applications, interactive music service [1], and so on. In particular, this letter introduces an algorithm for harmonic and percussive separation.

For source separation, nonnegative matrix factorization (NMF) has been recognized as a suitable machine learning technique in much previous research since [2]. Particularly for harmonic-percussive separation, significant research was conducted in [3], wherein Helen and Virtanen proposed an NMF-based approach with a support vector machine, and in [4], wherein Kim and others adopted nonnegative matrix partial co-factorization.

Manuscript received Mar. 7, 2012; revised July 15, 2012; accepted Aug. 27, 2012.

This work was supported by the Ministry of Knowledge Economy grant funded by the Korea government (No. 10037244).

Keunwoo Choi (phone: +82 42 860 0767, gnu@etri.re.kr) and Kyeongok Kang (kokang@etri.re.kr) are with the Broadcasting & Telecommunications Convergence Research Laboratory, ETRI, Daejeon, Rep. of Korea.

Sang Bae Chon (strlen03@gmail.com) is with the DMC Research Center, Samsung Electronics, Rep. of Korea.

http://dx.doi.rog/10.4218/etrij.12.0212.0107

On the other hand, there has been other research that has focused on the property of the time-frequency domain. Ono and others introduced a separation algorithm based on the assumptions of a spectrogram, whose smooth temporal envelopes with parallel ridges can be considered harmonic parts and wideband spectral envelopes concentrated within a short time period can be considered percussive instruments [5]. Fitzgerald introduced a faster harmonic and percussive separation algorithm using a median filter based on the same assumption [6]. While those algorithms provide a good separation performance for pop music, they are inappropriate for jazz music, which consists of the frequent use of ride cymbals and brushes, as such sounds result not in impulsive envelopes but noise-like energy distributions.

This letter, therefore, introduces a new harmonic and percussive separation algorithm for jazz music. The algorithm is based on both NMF and tonality masks. In section II, the details of the algorithm are introduced. Experiments and the results are presented in section III, and the conclusion follows in section IV.

II. Separation Algorithm

Using NMF, the audio signal is first decomposed into multiple bases. Then, two kinds of masks, that is, harmonic and percussive, are generated depending on the tonality of each spectral band and applied to every base. Figure 1 describes a

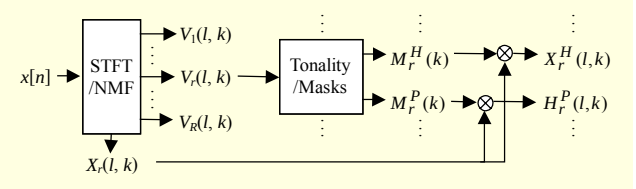

Fig. 1. Block diagram of proposed algorithm.

block diagram of the proposed algorithm explained so far. By grouping the partial bands of the components, the harmonic and percussive parts are then separated.

1. Extraction of Bases Using NMF

The NMF algorithm introduced in [7] decomposes an *L*-by-*K* nonnegative matrix **V** into an *L*-by-*R* matrix **W** and *R*-by-*K* matrix **H**, which yield **V**≈**W**·**H**. The decomposition is performed in a way that minimizes the approximated reconstruction error between **V** and **W**·**H**.

Given that **V** is the magnitude for short-time Fourier transformed spectrum **X** of input signal $x(n)$, **W** and **H** come to represent the frequency and time envelope, respectively.

In the calculation of NMF, the decomposed matrices **W** and **H** are updated using Lee and Seung's multiplicative update algorithm [7].

As a distance measure of the reconstruction error, the Kullback-Leibler divergence, defined as (1), is used.

$$D_{KL}(\mathbf{V},\mathbf{WH}) = \sum_{l=1}^{L}\sum_{k=1}^{K}\left(v_{lk}\log\frac{v_{lk}}{(wh)_{lk}} - v_{lk} + (wh)_{lk}\right), \quad (1)$$

where v_{lk} and $(wh)_{lk}$ represent the elements at the *l*-th row and *k*-th column, that is, the *k*-th frequency component at the *l*-th time frame of matrix **V** and **W**·**H**, respectively. Here, *l*, *k*, and *r* represent the time frame index, frequency index, and basis index to be decomposed, respectively. After calculating **W** and **H**, the element of base $\mathbf{V}_r$ is computed as (2).

$$(v_r)_{lk} = w_{lr}\times h_{rk}\,. \quad (2)$$

The components decomposed using the NMF are considered the bases of a mixed signal. As a result, it provides temporal and spectral information, which will be utilized when calculating the tonality. However, NMF does not provide an identically separated signal for the following reasons. First, the spectrum of the basis yielded by NMF tends to be *fragments* of a single note spectrum or a *mixture* of note spectrums rather than one complete note spectrum. Second, interference exists, which is an inevitable problem of the *separate and detect* approach, as pointed out in [8]. To solve these problems, an additional mask is applied to classify each spectral band into harmonic or percussive parts, depending on their respective tonalities.

2. Tonality Calculation of Decomposed Components

To separate the harmonic and percussive components, we find that the tonality introduced in [9] is an appropriate analysis parameter, as the percussive parts tend to be more noise-like than the harmonic parts tend to be.

To estimate the tonality, the chaos measure matrix $\mathbf{c}_r$ is calculated first, in which the chaos measure is defined as the prediction error in the complex domain. Then, the tonality matrix $\mathbf{T}_r$ is calculated using a limitation function and logarithmic mapping of the chaos measure as (3) and (4) (for more details, see [9]). As a result, the tonality in [9] is a property of a *time-frequency bin* $X(l,k)$, describing how predictable the bin is in terms of amplitude and phase.

$$\mathbf{c}_{rL}(l,k) = \max\left(0.05, \min\left(0.5, \mathbf{c}_r(l,k)\right)\right). \quad (3)$$

$$\mathbf{T}_r(l,k) = -0.43\times\log_{10}\left(\mathbf{c}_{rL}(l,k)\right) - 0.299\,. \quad (4)$$

We find that within each basis of NMF, a spectral band (with whole length in time) $X(l, :)$ is a proper unit for the tonality calculation, rather than a time-frequency bin $X(l,k)$. This approach assumes that either the harmonic or percussive part dominates even though the basis of NMF is still a mixture of both parts. For a tonality calculation of the spectral bands, therefore, we propose the normalized weighted average of the tonality based on the magnitude, shown in (5), to take the significance of each bin into account. Note that the tonality matrix of the *r*-th basis $\mathbf{T}_r$ becomes vector T_r with frequency index *k* after the normalized weighted average.

$$T_r(k) = \left(\sum_{l=1}^{L}\mathbf{T}_r(l,k)\times\mathbf{V}_r(l,k)\right)\Bigg/\left(\sum_{l=1}^{L}\mathbf{V}_r(l,k)\right). \quad (5)$$

3. Separation Using a Mask

Based on the tonality, some of the basis matrix $\mathbf{V}_r$ bands are determined to be harmonic, while the rest are determined to be percussive. As a decision function, binary-type masks are used in the proposed algorithm. The masks used for the harmonic parts, $M_r^H(k)$, and percussive parts, $M_r^P(k)$, of the *r*-th basis are defined as (6) and (7), respectively.

$$M_r^H(k) = \begin{cases} 1 & \text{if } T_r(k) \geq threshold, \\ 0 & \text{otherwise,} \end{cases} \quad (6)$$

$$M_r^P(k) = 1 - M_r^H(k)\,. \quad (7)$$

The harmonic and percussive spectra $\mathbf{X}^H$ and $\mathbf{X}^P$ are synthesized using (8), applying the masks for the respective bases, the summation of the harmonic or percussive parts, and the use of the phase information of the mixture.

$$\mathbf{X}^{H(or\ P)}(l,k) = \left(\sum_{r=1}^{R}\mathbf{V}_r(l,k)\times M_r^{H(or\ P)}(k)\right)\angle\mathbf{X}(l,k)\,. \quad (8)$$

Finally, time domain signals $x^H(n)$ and $x^P(n)$ are reconstructed using an inverse short-time Fourier transform.

III. Experiments and Discussion

In our evaluations, the median-filter-based separation algorithm proposed in [6] is implemented using the same

parameters and is compared with the proposed algorithm. Note that it assumes a high impulsiveness for percussive sources, whereas the proposed algorithm focuses on the noise-like characteristics, which shows that the performance of a jazz excerpt is different from that of a non-jazz excerpt.

1. Outline

In the experiments, 44.1 kHz sampled signals with a length of 20 seconds are used as test excerpts. Jz1 is "A Cat" by Casual Visit, and Jz2 is mixed by the authors using samples on Apple Logic Pro 9. NJz1 and NJz2 are "Mixtape" and "Wreck," respectively, from BASS-dB [10].

In the implementation of the proposed algorithm, the input signal is analyzed using a Fourier transform with a frame length of 2,048, double zero padding, and an overlap factor of 0.875. These parameter settings provide sufficient information, a frequency resolution of 10.8 Hz, and a time resolution of 6 ms. The rank R of NMF is set to 40, which is large enough to represent the notes in the test excerpts, and $\mathbf{W}$ and $\mathbf{H}$ are updated for seventy iterations. The threshold of the tonality is set empirically to 0.

Figure 2 illustrates the true source and its estimation by the proposed and compared algorithms of excerpt Jz1. It can be seen in Fig. 2(f) that the noise-like components, for example, the ride cymbals and the brush-rolled snare, are separated into the harmonic part for the compared algorithm. On the other hand, Fig. 2(e) shows that the proposed algorithm does not miss such components, extracting the ride cymbals. For the hi-hats, the peaks are slightly clearer in the compared algorithm. The separation results can be heard at *http://keunwoochoi.blogspot.com/p/sound-examples.html*.

2. Objective Evaluation

In an objective evaluation, we measure the standard signal-to-distortion ratio (SDR), signal-to-interference ratio (SIR), and signal-to-artifact ratio (SAR), which were introduced in [11]. The median values are calculated on local frames [12].

The objective evaluation results are presented in Table 1, in which values for all pairs of the better evaluated algorithm are in bold. Table 2 summarizes Table 1 as the ratio of winning items for each genre and algorithm, in which the total comparison includes 12 stands for 2×3×2, two excerpts (Jz1 and Jz2 or NJz1 and NJz2), three metrics (SDR, SIR, and SIR), and two parts (harmonic and percussive). According to Table 2, the proposed algorithm shows a better performance than that of the compared algorithm in 75% of the jazz excerpts (9 out of 12). In particular, the SIR shows the most advanced performance of all criteria, 6.8 dB on average. On the other hand, it shows a worse performance for non-jazz excerpts in 66% of the pairs compared (8 out of 12).

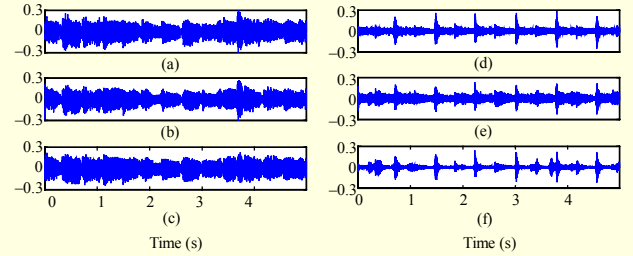

Fig. 2. Results of Jz1 for harmonic part (left) and percussive part (right), respectively: (a),(d) true source, (b),(e) estimation using proposed algorithm, and (c),(f) estimation using compared algorithm.

Table 1. Evaluation results of proposed (P) and compared (C) algorithms for jazz (Jz) and non-jazz (NJz) excerpts.

ID		Harmonic part (dB)			Percussive part (dB)		
		SDR	SIR	SAR	SDR	SIR	SAR
Jz1	P	12.6	**29.9**	12.8	**0.4**	**7.6**	**3.5**
	C	**14.2**	20.1	**16.5**	–1.0	5.8	2.9
Jz2	P	**15.6**	**30.6**	**16.3**	**–2.2**	**9.7**	0.2
	C	12.8	26.9	14.5	–7.5	–1.4	**1.2**
NJz1	P	7.6	20.0	9.2	–6.5	**–1.0**	0.8
	C	**10.3**	**21.1**	**11.3**	**–6.1**	–2.5	**3.8**
NJz2	P	7.2	**20.8**	8.5	–5.7	**3.7**	1.0
	C	**8.6**	14.9	**11.1**	**–3.6**	–2.1	**4.3**

Table 2. Comparison between proposed and compared algorithms based on objective evaluation.

	Performance achievement	
	Jazz	Non-jazz
Proposed	75% (9/12)	33% (4/12)
Compared	25% (3/12)	67% (8/12)

3. Subjective Evaluation

In a subjective test, seven experienced listeners from Seoul National University and the Electronics and Telecommunications Research Institute (ETRI), both of the Republic of Korea, participate. Experiments are conducted using Sennheiser HD650 reference headphones.

Using an integer grading scale of 1 to 5, the listeners are asked to evaluate the overall quality of separation, including interference and artifacts, compared with the reference signal (true source). The process and interface used for the listening

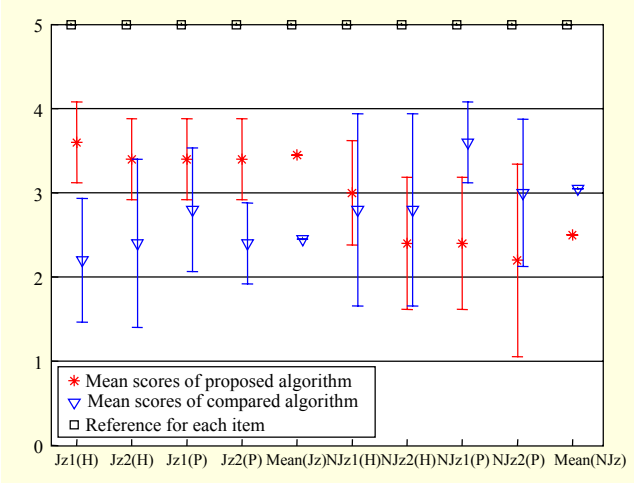

Fig. 3. Result of subjective evaluation. H and P indicate harmonic and percussive parts. Bar represents 95% confidence interval. Mean(Jz) and Mean(NJz) indicate total average score of jazz and non-jazz excerpts, respectively.

test are equivalent to ITU-R BS.1116 [13].

Figure 3 plots the results of the subjective evaluation, the mean, and the 95% confidence interval. For the jazz excerpts, the proposed algorithm acquires 3.4 on average, while the compared algorithm acquires 2.4 on average. These results agree with the objective evaluation results and indicate that the proposed algorithm performs better than the compared algorithm for jazz music.

IV. Conclusion

We introduced an algorithm for the blind separation of harmonic and percussive parts based on NMF and tonality, on the assumption that the tonality would be high in the harmonic part and low in the percussive part. We performed NMF to obtain the separated bases and then calculated their tonality and respective masks. The harmonic and percussive parts were acquired after applying the masks and summations.

During the experiments, the proposed algorithm showed a better performance for jazz excerpts in both objective and subjective evaluations. As a result, we show that the tonality can be used as a measure of percussive sources.

Future works will aim to increase the separation performance, reducing the calculation complexity and extending the algorithm to deal with various signals. The perceived quality of source separation will also be studied.

Acknowledgment

The authors are grateful to Dr. Hwan Shim and Casual Visit for sharing their song ‘A Cat’ as a test excerpt.

References

[1] I. Jang, J. Seo, and K. Kang, “File Format Design for Interactive Music Service,” *ETRI J.*, vol. 33, no. 1, Feb. 2011, pp. 128-131.

[2] P. Smaragdis and J.C. Brown, “Non-negative Matrix Factorization for Polyphonic Music Transcription,” *Appl. Signal Process. Audio Acoustics, IEEE Workshop*, 2003, pp. 177-180.

[3] M. Helen and T. Virtanen, “Separation of Drums from Polyphonic Music Using Non-negative Matrix Factorization and Support Vector Machine,” *European Signal Process. Conf.*, 2005.

[4] M. Kim et al., “Nonnegative Matrix Partial Co-Factorization for Spectral and Temporal Drum Source Separation,” *IEEE J. Sel. Topics Signal Process.*, vol. 5, 2011, pp. 1192-1204.

[5] N. Ono et al., “Separation of a Monaural Audio Signal into Harmonic/Percussive Components by Complementary Diffusion on Spectrogram,” *European Signal Process. Conf.*, Lausanne, Switzerland, 2008.

[6] D. Fitzgerald, “Harmonic/Percussive Separation Using Median Filtering,” *Int. Conf. Digital Audio Effects*, Graz, Austria, 2010.

[7] D.D. Lee and H.S. Seung, “Algorithms for Non-negative Matrix Factorization,” *Advances Neural Inf. Process. Syst.*, vol. 13, 2001.

[8] O. Gillet and G. Richard, “Transcription and Separation of Drum Signals from Polyphonic Music,” *IEEE Trans. Audio, Speech, Language Process.*, vol. 16, 2008, pp. 529-540.

[9] K. Brandenburg and J.D. Johnston, “Second Generation Perceptual Audio Coding: The Hybrid Coder,” *Audio Eng. Soc. (AES) Conv.*, 1990.

[10] R.G.E. Vincent et al., “BASS-dB: The Blind Audio Source Separation Evaluation Database.” Available: http://bass-db.gforge.inria.fr/BASS-dB

[11] E. Vincent, R. Gribonval, and C. Fevotte, “Performance Measurement in Blind Audio Source Separation,” *IEEE Trans. Audio, Speech, Language Process.*, vol. 14, 2006, pp. 1462-1469.

[12] E. Vincent, “Musical Source Separation Using Time-Frequency Source Priors,” *IEEE Trans. Audio, Speech, Language Process.*, vol. 14, 2006, pp. 91-98.

[13] International Telecommunication Union, “Methods for the Subjective Assessment of Small Impairments in Audio Systems Including Multi-channel Sound Systems,” Rec. ITU-R BS.1116, Geneva, Switzerland, 1994.