HARMONIC-PERCUSSIVE SOURCE SEPARATION USING HARMONICITY AND SPARSITY CONSTRAINTS

Jeongsoo Park, Kyogu Lee
Music and Audio Research Group
Seoul National University, Seoul, Republic of Korea
{psprink, kglee}@snu.ac.kr

ABSTRACT

In this paper, we propose a novel approach to harmonic-percussive sound separation (HPSS) using Non-negative Matrix Factorization (NMF) with sparsity and harmonicity constraints. Conventional HPSS methods have focused on temporal continuity of harmonic components and spectral continuity of percussive components. However, it may not be appropriate to use them to separate time-varying harmonic signals such as vocals, vibratos, and glissandos, as they lack in temporal continuity. Based on the observation that the spectral distributions of harmonic and percussive signals differ – *i.e.*, harmonic components have harmonic and sparse structure while percussive components are broadband – we propose an algorithm that successfully separates the rapidly time-varying harmonic signals from the percussive ones by imposing different constraints on the two groups of spectral bases. Experiments with real recordings as well as synthesized sounds show that the proposed method outperforms the conventional methods.

1. INTRODUCTION

Recently, musical signal processing has received a great deal of attention especially with the rapid growth of digital music sales. Automatic musical feature extraction and analysis for a large amount of digital music data has been enabled with the support of computational power. The major purposes of such tasks include extracting musical information such as melody extraction, chord estimation, onset detection, and tempo estimation.

Because most music signals often consist of both harmonic and percussive signals, the extraction of tonal attributes is often severely degraded by the presence of percussive interference. On the other hand, when we analyze rhythmic attributes such as tempo estimation, the harmonic signals act as interference that may prevent accurate analysis. Consequently, the separation of harmonic and percussive components in music signals will function as an important pre-processing step that allows efficient and precise analysis.

For these reasons, many researchers have focused on investigating HPSS using various approaches. Uhle *et al*. performed singular value decomposition (SVD) followed by independent component analysis (ICA) to separate drum sounds from the mixture [1]. Gillet *et al*. presented a drum-transcription algorithm based on band-wise decomposition using sub-band analysis [2].

Other researchers have employed matrix factorization techniques such as non-negative matric factorization (NMF). Helen *et al*. proposed a two-stage process composed of a matrix-factorization step and a basis-classification step [3]. Kim *et al*. employed the matrix co-factorization technique, where spectrograms of the mixture sound and drum-only sound are jointly decomposed [4]. NMF with smoothness and sparseness constraints was utilized by Canadas-Quesada *et al*. [5]. The algorithm was developed based on assumptions regarding the anisotropic characteristics of the harmonic and percussive components; harmonic components have temporal continuity and spectral sparsity, whereas percussive components have spectral continuity and temporal sparsity.

Most HPSS algorithms have employed the same assumption. Ono *et al*. presented a simple technique to represent a mixture sound spectrogram as a sum of harmonic and percussive spectrograms based on the Euclidean distance [6]. Their technique aims to minimize the temporal dynamics of harmonic components and the spectral dynamics of percussive components. They further extended their work to use an alternative cost function based on the Kullback-Leibler (KL) divergence [7]. More recently, FitzGerald presented a median filtering-based algorithm [8], where a median filter is applied to the spectrogram in a row-wise and column-wise manner for the extraction of harmonic and percussive sounds, respectively. Gkiokas *et al*. also proposed a non-linear filter-based HPSS algorithm [9].

However, the assumption regarding the temporal continuity, which is considered to be crucial for conventional harmonic-percussive studies, does not account for the rapidly time-varying harmonic signals often present in vocal sounds and musical expressions such as slides, vibratos, or glissandos. This is because their spectrograms often fluctuate over short periods of time. Thus, it may degrade the performance of the algorithms, particularly when

loud vocal components or such musical expressions are mixed.

In this paper, we propose a HPSS algorithm that is classified as a spectrogram decomposition-based method. We consider the spectrum of harmonic components to have a harmonic and sparse structure in the frequency domain, whereas the spectrum of percussive components to have an unsparse structure. To realize the successful separation of harmonic/percussive sounds, we apply constraints that impose a particular structure of the spectral bases. The novelty of the proposed method resides in the harmonicity constraint, which is an extension of the sparsity constraint presented in previous works [10]. The constraint is closely related to the Dirichlet prior, which is frequently used in probabilistic analysis. Because the proposed algorithm does not assume temporal continuity for the separation of harmonic signals, we can successfully separate harmonic signals from the mixture sound, even when there are significant fluctuations over time.

The rest of this paper is organized as follows. Section 2 explains in detail how the proposed method works. In Section 3, we present experimental results, and in Section 4, we conclude the paper.

2. PROPOSED METHOD

In this section, we present a detailed explanation of the proposed HPSS method. The proposed algorithm uses the spectrogram-decomposition technique, NMF, with the harmonicity and sparsity constraints based on the Dirichlet prior. For the efficient description of the proposed method, we first introduce the conventional NMF. Then, the algorithm description for the proposed method is presented. Finally, the theoretical relations of the proposed method to the Dirichlet prior are described.

2.1 Conventional NMF

Lee and Seung introduced the multiplicative update rule of NMF for KL divergence [11]. As we iteratively update the parameters, we can represent a non-negative matrix, which may correspond to a magnitude spectrogram, as a multiplication of two non-negative matrices that may contain spectral bases and temporal bases. The update rule can be represented as:

$$\mathbf{H}_{k,n} \leftarrow \frac{\mathbf{H}_{k,n} \sum_{m} \left\{ \mathbf{W}_{m,k} \mathbf{F}_{m,n} \Big/ \tilde{\mathbf{F}}_{m,n} \right\}}{\sum_{m'} \mathbf{W}_{m',k}} \tag{1}$$

$$\mathbf{W}_{m,k} \leftarrow \frac{\mathbf{W}_{m,k} \sum_{n} \left\{ \mathbf{H}_{k,n} \mathbf{F}_{m,n} \Big/ \tilde{\mathbf{F}}_{m,n} \right\}}{\sum_{n'} \mathbf{H}_{k,n'}} \tag{2}$$

where $\mathbf{F}$ and $\tilde{\mathbf{F}}$ denote the $M \times N$ magnitude spectrogram of an audio mixture, and its estimation, respectively, $\mathbf{W}$ and $\mathbf{H}$ denote the $M \times K$ matrix of the spectral bases and the $K \times N$ matrix of their activations.

2.2 Formulation of Harmonic-Percussive Separation

We present a modified NMF algorithm to impose the characteristics of harmonic/percussive sounds. The update rule is separately represented for the harmonic source basis and percussive source basis as follows:

$$\mathbf{H}_{k,n} \leftarrow \frac{\mathbf{H}_{k,n} \sum_{m} \left\{ \mathbf{W}_{m,k} \mathbf{F}_{m,n} \Big/ \tilde{\mathbf{F}}_{m,n} \right\}}{\sum_{m'} \mathbf{W}_{m',k}} \tag{3}$$

$$\mathbf{W}_{m,k} \leftarrow \frac{\mathbf{W}_{m,k} \sum_{n} \left\{ \mathbf{H}_{k,n} \mathbf{F}_{m,n} \Big/ \tilde{\mathbf{F}}_{m,n} \right\}}{\sum_{n'} \mathbf{H}_{k,n'}} \tag{4}$$

$$\mathbf{w}_k \leftarrow \left(1 - \gamma_H^H\right) \mathbf{w}_k + \gamma_H^H ifft\left(\left\{fft\left(\mathbf{w}_k\right)\right\}^p\right), k \in \Phi_H \tag{5}$$

$$\mathbf{w}_k \leftarrow \max\left(\mathbf{w}_k, 0\right), k \in \Phi_H \tag{6}$$

$$\begin{cases} \mathbf{w}_k \leftarrow \left(1 - \gamma_S^H\right) \mathbf{w}_k + \gamma_S^H (\mathbf{w}_k)^q, k \in \Phi_H \\ \mathbf{w}_k \leftarrow \left(1 - \gamma_S^P\right) \mathbf{w}_k + \gamma_S^P (\mathbf{w}_k)^r, k \in \Phi_P \end{cases} \tag{7}$$

where Φ_H and Φ_P denote a set of harmonic bases and percussive bases, respectively, $fft\left(\cdot\right)$ and $ifft\left(\cdot\right)$ denote the functions of the fast Fourier transform (FFT) and the inverse FFT (IFFT), respectively, $\mathbf{w}_k$ denotes the kth column of $\mathbf{W}$, γ_H^H denotes the harmonicity weight parameter for the harmonic signal, and γ_S^H and γ_S^P denote the sparsity weight parameters for harmonic and percussive signals, respectively. Note that Eqns (3) and (4) are identical to Eqns (1) and (2), respectively. Eqns (5)-(7) contribute to shaping the spectral bases as desired as the iteration proceeds.

Mixing weights that have values between 0 and 1 represent the importance of each constraint imposition, and indicate the degree to which we need to impose the characteristic. To enable the harmonic bases to have a harmonic and sparse structure while preserving the original figures of spectral bases, γ_H^H and γ_S^H are set to have small positive numbers, as the effect of the constraint is accumulated over the iteration.

The exponents p, q, and r have to be determined considering the range of each parameter, $0 \leq r \leq 1 \leq p, q$. Here, p and q respectively reflect the degree of harmonicity and sparsity of the destination, and they have to be controlled considering the spectral characteristics of the original harmonic sources. Likewise, r reflects the degree of "unsparsity" of the percussive sources.

Among the update equations shown above, the function of the conventional NMF update equations in Eqns (3) and (4) is to minimize the error between $\mathbf{F}$ and its estimation $\tilde{\mathbf{F}}$. On the other hand, the remainders of the equations aim to shape the spectral bases. The sparsity constraint in Eqn (7) has been similarly adopted for the matrix decomposition [10], and it is based on the fact that the square operation increases the differences among the vector components. If the square root operation is used instead, as

in the percussive case of Eqn (7), unsparsity can be imposed to the basis. Similarly, we can extend this concept to the harmonicity. The second term in Eqn (5) denotes the harmonics-emphasized basis, which is due to the fact that the *spectrum of the spectrum* is sparse. To prevent elements from being negative, the $\max(\cdot,\cdot)$ operation in Eqn (6) has to be jointly involved.

The harmonic and percussive sounds are reconstructed using the corresponding bases as follows:

$$\mathbf{F}^{(Harmonic)} = \sum_{k \in \Phi_H} \mathbf{w}_k \mathbf{h}_k \tag{8}$$

$$\mathbf{F}^{(Percussive)} = \sum_{k \in \Phi_P} \mathbf{w}_k \mathbf{h}_k \tag{9}$$

where $\mathbf{h}_k$ denotes the kth row of $\mathbf{H}$.

2.3 Relation to Dirichlet Prior

The proposed update equations can be intuitively comprehended. However, the equations are based on a firm theoretical background, not heuristically induced. In this subsection, we employ Dirichlet prior from the probability theory, and investigate its relations to the proposed method.

Priors were primarily adopted for the Bayesian probability theory, including the probabilistic latent component analysis (PLCA) or probabilistic latent semantic analysis (PLSA). Such spectrogram decomposition techniques often regard spectrogram components as histogram elements of multinomial distributions. Because the Dirichlet distribution is a conjugate prior of a multinomial distribution, it can be adopted as a prior knowledge of a multinomial distribution. By adopting the prior, we can modify our goal to be the maximizing posterior from the maximizing likelihood. For this reason, the Dirichlet prior has been adopted for the matrix factorization in the previous works [10], [12]. Our method employs one of the extensions of the Dirichlet prior for harmonicity imposition.

Because PLCA is a special case of NMF, where its cost function is KL divergence [13], we can generalize the Dirichlet prior of the PLCA [12] by applying it to the NMF algorithm as follows:

$$\mathbf{H}_{k,n} \leftarrow (1-\gamma_1) \frac{\mathbf{H}_{k,n} \sum_m \left\{ \mathbf{W}_{m,k} \mathbf{F}_{m,n} \Big/ \tilde{\mathbf{F}}_{m,n} \right\}}{\sum_{m'} \mathbf{W}_{m',k}} + \gamma_1 \mathbf{A}_{k,n} \tag{10}$$

$$\mathbf{W}_{m,k} \leftarrow (1-\gamma_2) \frac{\mathbf{W}_{m,k} \sum_n \left\{ \mathbf{H}_{k,n} \mathbf{F}_{m,n} \Big/ \tilde{\mathbf{F}}_{m,n} \right\}}{\sum_{n'} \mathbf{H}_{k,n'}} + \gamma_2 \mathbf{B}_{m,k} \tag{11}$$

where $\mathbf{A}$ and $\mathbf{B}$ denote the matrices of hyper parameters with respect to $\mathbf{H}$ and $\mathbf{W}$, respectively, and γ_1 and γ_2 denote the mixing weights. In our research, we focus only on the spectral bases, and thus Eqn (10) is discarded. As can be observed, the proposed update equations, Eqns (3)-(7), have the same form as Eqn (11), and the way in which we shape the spectral bases depends on the form of $\mathbf{B}$ matrix.

Frequency-domain sparsity imposition can be easily achieved by setting the hyper parameter $\mathbf{B}$ as [10]

$$\mathbf{b}_k = (\mathbf{w}_k)^u \tag{12}$$

where $\mathbf{b}_k$ denotes the kth column of $\mathbf{B}$, and u denotes an exponent that controls the degree of sparsity of $\mathbf{b}_k$.

On the other hand, harmonicity imposition can be achieved when the hyper parameter is represented as

$$\mathbf{b}_k = ifft\left(\{fft(\mathbf{w}_k)\}^v\right) \tag{13}$$

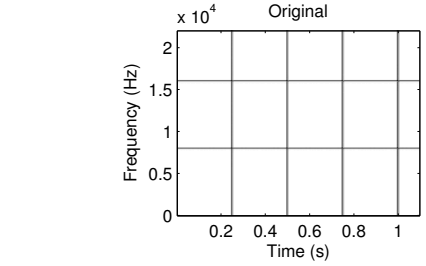
where v denotes the exponent that controls the degree of harmonicity of $\mathbf{b}_k$. This is because a periodic signal can be represented as a sum of sinusoids, and the spectrum of the periodic signal is sparse. Conversely, if a spectrum is sparse, we can assume that the original signal has a strongly periodic characteristic. Thus, we aim to make the *spectrum of the spectrum* to be sparse in order to shape a signal such that it has a harmonic structure. Note that in order to prevent destructive interference caused by phase distortion, we have to manipulate only the magnitudes within the IFFT function, preserving the original phases of $fft(\mathbf{w}_k)$.

3. PERFORMANCE EVALUATION

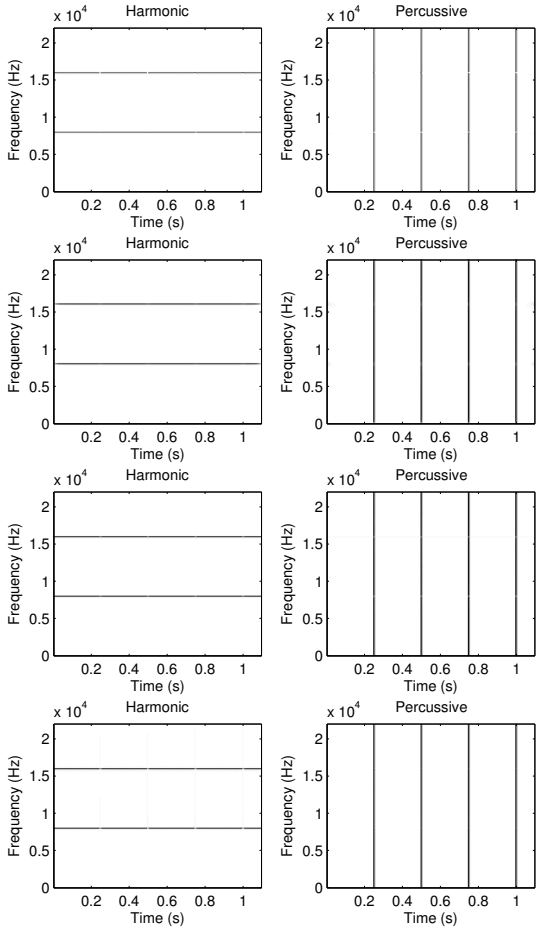
3.1 Sample Problem

In this section, we apply the proposed method and the conventional methods to simple sample examples, which is suitable for showing the novelty and validity of the proposed method. Spectrograms of synthesized sounds that consist of horizontal and vertical lines are presented in Figure 1(a) and Figure 2(a). Figure 1(a) models the case where a pitched harmonic sound is sustained for a certain period. The sounds of harmonic instruments such as guitars, pianos, flutes, and violins fall within this scenario. On the other hand, Figure 2(a) illustrates the case where a harmonic signal alters its frequency over time. In this case, vibratos, glissandos, and vocal signals correspond to the harmonic components. We compare the performance of the proposed method to the separation results obtained using three conventional methods: Ono *et al.*'s Euclidean distance-based method [6], Ono *et al.*'s KL divergence-based method [7], and FitzGerald's method [8].

As shown in Figure 1(b), both the conventional methods and the proposed method are able to successfully separate the sounds. This is because the horizontal lines in this example have horizontally continuous characteristics, which are assumed by the conventional methods to be present. However, when the harmonic sound vibrates and the horizontal lines fluctuate, as shown in Figure 2(a), conventional methods cannot distinguish the horizontal lines from vertical lines. As we can see in Figure 2(b), the estimated percussive components of conventional methods contain harmonic partials, and only the proposed method can successfully separate them. Thus, we can claim that the pro-

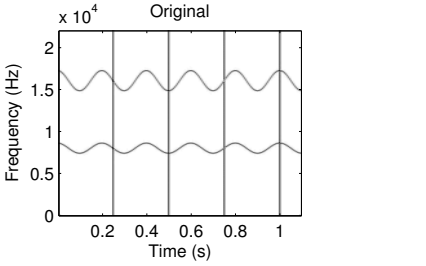


(a) Original figure

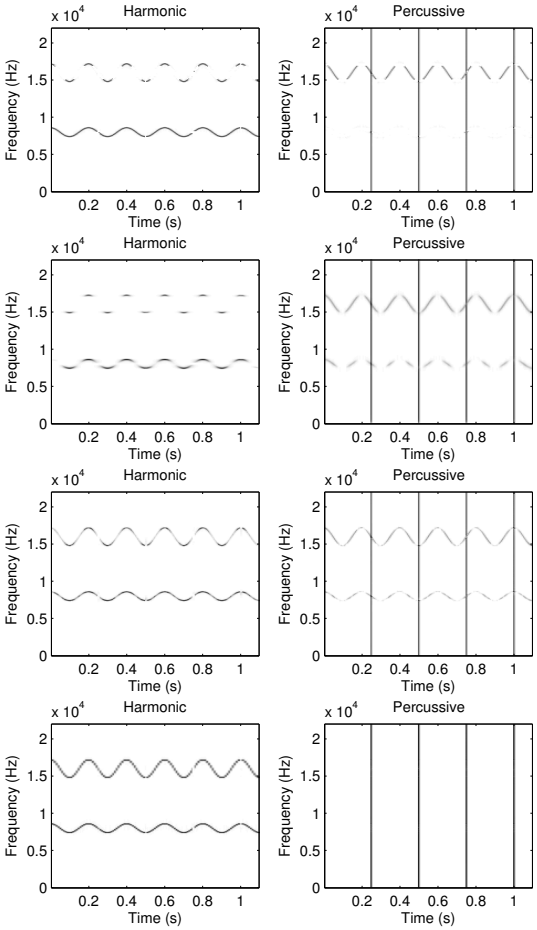


(b) Separation results of Ono's Euclidean distance-based method, Ono's KL divergence-based method, FitzGerald's method, and the proposed method (from top to bottom)

Figure 1. Sample example of separating horizontal lines and vertical lines.



(a) Original figure



(b) Separation results of Ono's Euclidean distance-based method, Ono's KL divergence-based method, FitzGerald's method, and the proposed method (from top to bottom)

Figure 2. Sample example of separating fluctuating horizontal lines and vertical lines.

posed method is not affected by variations in the pitch because it relies on the harmonic structure of the vertical axis, and not the degree of horizontal transition.

3.2 Qualitative Analysis

We evaluated the performance of the proposed method using a real recording example. Figure 3 shows a log-scale plot of the spectrogram of an excerpt from "*Billie Jean*," by *Michael Jackson*. The signal was sampled at 22,050 Hz, and the frame size and overlap size were set to 1,024 and 512, respectively. We can observe from the spectrogram that the excerpt contains both harmonic and percussive components. The harmonic components can be seen as horizontally connected lines, whereas the percussive components are seen as vertical lines as in the sample examples.

Figure 4(a) and (b) show the separation results of the harmonic sound (up) and percussive sound (down), which were obtained using Ono *et al*.'s Euclidean distance-based method and KL divergence-based method, respectively. Here, we set the parameters to the values recommended in the references. We observe that the estimated percussive

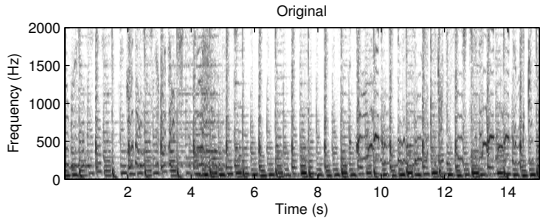


Figure 3. Spectrogram of a real audio recording example ("Billie Jean" by Michael Jackson).

components still contain harmonic components that may correspond to the vocal components. This is because Ono *et al*.'s algorithms aim to minimize the temporal transition of the harmonic spectrogram. However, vocal components in the original spectrogram do not match well with the underlying assumption.

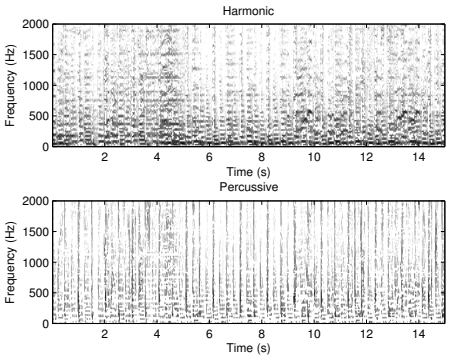
Figure 4(c) shows the result of FitzGerald's method with a median filter length of 17 and when the exponent for the Wiener filter-based soft mask is two, as recommended by FitzGerald [8]. We also observe that the separated percussive components still contain harmonic components, as in the previous case. This is because of the use of a one-dimensional median filter, which assumes that the harmonic components are sustained for several periods.

Figure 4(d) shows the performance of the proposed method. We observe that the harmonic and percussive components are clearly separated, and the percussive components do not have any vocal components in these results. This is because unlike conventional methods, the proposed algorithm does not rely on the horizontal continuity principle. Rather, the proposed algorithm tries to account for the harmonic components using the harmonic and sparse spectral bases.

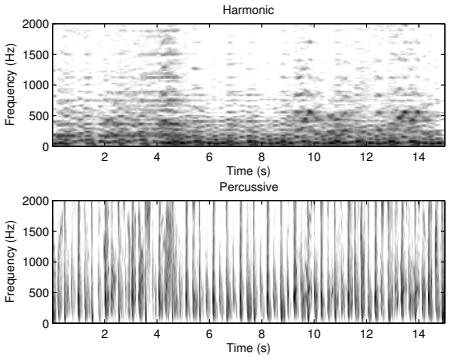
3.3 Quantitative Analysis

We performed a quantitative analysis to verify the validity of the proposed algorithm. First, we compiled a dataset that consists of 10 audio samples, which is a subset of the MASS database [14], but two sets of data, namely *tamy-que_pena_tanto_faz_6-19* and *tamy-que_pena_tanto_faz_46-57*, were excluded in this experiment because they lack percussive signals. Then, we obtained a spectrogram for each audio sample with the frame size and hop size set to 2,048 samples and 1,024 samples, respectively. Note that the sampling rate of the songs in the MASS dataset is 44,100 Hz. Finally, we measured the signal-to-distortion ratio (SDR), signal-to-interference ratio (SIR), and signal-to-artifact ratio (SAR) using the BSS_EVAL toolbox (http://bass-db.gforce.inria.fr /bss_eval/) supported by [15]. Table 1 shows the parameter values of the proposed method used in this experiment. The parameters of the conventional methods are set to the recommended values, as in the previous experiment.

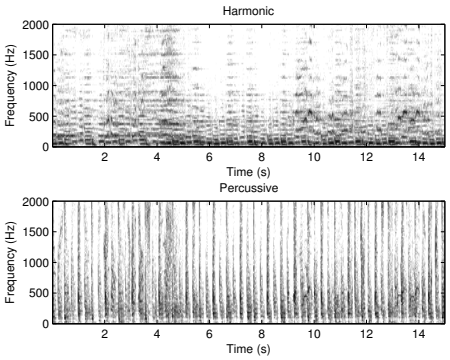
The evaluation results are summarized in Figure 5. We can see that the proposed method guarantees a better av-



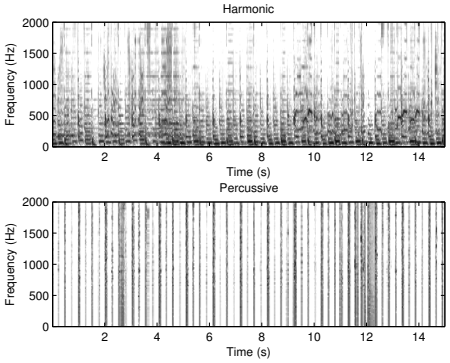
(a) Ono's Euclidean distance-based method



(b) Ono's KL divergence-based method



(c) FitzGerald's method



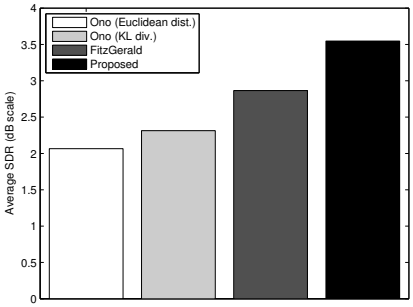
(d) Proposed method

Figure 4. Qualitative performance comparison of conventional and proposed methods.

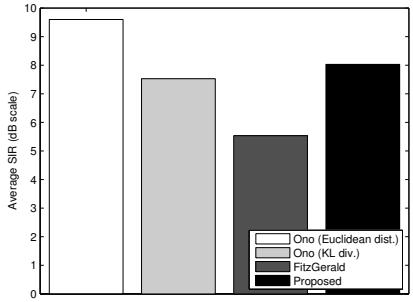
erage SDR result compared to conventional methods, even though the proposed method has a lower SIR performance than Ono *et al*.'s Euclidean distance-based method. This is because the proposed method far outperforms other methods with respect to the SAR, which has a trade-off relation with the SIR [16].

Parameter	Value
p	1.1
q	1.1
r	0.5
γ_H^H	0.001
γ_S^H	0.001
γ_S^P	0.1
Number of bases (H,P)	(300,200)

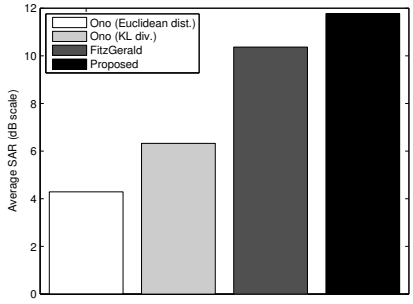
Table 1. Experimental parameters.



(a) SDR



(b) SIR



(c) SAR

Figure 5. Quantitative performance comparison of conventional and proposed methods.

4. CONCLUSION

In this paper, we proposed a novel HPSS algorithm based on NMF with harmonicity and sparsity constraints. Conventional methods assumed that the harmonic components were represented as horizontal lines with temporal continuity. However, such an assumption could not be applied to the vocal components or various musical expressions of harmonic instruments. To overcome this problem, we presented a harmonicity constraint, which is a generalized Dirichlet prior. By letting the spectrum of the spectrum be harmonic and sparse, we could refine the harmonic components and eliminate inharmonic components. The experimental results showed the validity of the proposed method by comparing it with conventional methods.

5. ACKNOWLEDGEMENTS

This research was partly supported by the MSIP(Ministry of Science, ICT and Future Planning), Korea, under the ITRC(Information Technology Research Center) support program (IITP-2015-H8501-15-1016) supervised by the IITP(Institute for Information & communications Technology Promotion). Also, this research was supported in part by the A3 Foresight Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology.

6. REFERENCES

[1] C. Uhle, C. Dittmar, and T. Sporer, "Extraction of drum tracks from polyphonic music using independent subspace analysis," *Proceedings of the ICA*, pp. 843–847, April, 2003.

[2] O. Gillet and G. Richard, "Drum track transcription of polyphonic music using noise subspace projection," *Proceedings of the ISMIR*, pp. 92–99, September, 2005.

[3] M. Helen and T. Virtanen, "Separation of drums from polyphonic music using non-negative matrix factorization and support vector machine," *Proceedings of the EUSIPCO*, September 2005.

[4] M. Kim, J. Yoo, K. Kang, and S. Choi, "Nonnegative matrix partial co-factorization for spectral and temporal drum source separation," *IEEE Journal of Selected Topics in Signal Processing*, Vol.5, No.6, pp. 1192–1204, 2011.

[5] F. J. Canadas-Quesada, P. Vera-Candeas, N. Ruiz-Reyes, J. Carabias-Orti, and P. Cabanas-Molero, "Percussive/harmonic sound separation by non-negative matrix factorization with smoothness/sparseness constraints," *EURASIP Journal on Audio, Speech, and Music Processing*, Vol.2014, No.1, pp. 1–17, 2014.

[6] N. Ono, K. Miyamoto, J. Le Roux, H. Kameoka, and S. Sagayama, "Separation of a monaural audio signal

into harmonic/percussive components by complementary diffusion on spectrogram," *Proceedings of the EUSIPCO*, August, 2008.

[7] N. Ono, K. Miyamoto, H. Kameoka, and S. Sagayama, "A real-time equalizer of harmonic and percussive components in music signals," *Proceedings of the ISMIR*, pp. 139–144, September, 2008.

[8] D. FitzGerald, "Harmonic/percussive separation using median filtering," *Proceedings of the 13th International Conference on Digital Audio Effects (DAFx-10)*, 2010.

[9] A. Gkiokas, V. Papavassiliou, V. Katsouros, and G. Carayannis, "Deploying nonlinear image filters to spectrogram for harmonic/percussive separation," *Proceedings of the 15th International Conference on Digital Audio Effects (DAFx-12)*, 2012.

[10] M. Kim and P. Smaragdis, "Manifold preserving hierarchical topic models for quantization and approximation," *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, pp. 1373–1381, 2013.

[11] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," *Proceedings of the Advances in Neural Information Processing Systems*, November, 2000.

[12] P. Smaragdis and G. J. Mysore, "Separation by humming: User-guided sound extraction from monophonic mixtures," *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, October, 2009.

[13] C. Ding, T. Li, and W. Peng, "On the equivalence between non-negative matrix factorization and probabilistic latent semantic indexing," *Computational Statistics & Data Analysis*, Vol.52, No.8, pp. 3913–3927, 2008.

[14] M. Vinyes, "MTG MASS database," http://www.mtg.upf.edu/static/mass/resources, 2008.

[15] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, Vol.14, No.4, pp. 1462–1469, 2006.

[16] D. L. Sun, and G. J. Mysore, "Universal speech models for speaker independent single channel source separation," *Proceedings of the ICASSP*, 2013.