SEPARATION OF A MONAURAL AUDIO SIGNAL INTO HARMONIC/PERCUSSIVE COMPONENTS BY COMPLEMENTARY DIFFUSION ON SPECTROGRAM

Nobutaka Ono, Kenichi Miyamoto, Jonathan Le Roux, Hirokazu Kameoka, and Shigeki Sagayama

Department of Information Physics and Computing,
Graduate School of Information Science and Technology, The University of Tokyo
7-3-1 Hongo Bunkyo-ku, Tokyo, 113-8656, Japan
Email: {onono, miyamoto, leroux, kameoka, sagayama}@hil.t.u-tokyo.ac.jp
web: http://hil.t.u-tokyo.ac.jp/index-e.html

ABSTRACT

In this paper, we present a simple and fast method to separate a monaural audio signal into harmonic and percussive components, which is much useful for multi-pitch analysis, automatic music transcription, drum detection, modification of music, and so on. Exploiting the differences in the spectrograms of harmonic and percussive components, the objective function is defined in a quadrature form of the spectrogram gradients. Applying the auxiliary function approach to that, simple and fast update equations are derived, which guarantee the decrease of the objective function at each iteration. We show some experimental results by applying our method to popular and jazz music songs.

1. INTRODUCTION

Recently, music signal has become a significant target in the signal processing field and various tasks have been discussed like audio music retrieval, audio onset detection, multiple fundamental frequency estimation, and so on [1]. The music signal often consists of two different components: harmonic one and percussive one. The simultaneous presence of them makes some tasks harder because of their much different spectral structures. For instance, most of the multipitch analysis are interfered by percussive tones, while suppression of harmonic components will facilitate the drum detection or the rhythm analysis.

In this paper, aiming to efficient pre-processing of music signal analysis, we present a simple and fast method to separate a monaural audio signal into harmonic and percussive components. This kind of separation problem has been discussed in several pilot research. Uhle et al. applied Independent Component Analysis (ICA) to the power spectrogram, and classified the extracted independent components into a harmonic and a percussive groups based on the several features like percussiveness, noise-likeness, etc [2]. Helen and Virtanen utilized Non-negative Matrix Factorization (NMF) for decomposing the spectrogram into elementary patterns and classified them by pre-trained Support Vector Machine (SVM) [3]. Through modeling harmonic and inharmonic tones on spectrogram, Itoyama et al. proposed separation of an audio signal to each track part based on the MIDI information synchronized to the input audio signal [4]. Daudet reviewed recent separation algorithms of music signals into steady and transient components [5]. Other kinds of single channel separation problem have been recently developed in [6, 7, 8, 9]. The contribution of this paper is to derive a simple and fast algorithm specifically for the harmonic/percussive separation based on the anisotropy of them on spectrogram. We present the formulation of the separation as an optimization problem, derive the fast iterative solution to it by auxiliary function approach, and examine the performance by experiments to popular and jazz music songs.

2. FORMULATION OF HARMONIC/PERCUSSIVE SEPARATION

Let $F_{h,i}$ be a Short Time Fourier Transform (STFT) of a monaural audio signal $f(t)$ and $W_{h,i} = |F_{h,i}|^2$ be its power spectrogram, where h and i represent indices of frequency and time bins. A typical spectrogram of an audio signal is shown in Fig. 1. In it, the vertical and horizontal structures are clearly observed. The harmonic component usually has a stable pitch and forms parallel ridges with smooth temporal envelopes on the spectrogram, while the energy of a percussive tone is concentrated in a short time frame, which forms a vertical ridge with a wideband spectral envelope. Exploiting the anisotropy, we decompose the original power spectrogram $W_{h,i}$ into the harmonic component $H_{h,i}$ and the percussive component $P_{h,i}$ on the spectrogram. For evaluating their anisotropic smoothness, L_2 norm of the power spectrogram gradients is examined here, that is, $H_{h,i}$ and $P_{h,i}$ are found by minimizing

$$\begin{aligned} J(\boldsymbol{H},\boldsymbol{P}) &= \frac{1}{2\sigma_H^2}\sum_{h,i}(H_{h,i-1}-H_{h,i})^2 \\ &\quad+\frac{1}{2\sigma_P^2}\sum_{h,i}(P_{h-1,i}-P_{h,i})^2 \qquad (1) \end{aligned}$$

under the constraint as

$$\begin{aligned} H_{h,i}+P_{h,i} &= W_{h,i} \qquad (2)\\ H_{h,i} &\geq 0, \quad P_{h,i} \geq 0, \qquad (3) \end{aligned}$$

where $\boldsymbol{H}$ and $\boldsymbol{P}$ are sets of $H_{h,i}$ and $P_{h,i}$, respectively, and σ_H and σ_P are parameters to control the weights of the horizontal and vertical smoothness. Minimizing eq. (1) is equivalent to maximum likelihood estimation under the assumption that the spectrogram gradients $(H_{h,i-1} - H_{h,i})$ and $(P_{h-1,i} - P_{h,i})$ follow independent Gaussian distributions. Although the actual distributions of the spectrogram gradients are different from them, the assumption leads us to simple and comprehensive formulation and solution. As confirmed later, replacing the power spectrogram $W_{h,i}$ by the range-compressed version: $\tilde{W}_{h,i} = |F_{h,i}|^{2\gamma}(0 < \gamma \leq 1)$ partially bridges a gap between the assumption and the real situation.

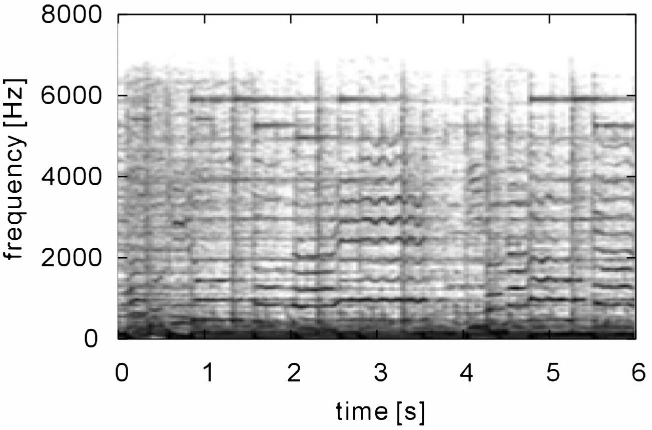

Figure 1: A spectrogram of a popular music song

3. DESIGN OF AUXILIARY FUNCTION

Since the objective function defined by eq. (1) is quadratic for all variables, it is monomodal and has a single global minimum, which can be directly obtained by solving $\partial J/\partial H_{h,i}=0$ and $\partial J/\partial P_{h,i}=0$. But they yield much large-size simultaneous equations of $H_{h,i}$ and $P_{h,i}$ (the order of the equations is equal to the number of all time-frequency bins of (h,i)). To avoid it and derive a simple iterative solution, we adopt an auxiliary function approach, which is used in other signal processing methods as NMF [10] or Harmonic-Temporal Clustering (HTC) [11].

In order to design the auxiliary function of our problem, note that

$$(A-B)^2 \leq 2(A-X)^2+2(B-X)^2 \tag{4}$$

for any A, B, and X since

$$\begin{aligned} & 2(A-X)^2+2(B-X)^2-(A-B)^2 \\ & \quad = 4\left(X-\frac{A+B}{2}\right)^2 \end{aligned} \tag{5}$$

is obviously nonnegative and equal to zero where $X=(A+B)/2$. Applying the inequality to eq. (1), we have

$$(H_{h,i-1}-H_{h,i})^2 \leq 2(H_{h,i-1}-U_{h,i})^2+2(H_{h,i}-U_{h,i})^2, \tag{6}$$

$$(P_{h-1,i}-P_{h,i})^2 \leq 2(P_{h-1,i}-V_{h,i})^2+2(P_{h,i}-V_{h,i})^2, \tag{7}$$

for any $U_{h,i}$ and $V_{h,i}$. The equalities are valid for $U_{h,i}=(H_{h,i-1}+H_{h,i})/2$ and $V_{h,i}=(P_{h-1,i}+P_{h,i})/2$. Hence, the auxiliary function:

$$\begin{aligned} & Q(\boldsymbol{H},\boldsymbol{P},\boldsymbol{U},\boldsymbol{V}) \\ = & \frac{1}{\sigma_H^2}\sum_{h,i}\left\{(H_{h,i-1}-U_{h,i})^2+(H_{h,i}-U_{h,i})^2\right\} \\ & +\frac{1}{\sigma_P^2}\sum_{h,i}\left\{(P_{h-1,i}-V_{h,i})^2+(P_{h,i}-V_{h,i})^2\right\} \end{aligned} \tag{8}$$

satisfies

$$J(\boldsymbol{H},\boldsymbol{P}) \leq Q(\boldsymbol{H},\boldsymbol{P},\boldsymbol{U},\boldsymbol{V}), \tag{9}$$

$$J(\boldsymbol{H},\boldsymbol{P}) = \min_{\boldsymbol{U},\boldsymbol{V}} Q(\boldsymbol{H},\boldsymbol{P},\boldsymbol{U},\boldsymbol{V}). \tag{10}$$

Then, the following updates:

$$\{\boldsymbol{U}^{(k+1)},\boldsymbol{V}^{(k+1)}\} = \min_{\boldsymbol{U},\boldsymbol{V}} Q(\boldsymbol{H}^{(k)},\boldsymbol{P}^{(k)},\boldsymbol{U},\boldsymbol{V}), \tag{11}$$

$$\{\boldsymbol{H}^{(k+1)},\boldsymbol{P}^{(k+1)}\} = \min_{\boldsymbol{H},\boldsymbol{P}} Q(\boldsymbol{H},\boldsymbol{P},\boldsymbol{U}^{(k+1)},\boldsymbol{V}^{(k+1)}), \tag{12}$$

decrease J monotonically, where k represents the number of iterations, and $\boldsymbol{U}$ and $\boldsymbol{V}$ are sets of $U_{h,i}$ and $V_{h,i}$, respectively.

4. DERIVATION OF UPDATE RULES

First, we here derive $\boldsymbol{H}^{(k+1)}$ and $\boldsymbol{P}^{(k+1)}$ satisfying eq. (12) under the constraint of eq. (2). With introducing Lagrange multipliers $\lambda_{h,i}$, consider

$$\begin{aligned} \tilde{Q}(\boldsymbol{H},\boldsymbol{P}) &= Q(\boldsymbol{H},\boldsymbol{P},\boldsymbol{U}^{(k+1)},\boldsymbol{V}^{(k+1)}) \\ &\quad +\sum_{h,i}\lambda_{h,i}(H_{h,i}+P_{h,i}-W_{h,i}). \end{aligned} \tag{13}$$

Differentiating it by $H_{h,i}$, $P_{h,i}$, and $\lambda_{h,i}$ yields, respectively

$$\frac{2}{\sigma_H^2}(2H_{h,i}-U_{h,i+1}^{(k+1)}-U_{h,i}^{(k+1)})+\lambda_{h,i} = 0, \tag{14}$$

$$\frac{2}{\sigma_P^2}(2P_{h,i}-V_{h+1,i}^{(k+1)}-V_{h,i}^{(k+1)})+\lambda_{h,i} = 0, \tag{15}$$

$$H_{h,i}+P_{h,i}-W_{h,i} = 0. \tag{16}$$

Solving them, we obtain

$$\begin{aligned} H_{h,i}^{(k+1)} &= \frac{\alpha}{2}(U_{h,i+1}^{(k+1)}+U_{h,i}^{(k+1)}) \\ &\quad +\frac{(1-\alpha)}{2}(2W_{h,i}-V_{h+1,i}^{(k+1)}-V_{h,i}^{(k+1)}), \end{aligned} \tag{17}$$

$$\begin{aligned} P_{h,i}^{(k+1)} &= \frac{(1-\alpha)}{2}(V_{h+1,i}^{(k+1)}+V_{h,i}^{(k+1)}) \\ &\quad +\frac{\alpha}{2}(2W_{h,i}-U_{h,i+1}^{(k+1)}-U_{h,i}^{(k+1)}), \end{aligned} \tag{18}$$

where

$$\alpha = \frac{\sigma_P^2}{\sigma_H^2+\sigma_P^2}. \tag{19}$$

While, the auxiliary parameters $\boldsymbol{U}^{(k+1)}$ and $\boldsymbol{V}^{(k+1)}$ satisfying eq. (11) are easily given by

$$U_{h,i}^{(k+1)} = \frac{H_{h,i-1}^{(k)}+H_{h,i}^{(k)}}{2}, \quad V_{h,i}^{(k+1)} = \frac{P_{h-1,i}^{(k)}+P_{h,i}^{(k)}}{2}. \tag{20}$$

By substituting eq. (20) into eq. (17) and eq. (18), we can remove the auxiliary parameters $U_{h,i}$ and $V_{h,i}$ from the update rules as

$$H_{h,i}^{(k+1)} = H_{h,i}^{(k)}+\Delta^{(k)}, \tag{21}$$

$$P_{h,i}^{(k+1)} = H_{h,i}^{(k)}-\Delta^{(k)}, \tag{22}$$

where

$$\begin{aligned} \Delta^{(k)} &= \alpha\left(\frac{H_{h,i-1}^{(k)}-2H_{h,i}^{(k)}+H_{h,i+1}^{(k)}}{4}\right) \\ &\quad -(1-\alpha)\left(\frac{P_{h-1,i}^{(k)}-2P_{h,i}^{(k)}+P_{h+1,i}^{(k)}}{4}\right). \end{aligned} \tag{23}$$

Introducing a process to restrict the obtained solution to satisfy eq. (3) and binarizing a separation result, which is experimentally confirmed to yield better separation performance, the separation algorithm is consequently summarized as follows.

1. Calculate $F_{h,i}$, the STFT of an input signal $f(t)$.
2. Calculate a range-compressed version of the power spectrogram by

$$W_{h,i} = |F_{h,i}|^{2\gamma} \quad (0 < \gamma \leq 1). \tag{24}$$

3. Set initial values as

$$H_{h,i}^{(0)} = P_{h,i}^{(0)} = \frac{1}{2} W_{h,i}, \tag{25}$$

for all h and i and set $k = 0$.

4. Calculate the update variables $\Delta^{(k)}$ defined as eq. (23).
5. Update $H_{h,i}$ and $P_{h,i}$ as

$$\begin{aligned} H_{h,i}^{(k+1)} &= \min(\max(H_{h,i}^{(k)} + \Delta^{(k)}, 0), W_{h,i}), \quad (26) \\ P_{h,i}^{(k+1)} &= W_{h,i} - H_{w,i}^{(k+1)}. \quad (27) \end{aligned}$$

6. Increment k. If $k < k_{max} - 1$ (k_{max}: the maximum number of iterations), then, go to step 4, else, go to step 7.
7. Binarize the separation result as

$$\begin{aligned} &(H_{h,i}^{(k_{max})}, P_{h,i}^{(k_{max})}) \\ &= \begin{cases} (0, W_{h,i}) & (H_{h,i}^{(k_{max}-1)} < P_{h,i}^{(k_{max}-1)}) \\ (W_{h,i}, 0) & (H_{h,i}^{(k_{max}-1)} \geq P_{h,i}^{(k_{max}-1)}) \end{cases} \quad (28) \end{aligned}$$

8. Convert $H_{h,i}^{(k_{max})}$ and $P_{h,i}^{(k_{max})}$ into waveforms by

$$\begin{aligned} h(t) &= \mathrm{ISTFT}[(H_{h,i}^{(k_{max})})^{1/2\gamma} e^{j\angle F_{h,i}}], \quad (29) \\ p(t) &= \mathrm{ISTFT}[(P_{h,i}^{(k_{max})})^{1/2\gamma} e^{j\angle F_{h,i}}], \quad (30) \end{aligned}$$

where ISTFT represents the inverse STFT.

Since the update variable Δ consists of the second order derivative of $H_{h,i}$ and $P_{h,i}$, the update in step 4 has basically the same form as the diffusion equation: $\frac{df}{dt} = C\frac{d^2 f}{dx^2}$, which represents dynamics of diffusion phenomena like heat. But unlike the physical diffusion process, each update of $H_{h,i}$ and $P_{h,i}$ includes a negative diffusion term derived from the other. As iterations, the energy distribution of $H_{h,i}$ on the spectrogram diffuses horizontally and concentrates vertically. $P_{h,i}$ follows the inverse way. Both of which hold that $H_{h,i} + P_{h,i} = W_{h,i}$, $H_{h,i} \geq 0$, and $P_{h,i} \geq 0$. We denote the diffusion-like process of two energy distributions with a balance by *complementary diffusion*. The balance parameter α $(0 < \alpha < 1)$ controls the strength of the diffusion along the vertical and the horizontal directions.

5. EXPERIMENTAL EVALUATIONS

We here show several results by experiments. The target signals were chosen from the RWC Music Database (Popular Music Database) [12] and the sampling frequency was converted into 16kHz. The frame length of STFT is 1024 and the frame shift is 512. The balance parameter α was experimentally determined to be 0.3. The resultant spectrograms of the harmonic component $H_{h,i}^{(k)}$ and the percussive component $P_{h,i}^{(k)}$ to 6.25s fragment of RWC-MDB-P-2001 No.7 is shown in Fig. 2, where $\gamma = 0.3$ was used. We can see that the energy of spectrogram is splitting to two components as iterations, each of which is forming horizontal and vertical ridges, respectively. The computational time for a 6.25s-length signal with 50 iterations is about 2.3s at a laptop-PC with 1.20GHz Pentium in our implementation, which is nearly three times faster than real-time processing. Indeed, we have developed the real-time processing version of it.

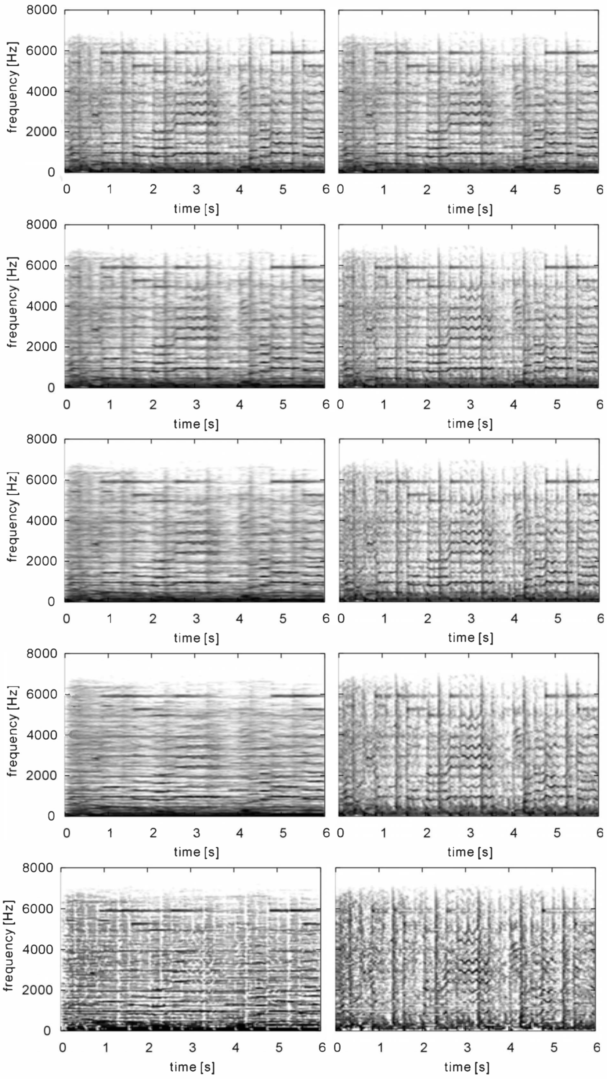

Figure 2: The spectrograms of the iteratively-updated harmonic component $H_{h,i}^{(k)}$(left) and the percussive component $P_{h,i}^{(k)}$ (right) at $k = 0$, $k = 3$, $k = 10$, $k = 50$, and after binarization, from top to bottom, respectively.

In auditory evaluation, pitched instrument tracks and drum tracks are well separated into $h(t)$ and $p(t)$, respectively. Nevertheless, the duration of some percussion, typically bass drum was almost separated into $h(t)$ since it has a smooth temporal envelope, too. Inversely, the attack of pitched tone had a tendency to belong to $p(t)$. A singing voice is also difficult to be perfectly classified to $h(t)$ due to the nature of time-varying pitch. In the condition, most of them except pitch-varying component belonged to $h(t)$. We think that utilizing wavelet transform instead of STFT has some potential to improve the performance for pitch-varying components. The resultant sound will be presented in the conference.

In order to quantitatively evaluate the separation performance of our algorithm, we prepared each track of audio signals in two songs (RWC-MDB-P-2001 No.18 and RWC-MDB-J-2001 No.16) by MIDI-to-WAV conversion. Let $f_i(t)$ be one audio track signal where i is an index of the track. and inputed the summation of all tracks to our algorithm. Then, the energy ratio of each track included in $h(t)$ and $p(t)$ was calculated as $r_h = E_h/(E_h+E_p)$, $r_p = E_p/(E_h+E_p)$, where $E_h =< f_i(t), h(t) >^2$, $E_p =< f_i(t),p(t) >^2$, and <> represents the cross correlation operation. The energy ratio of the harmonic component, r_h in each track is shown in Fig. 3 and Fig. 4, where several γs were examined. The tendency that pitched instrumental tracks and the bass drum track belong to $h(t)$ and other percussion tracks to $p(t)$ is consistent to the auditory evaluation. We can see that the separation performance gets worse as γ is larger, especially for drum tracks. Thus, the range compression facilitates the separation and $\gamma \simeq 0.3$ seems to give the best separation performance for these two songs. Probably, it is related to the actual distribution of spectrogram gradients of harmonic and percussive components.

6. CONCLUSION

In this paper, we derived a simple and fast method for a monaural audio signal into harmonic and percussive components, which is performed by the complementary diffusion on spectrogram. Despite of the simplicity of the algorithm, pitched instruments and drums are well separated. Applying the proposed algorithm as a pre-process to multi-pitch analysis or other tasks of music signal analysis is our current concern.

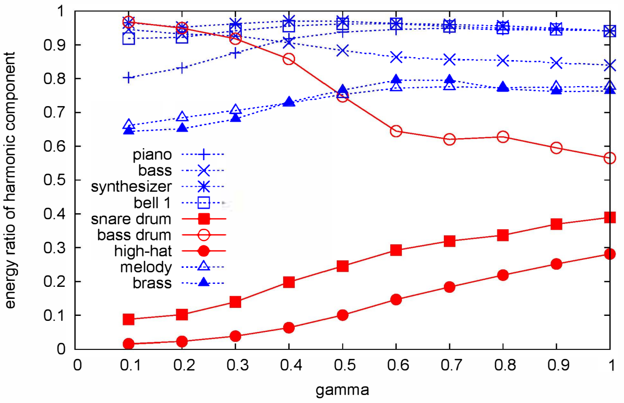

Figure 3: The energy ratio of the separated harmonic component in each track of RWC-MDB-P-2001 No.18

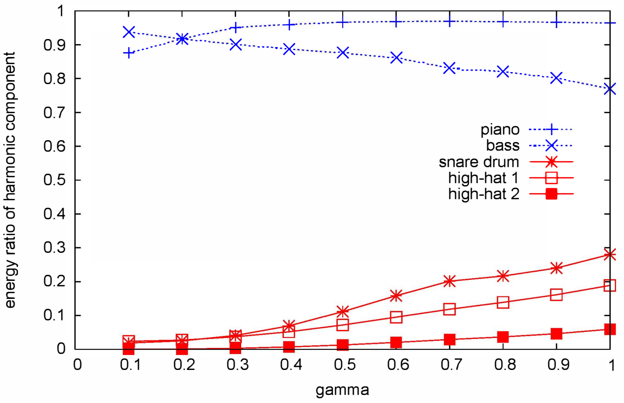

Figure 4: The energy ratio of the separated harmonic component in each track of RWC-MDB-J-2001 No.16

REFERENCES

[1] http://www.music-ir.org/mirex2007/index.php

[2] C. Uhle, C. Dittmar and T. Sporer, “Extraction of drum tracks from polyphonic music using independent subspace analysis,” *Proc. ICA*, pp. 843–847, Apr. 2003.

[3] M. Helen and T. Virtanen, “Separation of drums from polyphonic music using non-negative matrix factorization and support vector machine,” *Proc. EUSIPCO*, Sep. 2005.

[4] K. Itoyama, M. Goto, K. Komatani, T. Ogata, and H. Okuno, “Integration and Adaptation of Harmonic and Inharmonic Models for Separating Polyphonic Musical Signals,” *Proc, ICASSP*, pp. 57–60, Apr. 2007.

[5] L. Daudet, “A Review on Techniques for the Extraction of Transients in Musical Signals,” *Proc. CMMR*, pp. 219-232, 2005.

[6] D. P. W. Ellis and R. J. Weiss, “Model-based monaural source separation using a vector-quantized phase-vocoder representation,” *Proc. ICASSP*, pp. 957-960, May 2006.

[7] P. Smaragdis, B. Raj, and M. Shashanka, “Supervised and Semi-supervised Separation of Sounds from Single-Channel Mixtures,” *Proc. ICA* pp. 414-421, Sep. 2007.

[8] J. Le Roux, H. Kameoka, N. Ono, A. de Cheveigne, S. Sagayama, “Single Channel Speech and Background Segregation through Harmonic-Temporal Clustering,” *Proc. WASPAA*, Oct, 2007.

[9] 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, *Proc. ICASSP2008*, pp. 113-116, Apr. 2008.

[10] D. D. Lee and H. S. Seung, “Algorithms for Non-Negative Matrix Factorization” *Proc. NIPS*, pp. 556–562, 2000.

[11] H. Kameoka, T. Nishimoto, S. Sagayama, “A Multipitch Analyzer Based on Harmonic Temporal Structured Clustering,” *IEEE Trans. on Audio, Speech and Language Processing*, vol. 15, no. 3, pp.982-994, Mar. 2007.

[12] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka, “RWC music database: Popular, classical, and jazz music databases,” *Proc. ISMIR*, pp. 287-288, Oct. 2002.