A REAL-TIME EQUALIZER OF HARMONIC AND PERCUSSIVE COMPONENTS IN MUSIC SIGNALS

Nobutaka Ono, Kenichi Miyamoto, 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
E-mail: {onono,miyamoto,kameoka,sagayama}@hil.t.u-tokyo.ac.jp

ABSTRACT

In this paper, we present a real-time equalizer to control a volume balance of harmonic and percussive components in music signals without a priori knowledge of scores or included instruments. The harmonic and percussive components of music signals have much different structures in the power spectrogram domain, the former is horizontal, while the latter is vertical. Exploiting the anisotropy, our methods separate input music signals into them based on the MAP estimation framework. We derive two kind of algorithm based on a I-divergence-based mixing model and a hard mixing model. Although they include iterative update equations, we realized the real-time processing by a sliding analysis technique. The separated harmonic and percussive components are finally remixed in an arbitrary volume balance and played. We show the prototype system implemented on Windows environment.

1 INTRODUCTION

A graphic equalizer is one of the most popular tools on an audio player, which allows an user to control the volume balance between frequency bands as its preference by separating an input audio signal by several band-pass filters and remixing them with different gains. Recently, based on other kinds of separation, more advanced audio equalizations have been discussed and developed [1, 2, 3], which increase the variety of modifying audio sounds and enrich functions of audio players.

In this paper, focusing on two different components included in music signals: harmonic and percussive ones, we present a technique to equalize them in real-time without a priori knowledge of the scores or the included instruments. Not only as an extended audio equalizer, the technique should yield the useful pre-processing for various tasks related to music information retrieval from audio signals [4]. It can suppress percussive tones, which often interfere multipitch analysis, while, suppression of harmonic component will facilitate drum detection or rhythm analysis. We have currently applied this technique to automatic chord detection based on emphasized chroma features [5], rhythm pattern extraction and rhythm structure analysis [6], and melody extraction.

For independently equalizing the harmonic and percussive components, it is required to separate them. This kind of separation problem has been widely discussed in the literature. Uhle *et al.* applied Independent Component Analysis (ICA) to the magnitude spectrogram, and classified the extracted independent components into a harmonic and a percussive groups based on the several features like percussiveness, noise-likeness, etc [7]. Helen *et al.* utilized Nonnegative Matrix Factorization (NMF) for decomposing the spectrogram into elementary patterns and classified them by pre-trained Support Vector Machine (SVM) [8]. Through modeling harmonic and inharmonic tones on spectrogram, Itoyama *et al.* aimed to an instrument equalizer and proposed separation of an audio signal to each track based on the MIDI information synchronized to the input audio signal [1].

The contribution of this paper is to derive a simple and real-time algorithm specifically for the harmonic/percussive separation without any pre-learning or a priori knowledge of score or included instruments of the input audio signals. We present the formulation of the separation in Maximum A Priori (MAP) estimation framework, derive the fast iterative solution to it by auxiliary function approach, implement it with sliding update technique for real-time processing, and examine the performance by experiments to popular and jazz music songs.

2 FORMULATION OF HARMONIC/PERCUSSIVE SEPARATION

2.1 MAP Estimation Approach

Let $F_{\omega,\tau}$ be a Short Time Fourier Transform (STFT) of a monaural audio signal $f(t)$, and $W_{\omega,\tau} = |F_{\omega,\tau}|^2$ be a short time power spectrum, where ω and τ represent frequency and time bins. Let $H_{\omega,\tau}$ and $P_{\omega,\tau}$ be a harmonic and a percussive component of $W_{\omega,\tau}$, respectively. The variables $\boldsymbol{W}$, $\boldsymbol{H}$, and $\boldsymbol{P}$ denote a set of $W_{\omega,\tau}$, $H_{\omega,\tau}$, and $P_{\omega,\tau}$, respectively.

The separation of W into H and P is a kind of underdetermined blind source separation problem. One way to mathematically formulate this kind of problems is putting it on MAP (Maximum A Posteriori) estimation framework through representing desired source properties as a priori probabilities. Assuming that H and P are independent, the objective function of MAP estimation in our problem can be written as

$$\begin{aligned} & J(\boldsymbol{H}, \boldsymbol{P}) \\ = & \log p(\boldsymbol{H}, \boldsymbol{P}|\boldsymbol{W}) \\ = & \log p(\boldsymbol{W}|\boldsymbol{H}, \boldsymbol{P}) + \log p(\boldsymbol{H}, \boldsymbol{P}) + C \\ = & \log p(\boldsymbol{W}|\boldsymbol{H}, \boldsymbol{P}) + \log p(\boldsymbol{H}) + \log p(\boldsymbol{P}) + C, \quad (1) \end{aligned}$$

where the first term represents the log-likelihood, the second and the third terms represent the prior probabilities, and C is a constant term not including H and P, hereafter, we will omit it since it is not used for MAP estimation.

A harmonic component on the spectrogram usually has a stable pitch and form parallel ridges with smooth temporal envelopes, while the energy of a percussive tone is concentrated in a short time frame, which forms a vertical ridge with wideband spectral envelopes. Then typically, the vertical and horizontal structure emerges in the spectrogram of audio signals shown in the top of Fig. 3.

Focusing on the horizontal and vertical smoothed envelope of $H_{\omega,\tau}$ and $P_{\omega,\tau}$, we model their a priori distribution as functions of spectrogram gradients as:

$$p(\boldsymbol{H}) \propto \prod_{\omega,\tau} \frac{1}{\sqrt{2\pi}\sigma_H} \exp\left(-\frac{(H^{\gamma}_{\omega,\tau-1} - H^{\gamma}_{\omega,\tau})^2}{2\sigma_H^2}\right), \quad (2)$$

$$p(\boldsymbol{P}) \propto \prod_{\omega,\tau} \frac{1}{\sqrt{2\pi}\sigma_P} \exp\left(-\frac{(P^{\gamma}_{\omega-1,\tau} - P^{\gamma}_{\omega,\tau})^2}{2\sigma_P^2}\right), \quad (3)$$

where σ_H^2 and σ_P^2 are the variance of the spectrogram gradients, probably depending on the frame length or frame shift of STFT, and γ represents a range-compression factor such that $(0 < \gamma \leq 1)$, which we introduced for increasing the degree of freedom of our model with holding the assumption of the Gaussian distribution.

2.2 Method 1: I-divergence-based mixing model

Although $H_{\omega,\tau}$ and $P_{\omega,\tau}$ are the power spectrograms the additivity of them is not rigorously hold, $H_{\omega,\tau} + P_{\omega,\tau}$ should be close to the observation $W_{\omega,\tau}$. In several power-spectrogram-based signal processing methods NMF [9, 10, 11], the distance between power spectrograms $A_{\omega,\tau}$ and $B_{\omega,\tau}$ can be measured by I-divergence:

$$I(\boldsymbol{A}, \boldsymbol{B}) = \sum_{\omega,\tau} \left(A_{\omega,\tau} \log \frac{A_{\omega,\tau}}{B_{\omega,\tau}} - A_{\omega,\tau} + B_{\omega,\tau}\right), \quad (4)$$

which is almost equivalent to the assumption that $p(\boldsymbol{W}|\boldsymbol{H}, \boldsymbol{P})$ is Poisson distribution [11]. Assuming that observation at each time-frequency is independent, the log-likelihood term can be written as

$$\begin{aligned} & \log p(\boldsymbol{W}|\boldsymbol{H}, \boldsymbol{P}) - C \quad (5) \\ = & -\sum_{\omega,\tau} \left\{W_{\omega,\tau} \log \frac{W_{\omega,\tau}}{H_{\omega,\tau} + P_{\omega,\tau}} - W_{\omega,\tau} + H_{\omega,\tau} + P_{\omega,\tau}\right\}, \end{aligned}$$

where C is a constant term for normalization.

In the MAP estimation, the balance between a log-likelihood term and a prior distribution term is significant. Specifically in our problem, the relationship between them should be invariant for scaling. The property is satisfied by setting the range-compression factor as $\gamma = 0.5$. Then, the objective function can be written as

$$\begin{aligned} & J_1(\boldsymbol{H}, \boldsymbol{P}) \\ = & -\sum_{\omega,\tau} \left\{W_{\omega,\tau} \log \frac{W_{\omega,\tau}}{H_{\omega,\tau} + P_{\omega,\tau}} - W_{\omega,\tau} + H_{\omega,\tau} + P_{\omega,\tau}\right\} \\ & -\frac{1}{\sigma_H^2}(\sqrt{H_{\omega,\tau-1}} - \sqrt{H_{\omega,\tau}})^2 \\ & -\frac{1}{\sigma_P^2}(\sqrt{P_{\omega-1,\tau}} - \sqrt{P_{\omega,\tau}})^2). \quad (6) \end{aligned}$$

Note that, when $H_{\omega,\tau}$, $P_{\omega,\tau}$, and $W_{\omega,\tau}$ are multiplied by a scale parameter A, the objective function is also just multiplied by A and the function form is invariant.

2.3 Method 2: hard mixing model

Since the intersection of the horizontal and vertical ridges is small, we can make a more strong assumption that they are approximately disjoint. In the case, $W_{\omega,\tau} = H_{\omega,\tau}$ or $W_{\omega,\tau} = P_{\omega,\tau}$ are exclusively satisfied at each (ω, τ). However, the sparse mixing model leads us to a large number of combination problem. For avoiding it and obtaining an approximative solution, we cast it to a hard mixing model on the range-compressed power spectrum as

$$\hat{W}_{\omega,\tau} = \hat{H}_{\omega,\tau} + \hat{P}_{\omega,\tau}, \quad (7)$$

where

$$\hat{W}_{\omega,\tau} = W^{\gamma}_{\omega,\tau}, \ \ \hat{H}_{\omega,\tau} = H^{\gamma}_{\omega,\tau}, \ \ \hat{P}_{\omega,\tau} = P^{\gamma}_{\omega,\tau}. \quad (8)$$

Eq. (7) is hold if $H_{\omega,\tau}$ and $P_{\omega,\tau}$ are actually disjoint. Although the model is rough, this assumption leads us to simple formulation and solution. Since the deterministic mixing model of eq. (7) vanishes the log-likelihood term, the objective function is given by

$$\begin{aligned} J_2(\hat{\boldsymbol{H}}, \hat{\boldsymbol{P}}) \ = \ & -\frac{1}{2\sigma_H^2}\sum_{\omega,\tau}(\hat{H}_{\omega,\tau-1} - \hat{H}_{\omega,\tau})^2 \\ & -\frac{1}{2\sigma_P^2}\sum_{\omega,\tau}(\hat{P}_{\omega-1,\tau} - \hat{P}_{\omega,\tau})^2, \quad (9) \end{aligned}$$

under the constraint of eq. (7).

3 DERIVATION OF UPDATE EQUATIONS THROUGH AUXILIARY FUNCTION

3.1 Method 1

Maximizing eq. (6) is a nonlinear optimization problem. In order to derive an effective iterative algorithm, we introduce an auxiliary function approach, which has been recently utilized in several signal processing techniques such as NMF [9] and HTC (Harmonic Temporal Clustering) [10].

Note that the following auxiliary function:

$$\begin{aligned}
&Q_1(\boldsymbol{H}, \boldsymbol{P}, \boldsymbol{m_P}, \boldsymbol{m_H}) \\
&= -\sum_{\omega,\tau} m_{P\omega,\tau} W_{\omega,\tau} \log\left(\frac{m_{P\omega,\tau} W_{\omega,\tau}}{P_{\omega,\tau}}\right) \\
&\quad -\sum_{\omega,\tau} m_{H\omega,\tau} W_{\omega,\tau} \log\left(\frac{m_{H\omega,\tau} W_{\omega,\tau}}{H_{\omega,\tau}}\right) \\
&\quad -\frac{1}{\sigma_H^2}(\sqrt{H_{\omega,\tau-1}} - \sqrt{H_{\omega,\tau}})^2 \\
&\quad -\frac{1}{\sigma_P^2}(\sqrt{P_{\omega-1,\tau}} - \sqrt{P_{\omega,\tau}})^2) \qquad (10)
\end{aligned}$$

holds

$$J_1(\boldsymbol{H}, \boldsymbol{P}) \geq Q_1(\boldsymbol{H}, \boldsymbol{P}, \boldsymbol{m_P}, \boldsymbol{m_H}), \qquad (11)$$

for any $\boldsymbol{H}, \boldsymbol{P}, \boldsymbol{m_P}$, and $\boldsymbol{m_H}$ under the condition that

$$m_{P\omega,\tau} + m_{H\omega,\tau} = 1, \qquad (12)$$

where $m_{P\omega,\tau}$ and $m_{H\omega,\tau}$ are auxiliary variables and $\boldsymbol{m_P}$ and $\boldsymbol{m_H}$ are sets of $m_{P\omega,\tau}$ and $m_{H\omega,\tau}$, respectively. The equality of eq. (10) is satisfied for

$$m_{X\omega,\tau} = \frac{X_{\omega,\tau}}{H_{\omega,\tau} + P_{\omega,\tau}}, \qquad (13)$$

for $X = H$ or $X = P$. Then, updating $\boldsymbol{m_H}$ and $\boldsymbol{m_P}$ by eq. (13) increases the auxiliary function Q_1 and it achieves to J. After that, updating $\boldsymbol{H}$ and $\boldsymbol{P}$ by solving $\partial Q_1/\partial P_{\omega,\tau} = 0$ and $\partial Q_1/\partial H_{\omega,\tau} = 0$ increases Q_1 again and J_1 increases together because of the inequality of eq. (10). Hence, the iterations increases J_1 monotonically.

From $\partial Q_1/\partial P_{\omega,\tau} = 0$, $\partial Q_1/\partial H_{\omega,\tau} = 0$, and eq. (13), we have the following update equations:

$$H_{\omega,\tau} \leftarrow \left(\frac{b_{H\omega,\tau} + \sqrt{b_{H\omega,\tau}^2 + 4a_{H\omega,\tau}c_{H\omega,\tau}}}{2a_{H\omega,\tau}}\right)^2 \qquad (14)$$

$$P_{\omega,\tau} \leftarrow \left(\frac{b_{P\omega,\tau} + \sqrt{b_{P\omega,\tau}^2 + 4a_{P\omega,\tau}c_{P\omega,\tau}}}{2a_{P\omega,\tau}}\right)^2 \qquad (15)$$

$$m_{H\omega,\tau} \leftarrow \frac{H_{\omega,\tau}}{H_{\omega,\tau} + P_{\omega,\tau}} \qquad (16)$$

$$m_{P\omega,\tau} \leftarrow \frac{P_{\omega,\tau}}{H_{\omega,\tau} + P_{\omega,\tau}} \qquad (17)$$

where

$$\begin{aligned}
a_{H\omega,\tau} &= \frac{2}{\sigma_H^2} + 2, \quad c_{H\omega,\tau} = 2m_{H\omega,\tau}W_{\omega,\tau}, \qquad (18) \\
b_{H\omega,\tau} &= \frac{(\sqrt{H_{\omega,\tau-1}} + \sqrt{H_{\omega,\tau+1}})}{\sigma_H^2}, \qquad (19) \\
a_{P\omega,\tau} &= \frac{2}{\sigma_P^2} + 2, \quad c_{P\omega,\tau} = 2m_{P\omega,\tau}W_{\omega,\tau}, \qquad (20) \\
b_{P\omega,\tau} &= \frac{(\sqrt{P_{\omega-1,\tau}} + \sqrt{P_{\omega+1,\tau}})}{\sigma_P^2}. \qquad (21)
\end{aligned}$$

3.2 Method 2

Since eq. (9) is a quadrature form of $H_{\omega,\tau}$ and $P_{\omega,\tau}$ with a linear constraint, the optimal $\boldsymbol{H}$ and $\boldsymbol{P}, \boldsymbol{m}$ can be obtained by solving a simultaneous equation but it includes a large number of variables equal to the number of time-frequency bins. To avoid it and derive a simple iterative solution, we derived the following auxiliary function:

$$\begin{aligned}
&Q_2(\boldsymbol{H}, \boldsymbol{P}, \boldsymbol{U}, \boldsymbol{V}) \\
&= -\frac{1}{\sigma_H^2}\sum_{\omega,\tau}\left\{(\hat{H}_{\omega,\tau-1} - U_{\omega,\tau})^2 + (\hat{H}_{\omega,\tau} - U_{\omega,\tau})^2\right\} \\
&\quad -\frac{1}{\sigma_P^2}\sum_{\omega,\tau}\left\{(\hat{P}_{\omega-1,\tau} - V_{\omega,\tau})^2 + (\hat{P}_{\omega,\tau} - V_{\omega,\tau})^2\right\} (22)
\end{aligned}$$

satisfies

$$J_2(\boldsymbol{H}, \boldsymbol{P}) \geq Q_2(\boldsymbol{H}, \boldsymbol{P}, \boldsymbol{U}, \boldsymbol{V}), \qquad (23)$$

where $U_{\omega,\tau}$ and $V_{\omega,\tau}$ are auxiliary variables and $\boldsymbol{U}$ and $\boldsymbol{V}$ are sets of $U_{\omega,\tau}$ and $V_{\omega,\tau}$, respectively. The equality of eq. (10) is satisfied for $U_{\omega,\tau} = (\hat{H}_{\omega,\tau-1} + \hat{H}_{\omega,\tau})/2$ and $V_{\omega,\tau} = (\hat{P}_{\omega-1,\tau} + \hat{P}_{\omega,\tau})/2$. By taking the constraint of eq. (7) into consideration and organizing variables, we have the following update rules, which guarantees to monotonically increase the objective function J_2. The detailed derivation is presented in [12].

$$\begin{aligned}
\Delta_{\omega,\tau} &\leftarrow \alpha\left(\frac{\hat{H}_{\omega,\tau-1} - 2\hat{H}_{\omega,\tau} + \hat{H}_{\omega,\tau+1}}{4}\right) \\
&\quad -(1-\alpha)\left(\frac{\hat{P}_{\omega-1,\tau} - 2\hat{P}_{\omega,\tau} + \hat{P}_{\omega+1,\tau}}{4}\right) \qquad (24) \\
\hat{H}_{\omega,\tau} &\leftarrow \min(\hat{W}_{\omega,\tau}, \max(\hat{H}_{\omega,\tau} + \Delta_{\omega,\tau}, 0)), \qquad (25) \\
\hat{P}_{\omega,\tau} &\leftarrow \hat{W}_{\omega,\tau} - \hat{H}_{\omega,\tau}, \qquad (26)
\end{aligned}$$

where

$$\alpha = \frac{\sigma_P^2}{\sigma_H^2 + \sigma_P^2}. \qquad (27)$$

In method 2, any γ is allowable. According to our experiments, setting γ to be about 0.3 gives a good performance.

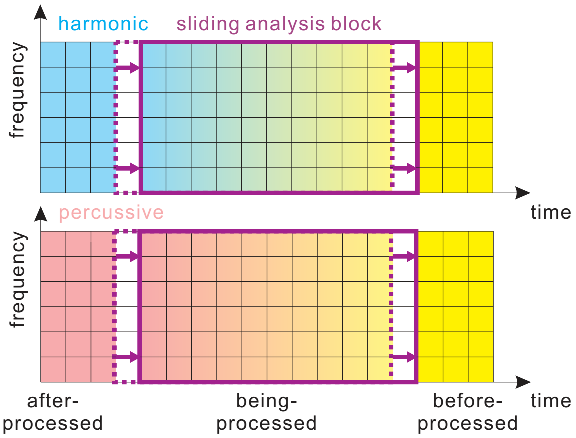


Figure 1. The process of the sliding block analysis

4 REAL-TIME PROCESSING BY SLIDING BLOCK ANALYSIS

Although the objective functions eq. (6) and eq. (9) should include all time-frequency bins, the iterative updates for the whole bins are much time-consuming. In order to obtain an approximate solution in real-time, we propose a sliding update algorithm. Based on the assumption that the separation of a certain time-frequency bin is weakly affected by far bins, we limit the processed frames to $n \leq \tau \leq n + B - 1$, where B is the size of the analysis block, and slide n iteratively. The real-time version of the Method 1 is summarized as follows.

1. Set the new frame as $H_{\omega,n+B-1} = P_{\omega,n+B-1} = W_{\omega,n+B-1}/2$.
2. Update variables by eq. (14), eq. (15), eq. (16), and eq. (17) for $n \leq \tau \leq n + B - 1$.
3. Convert the nth frame to a waveform by the inverse-STFT.
4. Increment n to slide the analysis block.

The real-time version of the Method 2 is in the same way. In step 3, the original phase is used for converting the STFT domain to the time domain. Note that the overlap of the frame shift should actually be considered for the conversion.

Each time-frequency bin is updated only once at step 2. Then, it is totally updated B times after passing through the analysis block shown in Fig. 1. Although the larger block size B shows better performance, the processing time from step 1 to step 4 must be less than the length of the frame shift for real-time processing.

5 IMPLEMENTATION AND EVALUATIONS

We implemented our algorithms in VC++ on Microsoft Windows environment. The GUI of the prototype system is shown in Fig. 2. After clicked a start button, the separation process begins. The processing steps are as folllows.

1. Loading a frame-shift-length fragment of the input audio signal from a WAV-formated file.
2. Calculating FFT for a new frame.
3. Updating stored frames as described in the previous section.
4. Calculating inverse FFT for the oldest frame.
5. Overlap-adding the waveform and playing it.
6. Go to Step 1.

The two bar graphs shown in Fig. 2 represent the power spectra of the separated harmonic and percussive component. The sliding bar named “P-H Balance” enables an user to change the volume balance between the harmonic and percussive components on play. The examples of the separated two spectrogram sequences are shown in Fig. 3. We can see that the input power spectrogram is sequentially separating in passing through the analysis block. In auditory evaluation, we observed:

- The pitched instrument tracks and the percussion tracks are well separated in both of method 1 and 2.
- Under the same analysis block size, the method 1 gives a little better performance than method 2.
- The method 1 requires about 1.5 $\sim$ 2 times computational time than the method 2 because of the calculation of square root. Thus, the method 2 allows the larger block size.
- The separation results depend on several parameters as σ_H, σ_P, the frame length, and the frame shift. But the dependency is not so large.

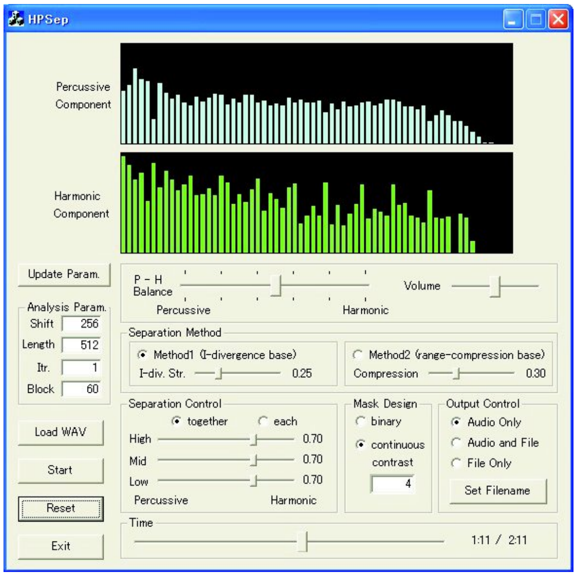
In order to quantitatively evaluate the performance of the harmonic/percussive separation and the relationship to the block size, we prepared each track data of two music pieces (RWC-MDB-J-2001 No.16 and RWC-MDB-P-2001 No.18 in [13]) by MIDI-to-WAV conversion and inputed the summation of all tracks to our algorithms. As a criterion of the performance, the energy ratio of the harmonic component $h(t)$ and the percussive component $p(t)$ included in each track was calculated as

$$r_h = \frac{E_h}{E_h + E_p}, \quad r_p = \frac{E_p}{E_h + E_p}, \tag{28}$$

where

$$E_h = < f_i(t), h(t) >^2, \quad E_p = < f_i(t), p(t) >^2, \tag{29}$$

and $<>$ represents the cross correlation operation and $f_i(t)$ represents a normalized signal of each track. The results are shown in Fig. 4. The pitched instrument tracks and the percussion tracks are represented by solid and dotted lines,



HPSep

Analysis Param.
Shift 256
Length 512
Itr. 1
Block 60

Separation Method
● Method1 (I-divergence base) I-div. Str. 0.25
○ Method2 (range-compression base) Compression 0.30

Separation Control
● together ○ each
High 0.70
Mid 0.70
Low 0.70

Mask Design
○ binary
● continuous
contrast 4

Output Control
● Audio Only
○ Audio and File
○ File Only

Time 1:11 / 2:11

Figure 2. The GUI of the harmonic/percussive equalizer

Table 1. Experimental conditions

signal length	10s
sampling rate	16kHz
frame size	512
frame shift	256
range-compression factor (method 1)	$\gamma = 0.5$
range-compression factor (method 2)	$\gamma = 0.3$
gradient variance	$\sigma_P = \sigma_H = 0.3$

respectively. We can see that the separation was almost well performed. Only the bass drum track has a tendency to belong to the harmonic component, which can be considered due to the long duration. Fig. 4 also shows that a large block size is not required and the separation performance converges at the block size of 30 or 40 frames in this condition.

6 CONCLUSION

In this paper, we presented a real-time equalizer of harmonic and percussive components in music signals without any a priori knowledge of score and included instruments. In auditory evaluation and experiments, we confirmed the good performance. Based on our equalizer, applying existing audio modification technique as conventional equalizing, reverb, pitch-shift, etc., to harmonic/percussive components independently will yield more interesting effect. Applying it as pre-processing for multi-pitch analysis, chord detection, rhythm pattern extraction, is another interesting future work.

7 REFERENCES

[1] 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.

[2] J. Woodruff, B. Pardo, and R. Dannenberg, "Remixing stereo music with score-informed source separation," *Proc.ISMIR*, 2006.

[3] K. Yoshii, M. Goto, and H. G. Okuno, "INTER:D: A drum sound equalizer for controlling volume and timbre of drums," *Proc. EWIMT*, pp. 205–212, 2005.

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

[5] Y. Uchiyama, K. Miyamoto, T. Nishimoto, N. Ono, and S. Sagayama, "Automatic Chord Detection Using Harmonic Sound Emphasized Chroma from Musical Acoustic Signal," *Proc. ASJ Spring Meeting*, pp.901-902, Mar., 2008. (in japanese)

[6] E. Tsunoo, K Miyamoto, N Ono, and S. Sagayama, "Rhythmic Features Extraction from Music Acoustic Signals using Harmonic/Non-Harmonic Sound Separation," *Proc. ASJ Spring Meeting*, pp.905-906, Mar., 2008. (in japanese)

[7] 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.

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

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

[10] H. Kameoka, T. Nishimoto, S. Sagayama, "A Multipitch Analyzer Based on Harmonic Temporal Structured Clustering," *IEEE Trans. ASLP*, vol. 15, no. 3, pp.982-994, Mar. 2007.

[11] J. Le Roux, H. Kameoka, N. Ono, A. de Cheveigne, S. Sagayama, "Single and Multiple F0 Contour Estimation Through Parametric Spectrogram Modeling of Speech in Noisy Environments," *IEEE Trans. ASLP*, vol. 15, no. 4, pp.1135-1145, May., 2007.

[12] 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," *Proc. EUSIPCO,* Aug., 2008. (to appear)

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

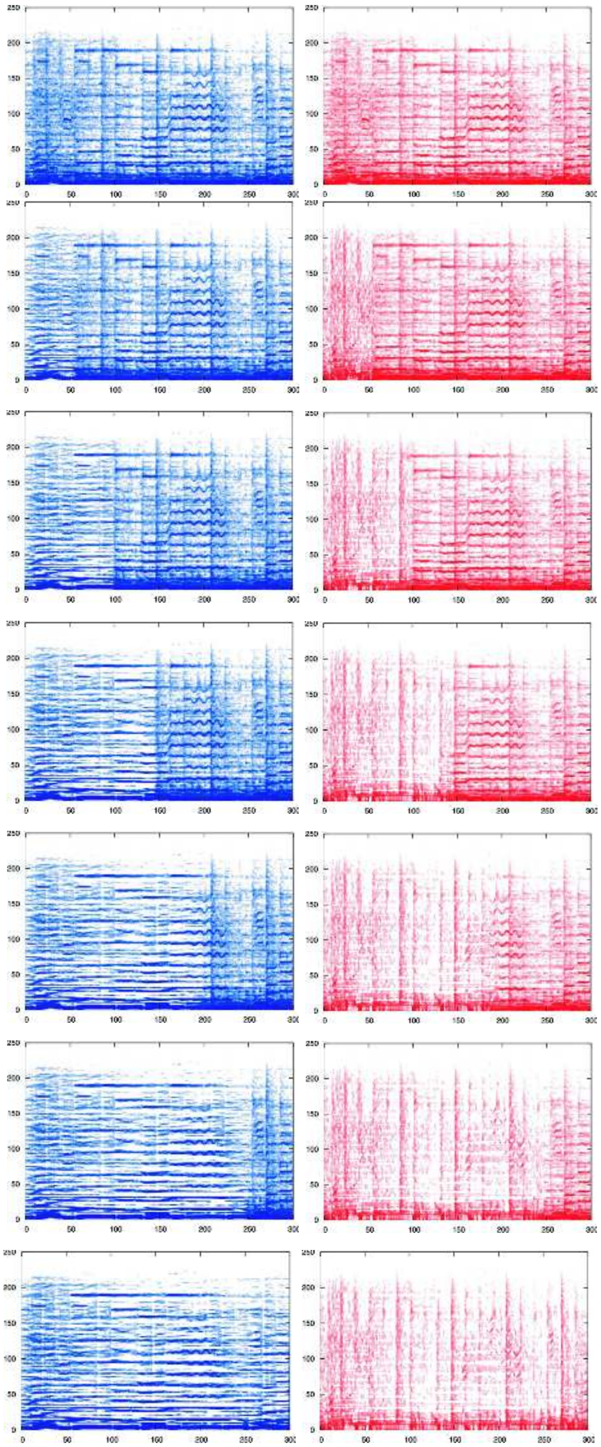


Figure 3. The spectrograms of separated harmonic component (left) and percussive component (right) by sliding block analysis. The first frame of the analysis block is 0, 10, 50, 100, 150, 200, and 250 from top to bottom, respectively.

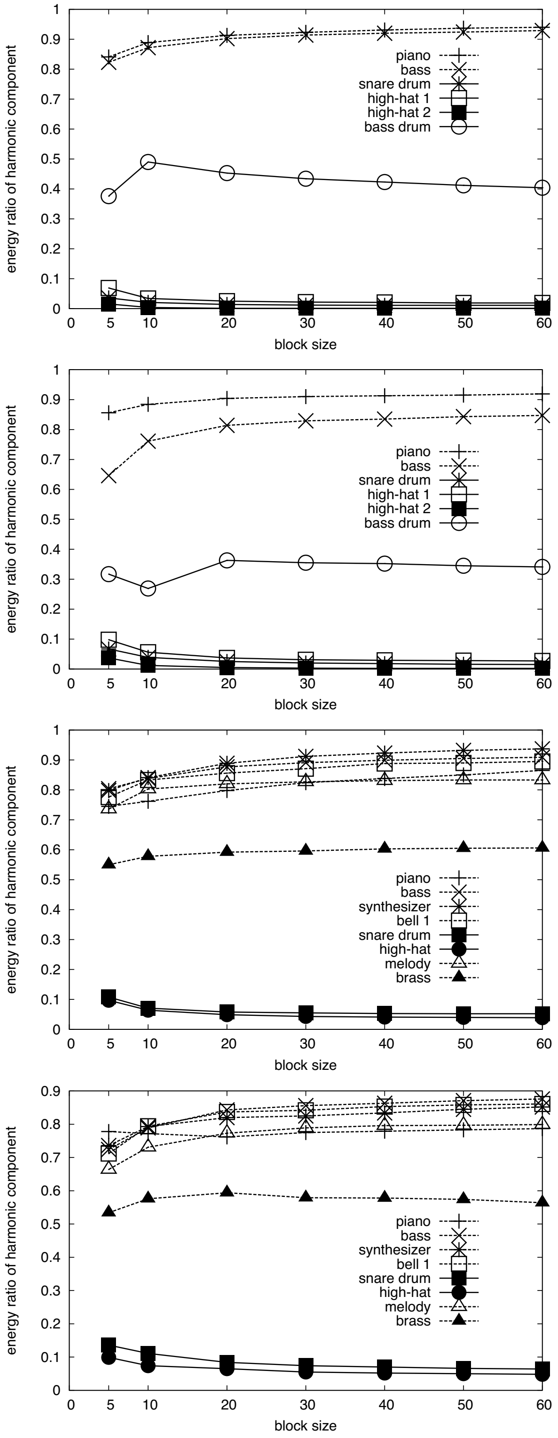


Figure 4. The energy ratio of the separated harmonic component in each track (r_h) for different block sizes. Their results from top to bottom are obtained by method 1 for RWC-MDB-J-2001 No.16, by method 2 for RWC-MDB-J-2001 No.16, by method 1 for RWC-MDB-P-2001 No.18, and by method 2 for RWC-MDB-P-2001 No.18, respectively.