TIMBRE REPLACEMENT OF HARMONIC AND DRUM COMPONENTS FOR MUSIC AUDIO SIGNALS

Tomohiko Nakamura[†‡], *Hirokazu Kameoka*[†], *Kazuyoshi Yoshii*[‡] *and Masataka Goto*[‡]

[†] Graduate School of Information Science and Technology, The University of Tokyo
[‡] National Institute of Advanced Industrial Science and Technology
{nakamura,kameoka}@hil.t.u-tokyo.ac.jp,{k.yoshii,m.goto}@aist.go.jp

ABSTRACT

This paper presents a system that allows users to customize an audio signal of polyphonic music (*input*), without using musical scores, by replacing the frequency characteristics of harmonic sounds and the timbres of drum sounds with those of another audio signal of polyphonic music (*reference*). To develop the system, we first use a method that can separate the amplitude spectra of the input and reference signals into harmonic and percussive spectra. We characterize frequency characteristics of the harmonic spectra by two envelopes tracing spectral dips and peaks roughly, and the input harmonic spectra are modified such that their envelopes become similar to those of the reference harmonic spectra. The input and reference percussive spectrograms are further decomposed into those of individual drum instruments, and we replace the timbres of those drum instruments in the input piece with those in the reference piece. Through the subjective experiment, we show that our system can replace drum timbres and frequency characteristics adequately.

Index Terms— Music signal processing, Harmonic percussive source separation, Nonnegative matrix factorization.

1. INTRODUCTION

Customizing existing musical pieces according to users' preferences is a challenging task in music signal processing. We would sometimes like to replace the timbres of instruments and audio textures of a musical piece with those of another musical piece. Professional audio engineers are able to perform such operations in the music production process by using effect units such as equalizers [1–5] that change the frequency characteristics of audio signals. However, sophisticated audio engineering skills are required for handling such equalizers effectively. It is therefore important to develop a new system that we can use intuitively without special skills.

Several highly functional systems have recently been proposed for intuitively customizing the audio signals of existing musical pieces. Itoyama *et al.* [6], for example, proposed an instrument equalizer that can change the volume of individual musical instruments independently. Yasuraoka *et al.* [7] developed a system that can replace the timbres and phrases of some instrument with users' own performances. Note that these methods are based on score-informed source separation techniques that require score information about the musical pieces (MIDI files). Yoshii *et al.* [8], on the other hand, developed a drum instrument equalizer called *Drumix* that can change the volume of bass and snare drums and replace their timbres and patterns with others prepared in advance. To achieve this, audio signals of bass and snare drums are separated from polyphonic audio signals without using musical scores. In this system, however, only the drum component can be changed or replaced. In addition, users would often need to prepare isolated drum sounds (called *reference*) with which they want to replace original drum sounds. Here we are concerned with developing an easier-to-handle system that only requires the users to specify a different musical piece as a reference.

In this paper, we propose a system that allows users to customize a musical piece (called *input*), without using musical scores, by replacing the timbres of drum instruments and the frequency characteristics of pitched instruments including vocals with those of another music piece (reference). We consider the problems of customizing the drum sounds and the pitched instruments separately, because they have different effects on audio textures. As illustrated in Fig. 1, the audio signals of the input and reference pieces are separated into harmonic and percussive components, respectively, by using a harmonic percussive source separation (HPSS) method [9] based on spectral anisotropy. The system then (1) analyzes the frequency characteristics of the spectra of the harmonic component (hereafter *harmonic spectra*) of the input piece and (2) adapts those characteristics to the frequency characteristics of the reference harmonic spectra. Moreover, (a) the spectrograms of the percussive components (hereafter *percussive spectrograms*) of the input and reference pieces are further decomposed into individual drum instruments such as bass and snare drums, and (b) the drum timbres of the input piece are replaced with those of the reference piece. In the following, we describe a replacement method of frequency characteristics for harmonic spectra and a replacement method of drum timbres for percussive spectrograms.

2. FREQUENCY CHARACTERISTICS REPLACEMENT

The goal is to modify the frequency characteristics of the harmonic spectra obtained with HPSS from an input piece by referring to those of a reference piece. The frequency characteristics of a musical piece are closely related to the timbres of the musical instruments used in that piece. If score information is available, a music audio signal could be separated into individual instrument parts [6, 7]. However, blind source separation is still difficult when score information is not available. We therefore take a different approach to avoid the need for perfect separation.

We here modify the input amplitude spectrum using two envelopes, named *bottom and top envelopes*, which trace the dips and peaks of the spectrum roughly as illustrated in Fig. 2. The bottom envelope expresses a flat and wide-band component in the spectrum, and the top envelope represents a spiky component in the spectrum. We can assume that the flat component corresponds to the spectrum of vocal consonants and attack sounds of musical instruments, while the spike component corresponds to the harmonic structures of musical instruments. Thus, individually modifying these envelopes allows us to approximately change the frequency characteristics of the musical instruments. The modified amplitude spectra are converted into an audio signal using the phases of the input harmonic spectra.

This study was supported in part by the JST OngaCREST project.

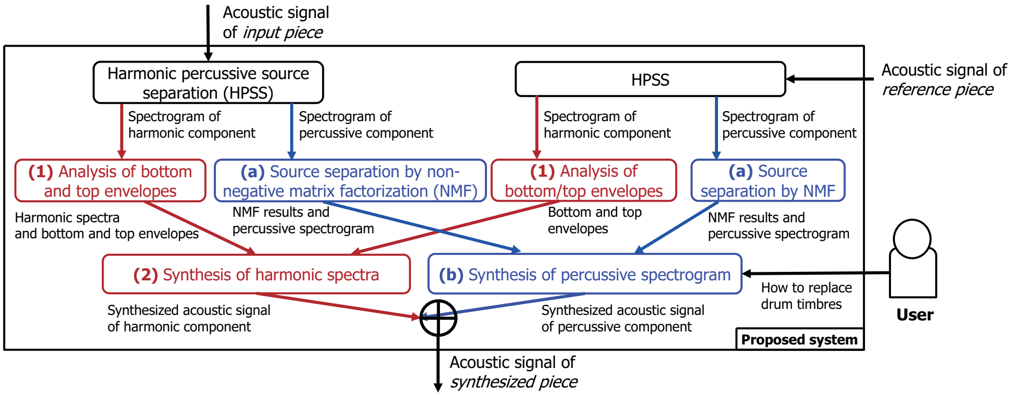
Acoustic signal of *input piece*
Harmonic percussive source separation (HPSS)
Spectrogram of harmonic component
Spectrogram of percussive component
(1) Analysis of bottom and top envelopes
(a) Source separation by non-negative matrix factorization (NMF)
Harmonic spectra and bottom and top envelopes
NMF results and percussive spectrogram
(2) Synthesis of harmonic spectra
(b) Synthesis of percussive spectrogram
Synthesized acoustic signal of harmonic component
Synthesized acoustic signal of percussive component
HPSS
Acoustic signal of *reference piece*
Spectrogram of harmonic component
Spectrogram of percussive component
(1) Analysis of bottom/top envelopes
(a) Source separation by NMF
Bottom and top envelopes
NMF results and percussive spectrogram
How to replace drum timbres
User
Proposed system
Acoustic signal of *synthesized piece*

Fig. 1. System outline for replacing drum timbres and frequency characteristics of the harmonic component. Red and blue modules relate to harmonic and percussive components of input and reference pieces.

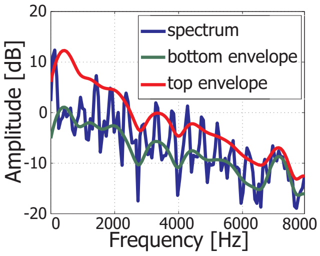

Fig. 2. Bottom (green) and top (red) envelopes of a spectrum (blue). The envelopes trace dips and peaks of a spectrum roughly.

2.1. Mathematical model for bottom and top envelopes

We describe each envelope using a Gaussian mixture model (GMM) as a function of the frequency ω:

$$\Psi(\omega; \mathbf{a}) := \sum_k a_k \psi_k(\omega), \ \psi_k(\omega) := \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big[-\frac{1}{2\sigma^2}\Big(\omega - \frac{k f_{\mathrm{nyq}}}{K}\Big)\Big] \tag{1}$$

where $\mathbf{a} := \{a_k\}_{k=1}^{K}$, and f_{nyq} stands for a Nyquist frequency. $a_k \geq 0$ denotes the power of the k-th Gaussian $\psi_k(\omega)$ with the average $k f_{\mathrm{nyq}}/K$ and the variance σ^2.

We first estimate $\mathbf{a}$ for the bottom envelopes of the input and reference pieces respectively by fitting $\Psi(\omega; \mathbf{a})$ to their harmonic spectra, and also estimate $\mathbf{a}$ for the top envelopes (see Sec. 2.3). We then design a filter that converts the input envelopes so that their time averages and variances equal those of the reference envelopes. Finally, by using the converted version of the input envelopes, we convert the input amplitude spectra.

2.2. Spectral synthesis via bottom and top envelopes

We consider converting the input piece so that the bottom and top envelopes of the converted version become similar to those of the reference piece. Let us define the averages and variances in time of the envelopes of the input and reference harmonic spectra as $\mu_\omega^{(l)}$ and $V_\omega^{(l)}$ for l = in, ref, respectively. Assuming that the envelopes follow normal distributions, the distributions of the converted input envelopes approach those of the reference envelopes by minimizing a measure between the distributions. As one such measure, we can use the Kullback-Leibler divergence, and derive the gains as

$$g_\omega = \frac{\mu_\omega^{(\mathrm{in})}\mu_\omega^{(\mathrm{ref})} + \sqrt{(\mu_\omega^{(\mathrm{in})}\mu_\omega^{(\mathrm{ref})})^2 - 4\{V_\omega^{(\mathrm{in})} + (\mu_\omega^{(\mathrm{in})})^2\}V_\omega^{(\mathrm{ref})}}}{2\{V_\omega^{(\mathrm{in})} + (\mu_\omega^{(\mathrm{in})})^2\}}. \tag{2}$$

Next, we show the conversion rule for the harmonic amplitude spectrum ($S_\omega^{(\mathrm{in})}$) of the input piece by using the gains for the bottom and top envelopes in the log-spectral domain. When modifying the bottom envelope, we want to modify only the flat component (and keep the spiky component fixed). On the other hand, when modifying the top envelope, we want to modify only the spiky component (and keep the flat component fixed). To do this, we multiply the spectral components above or near the top envelope by $g_{\mathrm{top},\omega}$ (the gain factor for the top envelope), and multiply the spectral components below or near the bottom envelope by $g_{\mathrm{bot},\omega}$ (the gain factor for the bottom envelope). One such rule is a threshold-based rule which means that we divide the set of spectral components into two sets, one consisting of the components above or near the top envelope and the other consisting of the components below or near the bottom envelope. We multiply the former and latter sets by $g_{\mathrm{top},\omega}$ and $g_{\mathrm{bot},\omega}$, respectively. Fig. 3 illustrates the rule where $S_\omega^{(\mathrm{synth})}$ is a synthesized amplitude spectrum and a threshold $\theta := \{\ln(\Psi(\omega; \mathbf{a}_{\mathrm{bot}})\Psi(\omega; \mathbf{a}_{\mathrm{top}}))\}/2$ is the midpoint of the bottom and top envelopes ($\Psi(\omega; \mathbf{a}_{\mathrm{bot}})$ and $\Psi(\omega; \mathbf{a}_{\mathrm{top}})$) of the input piece in the log-spectral domain. However, the rule changes spectral elements near θ with discontinuity. To avoid the discontinuity, we use the relaxed rule as shown in Fig. 3:

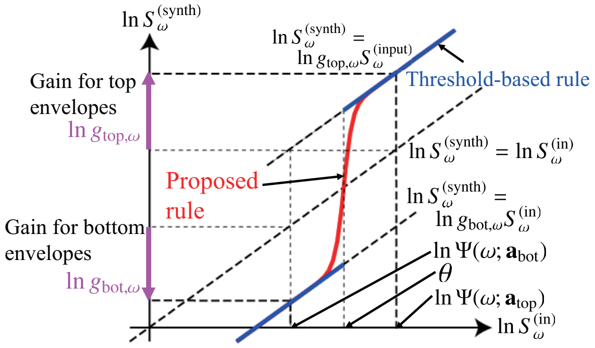
$\ln S_\omega^{(\mathrm{synth})}$
$\ln S_\omega^{(\mathrm{synth})} =$
$\ln g_{\mathrm{top},\omega} S_\omega^{(\mathrm{input})}$
Gain for top envelopes
$\ln g_{\mathrm{top},\omega}$
Threshold-based rule
$\ln S_\omega^{(\mathrm{synth})} = \ln S_\omega^{(\mathrm{in})}$
Proposed rule
$\ln S_\omega^{(\mathrm{synth})} =$
$\ln g_{\mathrm{bot},\omega} S_\omega^{(\mathrm{in})}$
Gain for bottom envelopes
$\ln g_{\mathrm{bot},\omega}$
$\ln \Psi(\omega; \mathbf{a}_{\mathrm{bot}})$
θ
$\ln \Psi(\omega; \mathbf{a}_{\mathrm{top}})$
$\ln S_\omega^{(\mathrm{in})}$

Fig. 3. The proposed (red curve) and threshold-based (blue lines) conversion rules of an input spectral element into a synthesized one in the log-spectral domain. The horizontal and vertical axes are an amplitude spectral elements of input and synthesized pieces.

$$\ln S_\omega^{(\mathrm{synth})} = \ln g_{\mathrm{bot},\omega} S_\omega^{(\mathrm{in})} + \ln \frac{g_{\mathrm{top},\omega}}{g_{\mathrm{bot},\omega}} f\Big(\frac{\ln S_\omega^{(\mathrm{in})} - \theta}{\rho \ln(\Psi(\omega; \mathbf{a}_{\mathrm{top}})/\Psi(\omega; \mathbf{a}_{\mathrm{bot}}))}\Big) \tag{3}$$

$$f(x) := \frac{1}{1 + \exp(-x)} = \begin{cases} 0 & (x \to -\infty) \\ 1 & (x \to \infty) \end{cases} \tag{4}$$

where $\rho > 0$. Note that (3) is equivalent to the threshold-based rule when $\rho \to 0$.

2.3. Estimation of bottom and top envelopes

2.3.1. Estimation of bottom envelopes

When estimating the bottom envelope $\Psi(\omega; \mathbf{a})$, we can use the Itakura-Saito divergence (IS divergence) [10] as a cost function. The estimation requires a cost function that is lower for the spectral dips than for the spectral peaks. The IS divergence meets the requirement as illustrated in Fig. 4. Let S_ω be an amplitude spectrum. The cost function is described as

$$\mathcal{J}_{\mathrm{bot}}(\mathbf{a}) := \sum_\omega D_{IS}(\Psi(\omega; \mathbf{a}) \| S_\omega), \tag{5}$$

$$D_{IS}(\Psi(\omega; \mathbf{a}) \| S_\omega) := \frac{\Psi(\omega; \mathbf{a})}{S_\omega} - \ln \frac{\Psi(\omega; \mathbf{a})}{S_\omega} - 1 \tag{6}$$

where $D_{IS}(\cdot||\cdot)$ is the IS divergence. Minimizing $\mathcal{J}_{\text{bot}}(\mathbf{a})$ directly is difficult, because of the non-linearity of the second term of (5).

We can use the auxiliary function method [11]. Given a cost function $\mathcal{J}$, we introduce an auxiliary variable λ and an auxiliary function $\mathcal{J}^{+}(x,\lambda)$ such that $\mathcal{J}(x) \leq \mathcal{J}^{+}(x,\lambda)$. We can then monotonically decrease $\mathcal{J}(x)$ indirectly by minimizing $\mathcal{J}^{+}(x,\lambda)$ with respect to x and λ iteratively.

The auxiliary function of $\mathcal{J}_{\text{bot}}(\mathbf{a})$ can be defined as

$$\mathcal{J}^{+}_{\text{bot}}(\mathbf{a},\boldsymbol{\lambda}) := \sum_{\omega}\Big\{\sum_{k}\Big(\frac{a_k\psi_k(\omega)}{S_{\omega}} - \lambda_k(\omega)\ln\frac{a_k\psi_k(\omega)}{\lambda_k(\omega)S_{\omega}}\Big) - 1\Big\} \tag{7}$$

where $\boldsymbol{\lambda} = \{\lambda_k(\omega)\}_{k=1,\omega=1}^{K,W}$ is a series of auxiliary variables such that $\forall\omega, \sum_k \lambda_k(\omega) = 1,\ \lambda_k(\omega) \geq 0$. The auxiliary function is obtained by Jensen's inequality based on the concavity of the logarithmic function in the second term of (5). By solving $\partial\mathcal{J}^{+}_{\text{bot}}(\mathbf{a},\boldsymbol{\lambda})/\partial a_k = 0$ and the equality condition of $\mathcal{J}_{\text{bot}}(\mathbf{a}) = \mathcal{J}^{+}_{\text{bot}}(\mathbf{a},\boldsymbol{\lambda})$, we can obtain

$$a_k \leftarrow \frac{\sum_{\omega}\lambda_k(\omega)}{\sum_{\omega}\psi_k(\omega)/S_{\omega}},\quad \lambda_k(\omega) \leftarrow \frac{a_k\psi_k(\omega)}{\sum_{k'}a_{k'}\psi_{k'}(\omega)}. \tag{8}$$

2.3.2. *Estimation of top envelopes*

The estimation of the top envelope $\Psi(\omega;\mathbf{a})$ requires a cost function that is higher for the spectral dips than for the spectral peaks. This is the opposite requirement for that in Sec. 2.3.1. The IS divergence is asymmetric as shown in Fig. 4, thus exchanging $\Psi(\omega;\mathbf{a})$ with S_{ω} of (6) leads to the opposite property to (6), and $D_{IS}(S_{\omega}||\Psi(\omega;\mathbf{a}))$ meets the requirement. Suppose that the bottom envelope $\Psi(\omega;\mathbf{a}_{\text{bot}})$ was estimated. The cost function is defined as

$$\mathcal{J}_{\text{top}}(\mathbf{a}) := P(\mathbf{a};\mathbf{a}_{\text{bot}}) + \sum_{\omega} D_{IS}(S_{\omega}||\Psi(\omega;\mathbf{a})) \tag{9}$$

where $P(\mathbf{a};\mathbf{a}_{\text{bot}}) := \sum_k \eta_k a_{\text{bot},k}/a_k$ is a penalty term for the closeness between the bottom and top envelopes, and $\eta_k \geq 0$ is the weight of $a_{\text{bot},k}/a_k$. Direct minimization of $\mathcal{J}_{\text{top}}(\mathbf{a})$ is also difficult because the IS divergence in the second term of (9) includes non-linear terms as described in (6).

Here we can define the auxiliary function of $\mathcal{J}_{\text{top}}(\mathbf{a})$ as

$$\begin{aligned}\mathcal{J}^{+}_{\text{top}}(\mathbf{a},\boldsymbol{\nu},\boldsymbol{h}) :=& P(\mathbf{a};\mathbf{a}_{\text{bot}}) + \sum_{\omega}\Big\{\sum_{k}\frac{(\nu_k(\omega))^2 S_{\omega}}{a_k\psi_k(\omega)} + \ln h(\omega)\\ &+ \frac{1}{h(\omega)}\Big(\sum_{k}a_k\psi_k(\omega) - h(\omega)\Big) - \ln S_{\omega} - 1\Big\}\end{aligned} \tag{10}$$

where $\boldsymbol{\nu} = \{\nu_k(\omega)\}_{k=1,\omega=1}^{K,W}$ and $\mathbf{h} = \{h(\omega)\}_{\omega=1}^{W}$ are series of auxiliary variables such that $\forall\omega, \sum_k \nu_k(\omega) = 1,\ \nu_k(\omega) \geq 0, h(\omega) > 0$. This inequality is derived from the following two inequalities for the non-linear terms:

$$\frac{1}{\sum_k x_k} \leq \sum_k \frac{\nu_k^2}{x_k},\ \ln x \leq \ln h + \frac{1}{h}(x-h). \tag{11}$$

where $\forall k, \nu_k \geq 0$ and $h > 0$ are auxiliary variables such that $\sum_k \nu_k = 1$. The first inequality is obtained by Jensen's inequality for $1/\sum_k x_k$ and the second inequality is a first-order Taylor-series approximation of $\ln x$ around h. By solving $\partial\mathcal{J}^{+}_{\text{top}}(\mathbf{a},\boldsymbol{\nu},\boldsymbol{h})/\partial a_k = 0$ and the equality condition of $\mathcal{J}_{\text{top}}(\mathbf{a}) = \mathcal{J}^{+}_{\text{top}}(\mathbf{a},\boldsymbol{\nu},\boldsymbol{h})$, update rules can be derived as

$$a_k \leftarrow \Big\{\frac{\eta_k a_{\text{bot},k} + \sum_{\omega}(\nu_k(\omega))^2 S_{\omega}/\psi_k(\omega)}{\sum_{\omega}\psi_k(\omega)/h(\omega)}\Big\}^{1/2}, \tag{12}$$

$$\nu_k(\omega) \leftarrow \frac{a_k\psi_k(\omega)}{\sum_{k'}a_{k'}\psi_{k'}(\omega)},\qquad h(\omega) \leftarrow \sum_k a_k\psi_k(\omega). \tag{13}$$

(12) does not guarantee $a_k \geq a_{\text{bot},k}$, and we set $a_k = a_{\text{bot},k}$ when $a_k < a_{\text{bot},k}$.

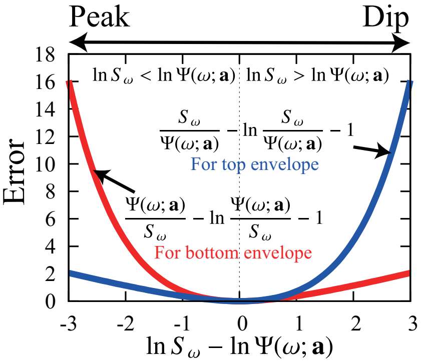

Fig. 4. The Itakura-Saito divergence for bottom and top envelopes.

3. DRUM TIMBRE REPLACEMENT

To replace drum timbres, we first decompose the percussive amplitude spectrograms into approximately those of individual drum instruments. The decomposition can be achieved by nonnegative matrix factorization (NMF) [12] and Wiener filtering. We call a component of the decomposed spectrograms *a basis spectrogram.* NMF approximates the amplitude spectrograms by a product of two nonnegative matrices, one of which is a basis matrix. Each column of the basis matrix corresponds to the amplitude spectrum of an individual drum sound, and the corresponding row of the activation matrix represents its temporal activity. The users are then allowed to specify which drum sounds (bases) in the input piece they want to replace with which drum sounds in the reference piece. According to this choice, the chosen drum timbres of the input piece are replaced with those of the reference piece for each basis.

3.1. Equalizing method

One simple method for replacing drum timbres, called *the equalizing (EQ) method*, is to apply gains to a basis spectrogram of the input piece such that the drum timbre of the input basis becomes similar to that of the reference basis. The input and reference bases represents the timbral characteristics of their drum sounds, and we use the gain that equalize the input and reference bases for each frequency bin. Let us define the complex basis spectrogram of the input piece and its basis as $Y^{(\text{in})}_{\omega,t}$ and $H^{(\text{in})}_{\omega}$. Using the corresponding reference basis $H^{(\text{ref})}_{\omega}$, we can obtain the synthesized complex spectrogram $Y^{(\text{synth})}_{\omega,t}$ for the basis as $Y^{(\text{synth})}_{\omega,t} = Y^{(\text{in})}_{\omega,t}H^{(\text{ref})}_{\omega}/H^{(\text{in})}_{\omega}$ for $\omega \in [1,W]$ and $t \in [1,T]$.

This method only requires applying gains to the input basis spectrograms uniformly in time. However, when there is a large difference between the timbres of the specified drum sounds, the method often amplifies low-energy frequency elements excessively, and so the resulting converted version would sound very noisy and the method fails to replace the drum timbres adequately.

3.2. Copy and paste method

To avoid the problem of the EQ method, we directly use basis spectrograms of the reference piece. The reference basis spectra include the drum timbre which we want, and by appropriately copying and pasting the reference basis spectra, we can obtain the percussive spectrogram with the reference drum timbres and the input temporal activities. We call the method *the copy and paste (CP) method.*

This method requires how to copy and paste the reference basis spectra with keeping the input temporal activities and how to reduce noise occured by this method. Features should be less sensitive to the drum timbres but reflect temporal activities. As the features, the NMF activations are available. Furthermore, there are three requirements related to the noise reduction. Noise occurs when previously remote high-energy spectra are placed adjacent to each other. To suppress the noise, (i) time-continuous segments should be used and (ii) the segment boundaries should be established when the activation is low. Since unsupervised source separation is still a challenging problem, the basis spectra may include a non-percussive component due to imperfect source separation, and (iii) the use of basis spectra that include non-percussive components should be avoided.

The problem can be formulated as an alignment problem. The requirements of (i), (ii), and (iii) are described as cost functions, and the cumulative cost $\mathcal{I}_t(\tau)$ can be written recursively as

$$\mathcal{I}_t(\tau) := \begin{cases} O_{t,\tau} & (t = 1) \\ O_{t,\tau} + \max_{\tau'}\{C_{\tau',\tau} + \mathcal{I}_{t-1}(\tau')\} & (t > 1) \end{cases}, \tag{14}$$

$$O_{t,\tau} := \alpha D(\tilde{U}_t^{(\mathrm{in})} \| \tilde{U}_\tau^{(\mathrm{ref})}) + \beta P_\tau \tag{15}$$

where τ is a time index of the reference piece, $\alpha > 0$ and $\beta > 0$ are the weights of $D(\tilde{U}_t^{(\mathrm{in})} \| \tilde{U}_\tau^{(\mathrm{ref})})$ and P_τ, and $\tilde{U}_t^{(l)} := U_t^{(l)} / \max_t\{U_t^{(l)}\}$ for $l = \mathrm{in}, \mathrm{ref}$. The first term of (15) indicates the generalized I-divergence between the two normalized activations. P_τ represents the degree to which the reference basis spectrum at the τ-th frame includes non-percussive components: the term becomes larger as the number of non-percussive components in the spectrum (requirement (iii)). $C_{\tau',\tau}$ is the transition cost from the τ'-th frame to the τ-th frame of the reference piece:

$$C_{\tau',\tau} = \begin{cases} 1 & (\tau = \tau' + 1) \\ c + \gamma(\tilde{U}_{\tau'}^{(\mathrm{ref})} + \tilde{U}_\tau^{(\mathrm{ref})}) & (\tau \neq \tau' + 1) \end{cases}. \tag{16}$$

The constant c expresses a cost for all other transitions except for a straight one. We set $c > 1$ and this ensures that a straight transition occurs more frequently than the others (requirement (i)). The second term of (16) for $\tau \neq \tau' + 1$ indicates that transitions to remote frames tend to occur when the activations are low (requirement (ii)), and $\gamma > 0$ is the weight of $\tilde{U}_{\tau'}^{(\mathrm{ref})} + \tilde{U}_\tau^{(\mathrm{ref})}$. We can obtain the alignment as an optimal path that minimizes the cumulative cost by the Viterbi algorithm [13].

The input basis spectra may include the non-percussive components because of imperfect source separation. In this case, the input basis spectra which may include the non-percussive components are replaced with the reference basis spectra by the CP method, and the input basis spectra loses the input non-percussive components. To recover the components, we make an extra processing. The components tend to have low energy, and they would probably be included in the input percussive spectra with low energy. We replace synthesized percussive spectra $\{Y_{\omega,t}^{(\mathrm{synth})}\}_\omega$ with the corresponding input percussive spectra $\{Y_{\omega,t}^{(\mathrm{in})}\}_\omega$ when $\sum_\omega Y_{\omega,t}^{(\mathrm{in})}$ is lower than a threshold ϵ.

4. EXPERIMENTAL EVALUATION

4.1. Experimental condition

We conducted an experiment to evaluate the performance of the system subjectively. We prepared three audio signals of musical pieces (10 s for each piece) from the RWC popular music and music genre databases [14] as input and reference pieces, and they were downsampled from 44.1 to 22.05 kHz. Then, we synthesized six pairs[1] of these musical audio signals. The signals of the input and reference pieces were converted into spectrograms with the short time Fourier transform (STFT) with a 512-sample Hanning window and a 256-sample frame shift, and the synthesized spectrograms were converted into audio signals by the inverse STFT with the same window and frame shift. The parameters of the frequency characteristics replacement were set at $\sigma = 240$ Hz and $(K, \rho, \eta_k) = (30, 0.2, 100/k)$ for $k \in [1, K]$. Then, the parameter a_k of the envelope model was initialized by $\sum_\omega S_\omega / K$ for $k \in [1, K]$, all frames and all pieces. For the NMF of the percussive spectrograms, we set the number of bases at 4, and used the generalized I-divergence. The CP method was compared with the EQ method, and one of the authors chose which drum sounds in the input piece were replaced with which drum sounds in the reference piece. The parameters for the drum timbre replacement were set at $(M, \alpha, \beta, \gamma, c, \epsilon) = (4, 0.5, 3, 10, 3, 100)$. A negative log posterior, which was computed by the L2-regularized L1-loss support vector classifier (SVC) [15], was used as P_τ, and the SVC was trained to distinguish between percussive and non-percussive instruments, using the RWC instrument database [14].

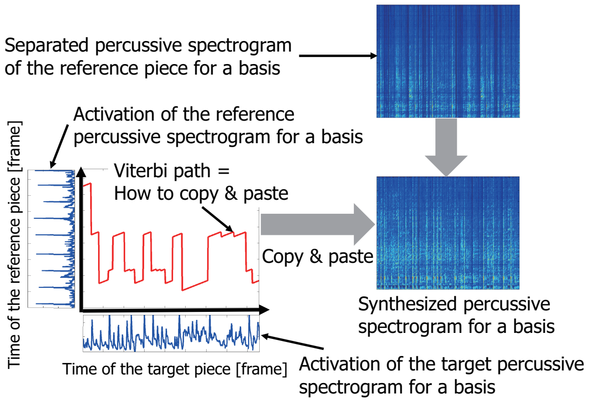
Separated percussive spectrogram of the reference piece for a basis
Activation of the reference percussive spectrogram for a basis
Time of the reference piece [frame]
Viterbi path = How to copy & paste
Copy & paste
Synthesized percussive spectrogram for a basis
Time of the target piece [frame]
Activation of the target percussive spectrogram for a basis

Fig. 5. Outline of the copy and paste method.

We asked 9 subjects how adequately they felt that (1) the drum timbres of the input piece were replaced with those of the reference piece and (2) the timbres of the input harmonic components were replaced with those of the reference piece. The subjects were allowed to listen to the input, reference, and synthesized pieces as well as their harmonic and percussive components as many times as they liked. They then evaluated (1) and (2) for each synthesized piece on a scale of 1 to 5. 1 point means that the timbres were not replaced and 5 points indicates that the timbres were replaced perfectly.

4.2. Result and discussion

The average scores of (1) with standard errors were 2.37 ± 0.15 and 2.83 ± 0.15 for the EQ and the CP methods. The CP method result was provided prior to that provided by the EQ method, in particular when the drum timbres were very different as we mentioned in Sec. 3. The average score of (2) with standard errors was 2.5 ± 0.1. The results show that the subjects perceived the replaced drum timbres and frequency characteristics, and that the system works well.

We asked the subjects to comment about the synthesized pieces. One subject said that he wanted to control the degree to which drum timbres and frequency characteristics were converted. This opinion indicates that it is important to enables users to adjust the conversions. Additionally, another subject mentioned that replacing vocal timbres separately would change the moods of the musical pieces more drastically. We plan to replace vocal timbres by using an extension of HPSS [16] for vocal extraction.

5. CONCLUSION

We have described a system that can replace the drum timbres and frequency characteristics of harmonic components in polyphonic audio signals without using musical scores. We have proposed an algorithm that can modify a harmonic amplitude spectrum via its bottom and top envelopes. We have also discussed two methods for replacing drum timbres. The EQ method applies gains to basis spectrograms by the proportions of the NMF bases of the input percussive spectrograms and those of the reference percussive spectrograms. The CP method copies and pastes the basis spectra of a reference piece, according to NMF activations of the input and reference pieces. Through the subjective experiment, we confirmed that the system can replace drum timbres and frequency characteristics adequately.

[1] Some synthesized sounds are available at `http://hil.t.u-tokyo.ac.jp/~nakamura/demo/TimbreReplacement.html`.

6. REFERENCES

[1] M. N. S. Swamy and K. S. Thyagarajan, “Digital bandpass and bandstop filters with variable center frequency and bandwidth,” *Proc. of IEEE*, vol. 64, no. 11, pp. 1632–1634, 1976.

[2] S. Erfani and B. Peikari, “Variable cut-off digital ladder filters,” *Int. J. Electron*, vol. 45, no. 5, pp. 535–549, 1978.

[3] E. C. Tan, “Variable lowpass wave-digital filters,” *Electron. Lett.*, vol. 18, pp. 324–326, 1982.

[4] P. A. Regalia and S. K. Mitra, “Tunable digital frequency response equalization filters,” *IEEE Trans. ASLP*, vol. 35, no. 1, pp. 118–120, 1987.

[5] S. J. Orfanidis, “Digital parametric equalizer design with prescribed Nyquist-frequency gain,” *J. of Audio Eng. Soc.*, vol. 45, no. 6, pp. 444–455, 1997.

[6] K. Itoyama, M. Goto, K. Komatani, T. Ogata, and H. G. Okuno, “Integration and adaptation of harmonic and inharmonic models for separating polyphonic musical signals,” in *Proc. of ICASSP*, 2007, vol. 1, pp. I–57–I–60.

[7] N. Yasuraoka, T. Abe, K. Itoyama, T. Takahashi, T. Ogata, and H. G. Okuno, “Changing timbre and phrase in existing musical performances as you like: manipulations of single part using harmonic and inharmonic models,” in *Proc. of ACM-MM*, 2009, pp. 203–212.

[8] K. Yoshii, M. Goto, K. Komatani, T. Ogata, and H. G. Okuno, “Drumix: An audio player with real-time drum-part rearrangement functions for active music listening,” *Trans. IPSJ*, vol. 48, no. 3, pp. 1229–1239, 2007.

[9] H. Tachibana, H. Kameoka, N. Ono, and S. Sagayama, “Comparative evaluation of multiple harmonic/percussive sound separation techniques based on anisotropic smoothness of spectrogram,” in *Proc. of ICASSP*, 2012, pp. 465–468.

[10] F. Itakura and S. Saito, “Analysis synthesis telephony based on the maximum likelihood method,” in *Proc. of ICA*, 1968, C-17–C-20.

[11] J. M. Ortega and W. C. Rheinboldt, *Iterative solution of nonlinear equations in several variables*, Number 30. 2000.

[12] D. Seung and L. Lee, “Algorithms for non-negative matrix factorization,” *Adv. Neural Inf. Process. Syst.*, vol. 13, pp. 556–562, 2001.

[13] A. Viterbi, “Error bounds for convolutional codes and an asymptotically optimum decoding algorithm,” *IEEE Trans. Inf. Theory*, vol. 13, no. 2, pp. 260–269, 1967.

[14] M. Goto, “Development of the RWC Music Database,” in *Proc. of ICA*, 2004, pp. l–553–556.

[15] R. Fan, K. Chang, C. Hsieh, X. Wang, and C. Lin, “LIBLINEAR: A library for large linear classification,” *JMLR*, vol. 9, pp. 1871–1874, 2008.

[16] H. Tachibana, T. Ono, N. Ono, and S. Sagayama, “Melody line estimation in homophonic music audio signals based on temporal variability of melodic source,” in *Proc. of ICASSP*, 2010, pp. 425–428.