REPLACING DRUM TIMBRES AND FREQUENCY CHARACTERISTICS BETWEEN DIFFERENT MUSICAL PIECES

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 (target) without using musical scores by replacing the timbres of drum sounds and the frequency characteristics of harmonic 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 target 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 target harmonic spectra are modified such that their envelopes become similar to those of the reference harmonic spectra. The target and reference percussive spectrograms are further decomposed into individual drum instruments, and we replace the timbres of those drum instruments in the target signal with those in the reference piece. The experimental results showed that our system can successfully let users perceive replacement of drum timbres and frequency characteristics.

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

1. INTRODUCTION

Customizing existing musical pieces according to users' preferences is one of the challenging tasks of music signal processing. Many people 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 can be intuitively used by ordinary people without special skills.

Several highly-functional systems have recently been proposed for intuitively customizing audio signals of existing musical pieces. Itoyama *et al.* [6], for example, proposed an instrument equalizer that can change the volumes of individual musical instrument parts independently. Yasuraoka *et al.* [7] developed a system that can replace the timbres and phrases of some instrument parts with users' own performances. Note that these methods are based on score-informed source separation techniques that require score information of musical pieces (MIDI files). Yoshii *et al.* [8], on the other hand, developed a drum instrument equalizer called *Drumix* that can change the volumes of bass and snare drums and replace their timbres and patterns with other ones 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 instruments can be changed or replaced. In addition, users would often suffer from having to prepare isolated drum sounds (called reference) with which they want to replace original drum sounds. Here we are concerned with developing a more easy-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 (target) 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 target and reference pieces are respectively separated into harmonic and percussive components 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 target piece and (2) adapts those characteristics to the frequency characteristics of the reference harmonic spectra. On the other hand, (a) the spectrograms of the percussive components (hereafter *percussive spectrograms*) of the target and reference pieces are further respectively decomposed into individual drum instruments such as bass and snare drums, and (b) the drum timbres of the target piece are replaced with those of the reference piece. In the following, we describe a frequency-characteristics replacer for harmonic spectra and a drum-timbre replacer for percussive spectrograms.

2. FREQUENCY CHARACTERISTICS REPLACER

The goal is to modify the frequency characteristics of the harmonic spectra obtained with HPSS from a target piece by referring to those of a reference piece. The frequency characteristics of a musical piece are closely related to the timbres of 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 a difficult open challenge when score information is not available. We therefore take a different approach to avoid the need for perfect separation.

We here modify the target 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 envelopes express a flat and wide-band component in the spectrum, and the top envelopes represent a spiky component in the spectrum. We can assume that this flat component corresponds to spectra 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 spectrum can be converted into an audio signal using the phase of the target harmonic spectra.

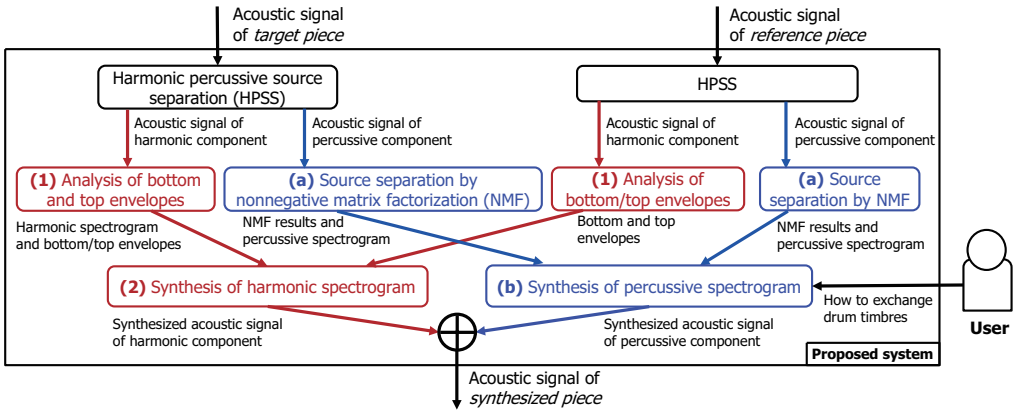

Fig. 1. System outline for replacing drum timbres and frequency characteristics of harmonic components. Red and blue modules relate to harmonic and percussive components of target and reference pieces.

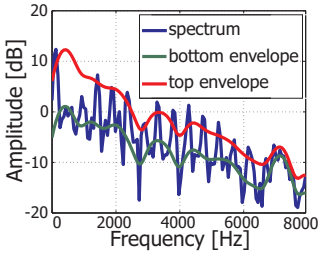

Fig. 2. Bottom (green) and top (red) envelopes of a spectrum (blue). The envelopes trace dips and peaks of a spectrum roughly.

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:

$$\Psi(\omega;\boldsymbol{a}) := \sum_k a_k \psi_k(\omega), \quad \psi_k(\omega) := \mathcal{N}\Big(\omega; \frac{k f_{\mathrm{nyq}}}{K}, \sigma^2\Big), \tag{1}$$

where $\boldsymbol{a} := \{a_k\}_{k=1}^K$ and f_{nyq} stands for a Nyquist frequency. a_k denotes a 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 a_k to fit $\Psi(\omega;\boldsymbol{a})$ to each spectrum for target and reference pieces, and obtain the bottom and top envelopes (see Sec. 2.3). We then design a filter that converts the target envelopes so that their time averages and variances become equal to those of the reference envelopes. Finally, by using the converted version of the target envelopes, we convert the target amplitude spectra.

2.2. Spectral synthesis via bottom and top envelopes

Here we show how to modify the harmonic spectra of a target piece such that their bottom and top envelopes become similar to those of the reference piece.

We consider converting the target piece so that the bottom and top envelopes of the converted version become similar to those of the reference piece. Let us define averages and variances in time of the envelopes of the harmonic spectra of the target and reference pieces as $\mu_\omega^{(l)}$ and $V_\omega^{(l)}$ for $l = \mathrm{tar}, \mathrm{ref}$. Assuming that the envelopes follow normal distributions, the distributions of the gained target envelopes approach to 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

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

Next, we show a conversion rule of the harmonic amplitude spectrum ($S_\omega^{(\mathrm{tar})}$) of the target piece by using the gains for the bottom and top envelopes in the log-spectral domain. Through the modification of the bottom envelope, we would want to modify the flat component only (and keep the spiky component fixed). On the other hand, through the modification of the top envelope, we would want to modify the spiky component only (and keep the flat component fixed). To do so, 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

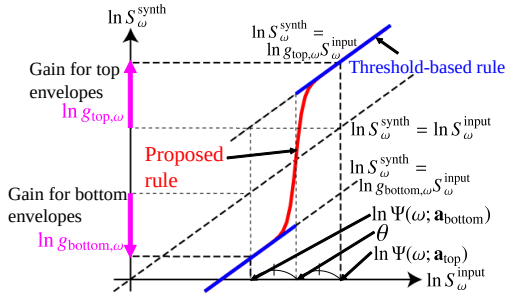

Fig. 3. The proposed (red curve) and threshold-based (blue lines) conversion rules of a target spectral element into a synthesized one in the log-spectral domain. The horizontal and vertical axes are an amplitude spectral element of target and synthesized pieces.

of spectral components into two sets, one consisting of the components that are above or near the top envelope and the other consisting of the components that are below or near the bottom envelope, and for the former set, we multiply them by $g_{\mathrm{top},\omega}$ and for the latter set, we multiply them by $g_{\mathrm{bot},\omega}$. Fig. 3 illustrates the rule where $S_\omega^{(\mathrm{synth})}$ is a synthesized amplitude spectrum and a threshold $\theta := \{\ln(\Psi(\omega;\boldsymbol{a}_{\mathrm{bot}})\Psi(\omega;\boldsymbol{a}_{\mathrm{top}}))\}/2$ is the midpoint of the bottom and top envelopes ($\Psi(\omega;\boldsymbol{a}_{\mathrm{bot}})$ and $\Psi(\omega;\boldsymbol{a}_{\mathrm{top}})$) of the target 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:

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

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

where $\rho > 0$. Notice 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*

Estimating the bottom envelope $\Psi(\omega;\boldsymbol{a})$ can uses the Itakura-Saito divergence (IS divergence) [10] as an objective function. The estimation requires an objective function that have smaller costs close to spectral dips than close to spectral peaks. The IS divergence meets the requirement as illustrated in Fig. 4. Let S_ω be an amplitude spec-

trum. The objective function is described as

$$\mathcal{J}_{\text{bot}}(\boldsymbol{a}) := \sum_{\omega}\Big(\frac{\Psi(\omega;\boldsymbol{a})}{S_\omega} - \ln\frac{\Psi(\omega;\boldsymbol{a})}{S_\omega} - 1\Big). \tag{5}$$

Minimizing $\mathcal{J}_{\text{bot}}(\boldsymbol{a})$ directly is difficult, because of the non-linearity of the second term of (5).

We can use the auxiliary function method [11]. Given an objective function $\mathcal{J}$, we introduce an auxiliary variable λ and an auxiliary function $\mathcal{J}^+(x,\lambda)$ such that $\mathcal{J}(x) \le \mathcal{J}^+(x,\lambda)$. We can then minimize $\mathcal{J}(x)$ indirectly by iteratively minimizing $\mathcal{J}^+(x,\lambda)$ for x and λ.

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

$$\mathcal{J}^+_{\text{bot}}(\boldsymbol{a},\boldsymbol{\lambda}) := \sum_{\omega}\Big\{\sum_{k}\Big(\frac{a_k\psi_k(\omega)}{S_\omega} - \lambda_k(\omega)\ln\frac{a_k\psi_k(\omega)}{\lambda_k(\omega)S_\omega}\Big) - 1\Big\} \tag{6}$$

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

$$a_k \leftarrow \frac{\sum_\omega \lambda_k(\omega)}{\sum_\omega \psi_k(\omega)/S_\omega},\ \lambda_k(\omega) \leftarrow \frac{a_k\psi_k(\omega)}{\sum_{k'} a_{k'}\psi_{k'}(\omega)}. \tag{7}$$

2.3.2. *Estimation of top envelopes*

For estimation of the top envelope $\Psi(\omega;\boldsymbol{a})$, the IS divergence can also be used as an objective function. The estimation requires an objective function that have larger costs close to spectral dips than those close to spectral peaks. The requirement is met by the IS divergence as shown in Fig. 4. Suppose that the bottom envelope $\Psi(\omega;\boldsymbol{a}_{\text{bot}})$ was estimated. The objective function is defined as

$$\mathcal{J}_{\text{top}}(\boldsymbol{a}) := \sum_{\omega}\Big(\frac{S_\omega}{\Psi(\omega;\boldsymbol{a})} - \ln\frac{S_\omega}{\Psi(\omega;\boldsymbol{a})} - 1\Big) + P(\boldsymbol{a};\boldsymbol{a}_{\text{bot}}), \tag{8}$$

where $P(\boldsymbol{a};\boldsymbol{a}_{\text{bot}}) := \sum_k \eta_k a_{\text{bot},k}/a_k$ is a penalty term for closeness between the bottom and top envelopes, and $\eta_k \ge 0$ weights a contribution of $a_{\text{bot},k}/a_k$ to $P(\boldsymbol{a};\boldsymbol{a}_{\text{bot}})$. The first and second terms of (8) are non-linear.

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

$$\begin{aligned}\mathcal{J}^+_{\text{top}}(\boldsymbol{a},\boldsymbol{\nu},\boldsymbol{h}) :=& P(\boldsymbol{a};\boldsymbol{a}_{\text{bot}}) + \sum_\omega\Big\{\sum_k \frac{(\nu_k(\omega))^2 S_\omega}{a_k\psi_k(\omega)} + \ln h(\omega)\\ &+ \frac{1}{h(\omega)}\Big(\sum_k a_k\psi_k(\omega) - h(\omega)\Big) - \ln S_\omega - 1\Big\}\end{aligned} \tag{9}$$

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

$$\frac{1}{\sum_k x_k} \le \sum_k \frac{\nu_k^2}{x_k}, \quad \ln x \le \ln h + \frac{1}{h}(x-h). \tag{10}$$

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

$$a_k \leftarrow \Big\{\frac{\eta_k a_{\text{bot},k} + \sum_\omega (\nu_k(\omega))^2 S_\omega/\psi_k(\omega)}{\sum_\omega \psi_k(\omega)/h(\omega)}\Big\}^{1/2} \tag{11}$$

where $\nu_k(\omega) \leftarrow (a_k\psi_k(\omega))/(\sum_{k'} a_{k'}\psi_{k'}(\omega))$ and $h(\omega) \leftarrow \sum_k a_k\psi_k(\omega)$. This rule does not guarantee $a_k \ge a_{\text{bot},k}$, and we set $a_k = a_{\text{bot},k}$ when $a_k < a_{\text{bot},k}$.

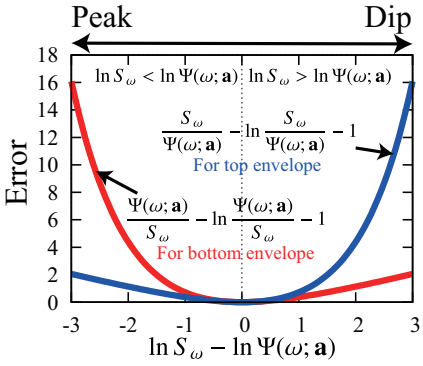

Fig. 4. The Itakura-Saito divergence for bottom and top envelopes.

3. DRUM TIMBRE REPLACER

To replace drum timbres, we first decompose the percussive amplitude spectrograms into those of individual drum instruments approximately. 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 being 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 target piece they want to replace with which drum sounds in the reference piece. According to this choice, the chosen drum timbres of the target 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 method*, is to gain a basis spectrogram of the target piece by the proportion of the target basis and the corresponding reference basis in all frequency bins. Let us define the complex basis spectrogram of the target piece and its basis as $Y^{(\text{tar})}_{\omega,t}$ and $H^{(\text{tar})}_{\omega}$. 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 $Y^{(\text{synth})}_{\omega,t} = Y^{(\text{tar})}_{\omega,t} H^{(\text{ref})}_{\omega}/H^{(\text{tar})}_{\omega}$ for $\omega \in [1,W]$ and $t \in [1,T]$.

This method gains the target basis spectrograms uniformly in time. When there is a large difference between the timbres of the specified drum sounds, the method often excessively amplifies low-energy frequency elements, and so the resulting converted version would sound very noisy and replacing the drum timbres tend not to be perceived by the users.

3.2. Alignment method

To avoid the problem of the equalizing method, we can directly use basis spectrograms of the reference piece. For this purpose, we need to align temporal activities of the target and reference basis spectra, since they correspond to different rhythms of drum instruments. By appropriately aligning the basis spectra of the reference piece, percussive spectrograms with the reference drum timbres and the target temporal activities can be obtained. We call the method *the alignment method*.

The alignment method requires features representing temporal activities, and we can use the activations as the features. Furthermore, three requirements arise to reduce noise. Noise occurs when previously remote high-energy spectra are placed adjacently. To suppress the noise, (i) time-continuous segments should be used and (ii) the segment boundaries should be placed when the activations are 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) it should be avoided to use the basis spectra including the non-percussive component.

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

$$\mathcal{I}_t(\tau) := \begin{cases} O_{t,\tau} & (t = 1) \\ O_{t,\tau} + \max_{\tau'}\{C_{\tau',\tau} + \mathcal{I}_{t-1}(\tau')\} & (t > 1) \end{cases}, \tag{12}$$

$$O_{t,\tau} := \alpha D(\tilde{U}_t^{(\mathrm{tar})} \| \tilde{U}_\tau^{(\mathrm{ref})}) + \beta P_\tau \tag{13}$$

where τ is a time index of the reference piece, $\alpha > 0$ and $\beta > 0$ weight contributions of the terms to $\mathcal{I}_t(\tau)$, and $\tilde{U}_t^{(l)} := U_t^{(l)} / \max_t\{U_t^{(l)}\}$ for $l = \mathrm{tar}, \mathrm{ref}$. The first term of (13) indicates the generalized I-divergence between the two normalized activations. P_τ represents how much the reference basis spectrum at the τ-th frame includes the non-percussive components, and the term becomes larger when the spectrum includes more non-percussive components (requirement (iii)). $C_{\tau',\tau}$ is a transition cost from the τ'-th frame to the τ-th frame of the reference piece:

$$C_{\tau',\tau} = \begin{cases} 1 & (\tau = \tau' + 1) \\ c + \gamma(\tilde{U}_{\tau'}^{(\mathrm{ref})} + \tilde{U}_\tau^{(\mathrm{ref})}) & (\tau \neq \tau' + 1) \end{cases}. \tag{14}$$

The constant $c > 1$ ensures that a straight transition occurs prior to other ones (requirement (i)). The second term of (14) for $\tau \neq \tau' + 1$ indicates that transitions to remote frames tend to occur when the activations are low (requirement (ii)), and $\gamma > 0$ weights a contribution of the term to $C_{\tau',\tau}$. We can obtain the alignment as an optimal path which minimizes the cumulative cost by the Viterbi algorithm [13].

The non-percussive components may be included in the target basis spectra due to imperfect source separation, and the synthesized sounds become thin if the components are lost. The alignment method loses the components, because the method does not use the target basis spectra for synthesizing percussive spectra, and we need to recover the non-percussive components. They tend to have low energy, and we replace a synthesized percussive spectrum with the corresponding target percussive spectrum when a summation over all the frequency bins of the percussive amplitude spectrum of the target piece 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) in the RWC popular music and music genre databases [14] as target and reference pieces, which were downsampled from 44.1 to 22.05 kHz. Then, we synthesized six pairs[1] of these musical audio signals. The signals of the target and reference pieces were converted into spectrograms by 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 replacer were set as $\sigma = 240$ 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 NMF of percussive spectrograms, we set the number of bases as four, and used the generalized I-divergence. The alignment method was compared with the equalizing method, and which drum sounds in the target piece were replaced with which drum sounds in the reference piece was chosen by one of the authors. The parameters for the drum timbre replacer were set as $(M, \alpha, \beta, \gamma, c, \epsilon) = (4, 0.5, 3, 10, 3, 100)$. A negative log posterior which was computed by the L2-regularized L1-loss support vector classifier (SVC) of LIBLINEAR [15] was used as P_τ, and the SVC was trained to distinguish percussive instruments from other ones, using the RWC instrument database [14].

We asked nine subjects how adequately they felt that (1) the drum timbres of the target piece were replaced with those of the reference piece and (2) the timbres of the target harmonic components were replaced with those of the reference piece. The subjects could listen to the target, 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 one to five. One point means that the timbres were not replaced and five point indicates that the timbres were replaced perfectly.

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

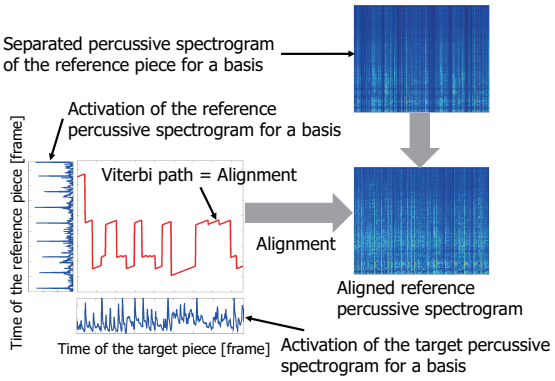

Fig. 5. Outline of the alignment method.

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 equalizing and the alignment methods. The result by the alignment algorithm was prior to that by the equalizing algorithm, in particular when drum timbres were largely different as we mentioned in Sec. 3. The average score of (2) with standard errors was 2.5 ± 0.1. The scores tended to increase when high-frequency elements were gained or reduced. The results show that subjects perceived replacements of drum timbres and frequency characteristics, and the system works well.

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

5. CONCLUSION

We have presented a system that can replace drum timbres and frequency characteristics of harmonic components between polyphonic audio signals, without using musical scores. We have proposed an algorithm that modifies a harmonic amplitude spectrum via bottom and top envelopes. We have also discussed two methods for replacing drum timbres. The equalizing method gains basis spectrograms by the proportions of the NMF bases of the target percussive spectrograms and those of the reference percussive spectrograms. The alignment method aligns basis spectra of the reference piece, referring to NMF activations of the target and reference pieces. Through an experiment, we confirmed that the system can let us perceive replacements of the drum timbres and the frequency characteristics.

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.