INTEGRATION AND ADAPTATION OF HARMONIC AND INHARMONIC MODELS FOR SEPARATING POLYPHONIC MUSICAL SIGNALS

Katsutoshi Itoyama,[†] *Masataka Goto,*[‡] *Kazunori Komatani,*[†] *Tetsuya Ogata,*[†] *Hiroshi G. Okuno*[†]

[†]Dept. of Intelligence Science and Technology
Graduate School of Informatics, Kyoto University
Sakyo-ku, Kyoto 606-8501 Japan
{itoyama, komatani, ogata, okuno} [at] kuis.kyoto-u.ac.jp

[‡]National Institute of Advanced Industrial
Science and Technology (AIST)
Tsukuba, Ibaraki 305-8568 Japan
m.goto [at] aist.go.jp

ABSTRACT

This paper describes a sound source separation method for polyphonic sound mixtures of music to build an instrument equalizer for remixing multiple tracks separated from compact-disc recordings by changing the volume level of each track. Although such mixtures usually include both harmonic and inharmonic sounds, the difficulties in dealing with both types of sounds together have not been addressed in most previous methods that have focused on either of the two types separately. We therefore developed an integrated weighted-mixture model consisting of both harmonic-structure and inharmonic-structure tone models (generative models for the power spectrogram). On the basis of the MAP estimation using the EM algorithm, we estimated all model parameters of this integrated model under several original constraints for preventing over-training and maintaining intra-instrument consistency. Using standard MIDI files as prior information of the model parameters, we applied this model to compact-disc recordings and achieved the instrument equalizer.

Index Terms— Music, separation, equalizers, sound source separation, music understanding.

1. INTRODUCTION

Our goal is to build an instrument equalizer, *INTER* [1], that enables a user to remix multiple audio tracks corresponding to different instruments in compact-disc (CD) recordings by changing the volume of each track. While most previous equalizers, such as graphic equalizers and tone controls for bass and treble, change the volume of each frequency band to adjust frequency characteristics, INTER can change the volume of each instrument to adjust the mixing balance of instruments in CD recordings. Because it was difficult to extract (i.e., demix) tracks from CD recordings, however, the proposed INTER was partially achieved only for drums: a drum-sound equalizer, called *INTER:D* [1], that changes the volume of drum sounds (inharmonic sounds). We therefore aim to built fully functional INTER, i.e., to build an advanced instrument equalizer for more general sound mixtures, including both harmonic and inharmonic sounds. Our equalizer, for example, enables a user to boost the volume of the guitar part, cut that of the bass part, and apply an additional sound effect to the saxophone melody part.

Our equalizer requires that the multiple audio tracks to be equalized can be separated from polyphonic sound mixtures, such as those recorded on CDs. Such mixtures consist of both harmonic sounds from pitched instruments, such as the piano and flute, and inharmonic sounds from unpitched instruments, such as drums. Even pitched instruments generate inharmonic sounds around their attacks (right after their onset times). For example, the piano sounds have inharmonic sounds when a hammer hits a piano string. A sound source separation method for our equalizer must therefore be able to deal with both types of sounds.

Most previous sound-source-separation methods, however, individually dealt with either of the two types and had difficulties dealing with both types together. For example, sound source separation methods for harmonic sounds have been reported [2, 3, 4, 5], and those for inharmonic sounds, such as drum sounds, have also been reported [6, 7]. Goto [8] mentioned a theoretical way of integrating inharmonic sound models with harmonic sound models, but has not evaluated it yet. Although there is another approach of using the blind sound source separation without any assumptions on sound sources [9], it still has difficulty dealing with sound mixtures consisting of many sounds and has not been successfully applied to CD recordings, including complex audio signals of musical pieces.

We therefore propose a sound source separation method using an integrated weighted-mixture model that consists of both harmonic-structure and inharmonic-structure tone models (generative models for the power spectrogram). These tone models are complementary because inharmonic-structure tone models can capture drum sounds and attacks of instrument sounds. The harmonic-structure tone model is based on a parametric model that represents the harmonic components of a pitched sound (e.g., a guitar sound). It is represented by model parameters such as the overall amplitude, F0 (fundamental frequency) trajectory, onset time, duration, timbre (relative amplitude of harmonic components), temporal power envelope for each pitched sound. The inharmonic-structure tone model, on the other hand, is based on a non-parametric model that directly represents the power spectrogram of an unpitched sound (e.g., a drum sound or the attack of a guitar sound) in the time-frequency domain.

Our method estimates all model parameters of this integrated model for each musical note in the audio signal of a musical piece on the basis of the MAP (Maximum *A Posteriori* Probability) estimation using the *Expectation-Maximization* algorithm. Because the inharmonic-structure tone model is too flexible, however, it can be fitted to any sounds, and the input sound mixture is sometimes only represented by the inharmonic-structure models. To solve this over-training problem, we estimated the model parameters by using their prior information and initial values that are given on the basis of multiple tracks in the standard MIDI file (SMF) corresponding to the target piece.[1] Moreover, we used several original constraints based on the Kullback-Leibler (KL) divergence. For example, if the inharmonic-structure tone model captures harmonic components of a sound, it is penalized. In addition, to maintain the intra-instrument consistency, we penalized the tone model of a sound if it is too different from that of other sounds from the same instrument. Because the power spectrograms of both harmonic and inharmonic sounds were thus obtained, the instrument equalizer was achieved by regenerating and remixing separated tracks.

2. PROBLEMS AND APPROACHES

Given the input audio signal of a musical piece and its standard MIDI file (SMF) is synchronized with the input signal (i.e., audio-synchronized transcription), the goal of our method is to separate the input signal into multiple audio tracks corresponding to the MIDI tracks of the SMF. Each MIDI track usually corresponds to a different musical instrument part. In other words, our method estimates all parameters of the harmonic-structure and inharmonic-structure

[1] We assume that the SMF has already been synchronized with the input mixture by using audio-to-score alignment methods [10, 11, 12].

tone models corresponding to all the notes in these separated tracks. By playing back each track of the SMF on a MIDI sound module, we prepared a sampled sound for each note. We call this a template sound and used this as prior information (and initial values) in the estimation. Even if the SMF and template sounds are available, we still have the following difficulties in separating complex sound mixtures.

1. The timbre of a template sound from the MIDI sound module is always different from that of the corresponding actual sound in the input signal because the different instrument is used.
2. Even if the same musical instrument is performed at the same F0 and duration, the produced sounds are different because of playing styles and expressions, such as vibrato and dynamics. These sounds still have consistency, though, compared to the sounds produced from a different instrument.

We address these issues as follows.

1. Starting from the tone model initialized with a template sound, the method tries to minimize the timbral difference between the tone model and the spectrogram of the corresponding actual sound through the estimation of the model parameters. This estimation can be considered the model adaptation. We use the integrated weighted-mixture model for all notes of both pitched and unpitched instruments.
2. The method tries to maintain the intra-instrument consistency while allowing for differences in the tone models for individual notes. This can be achieved by estimating the model parameters under constraint by using the KL divergence between the model parameters of each note and those averaged over all the notes of the same instrument (in the same MIDI track).

3. FORMULATION

The problem is to decompose the input power spectrogram, $g^{(O)}(c,f,t)$, into the power spectrogram corresponding to each musical note, where c is the channel (e.g., left and right), f is the frequency, and t is the time. Our method can theoretically deal with sound recordings of any number of sources/channels (even monaural). We assume there are K musical instruments in the $g^{(O)}$, and each instrument plays L_k musical notes. Here, the spectrogram of a template sound of l-th musical note performed by k-th musical instrument is denoted by $g^{(T)}_{k,l}(f,t)$, and the model (the corresponding spectrogram) estimated for its musical note is denoted by $h_{k,l}(c,f,t)$. The template spectrogram is monaural because the sound localization information in SMFs is not reliable.

For the decomposition of $g^{(O)}(c,f,t)$ with the model, $h_{k,l}(c,f,t)$, we introduce a distribution (decomposition) function of the spectrogram, $m^{(O)}(k,l;c,f,t)$, which satisfies

$$0 \le m^{(O)}(k,l;c,f,t) \le 1 \quad \text{and} \quad \sum_{k,l} m^{(O)}(k,l;c,f,t) = 1.$$

Thus, $m^{(O)}(k,l;c,f,t)g^{(O)}(c,f,t)$ represents the separated spectrogram of the l-th note by the k-th instrument. To evaluate the 'effectiveness' of this separation, we use the KL divergence between this $m^{(O)}g^{(O)}$ and the estimated-model spectrogram, $h_{k,l}$:

$$J_1(k,l) = \sum_c \iint m^{(O)}(k,l;c,f,t)g^{(O)}(c,f,t) \log \frac{m^{(O)}(k,l;c,f,t)g^{(O)}(c,f,t)}{h_{k,l}(c,f,t)}\, df\, dt,$$

To evaluate the 'effectiveness' of the estimated model $h_{k,l}(c,f,t)$, we use the KL divergence between $g^{(T)}_{k,l}$ and $h_{k,l}$:

$$J_2(k,l) = \sum_c \iint g^{(T)}_{k,l}(f,t) \log \frac{g^{(T)}_{k,l}(f,t)}{h_{k,l}(c,f,t)}\, df\, dt.$$

We use the following sum over k and l of the KL divergences as the cost function of the 'effectiveness' of both the separation and the models:

$$J_0 = \sum_{k,l} \left(\alpha J_1 + (1-\alpha) J_2\right),$$

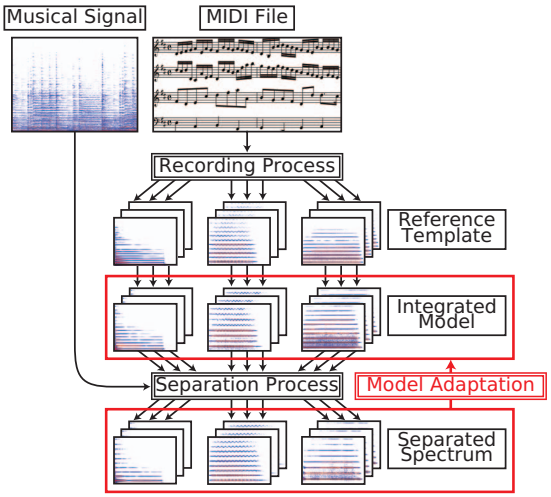

Fig. 1. Overview of separation and model adaptation.

where α $(0 \le \alpha \le 1)$ is a weight parameter. We can prevent the over-training of the models by gradually increasing α from 0 (i.e., the estimated model should first be close to the template spectrogram) through the iteration of the separation and adaptation (model estimation).

The overall process is depicted in Figure 1. We minimize J_0 by iteratively applying the separation process of estimating $m^{(O)}$ and the model adaptation of estimating $h_{k,l}$ one after the other; the other variables are fixed during the iterations. Here, J_0 is written as

$$\begin{aligned} J_0 = &\sum_{c,k,l} \iint m^{(O)}(k,l;c,f,t)g^{(O)}(c,f,t) \log \frac{m^{(O)}(k,l;c,f,t)g^{(O)}(c,f,t)}{h_{k,l}(c,f,t)}\, df\, dt \\ &+ \sum_{c,k,l} \iint g^{(T)}_{k,l}(f,t) \log \frac{g^{(T)}_{k,l}(f,t)}{h_{k,l}(c,f,t)}\, df\, dt \\ &- \lambda(c,f,t) \sum_{k,l} \left(\sum_c \iint m^{(O)}(k,l;c,f,t)\, df\, dt - 1 \right), \end{aligned}$$

where λ is a Lagrange undetermined multiplier. First, we minimize J_0 by optimizing the $m^{(O)}$ with the $h_{k,l}$ fixed. The partial derivatives of J_0 are

$$\frac{\partial J_0}{\partial m^{(O)}} = g^{(O)}(c,f,t) \log \frac{m^{(O)}(k,l;c,f,t)g^{(O)}(c,f,t)}{h_{k,l}(c,f,t)} - \lambda$$

$$\text{and} \quad \frac{\partial J_0}{\partial \lambda} = \sum_{k,l} m^{(O)}(k,l;c,f,t) - 1.$$

By solving the simultaneous equations:

$$\frac{\partial J_0}{\partial m^{(O)}} = 0 \quad \text{and} \quad \frac{\partial J_0}{\partial \lambda} = 0,$$

we get the optimal $m^{(O)}(k,l;c,f,t)$:

$$m^{(O)}(k,l;c,f,t) = \frac{h_{k,l}(c,f,t)}{\sum_{k,l} h_{k,l}(c,f,t)}.$$

Second, we minimize J_0 by optimizing the $h_{k,l}$ with the $m^{(O)}$ fixed. Because this optimization depends on the model definition of $h_{k,l}$, it is described in Sections 4 and 5.

Note that the above optimization of $m^{(O)}$ and $h_{k,l}$ is equivalent to the MAP estimation using the EM algorithm. This fact becomes

Table 1. Parameters of harmonic-structure tone model.

Symbol	Description
$w_{k,l}$	overall amplitude
$\mu_{k,l}(t)$	F0 trajectory
$v_{k,l,n}$	relative amplitude of n-th harmonic component
$u_{k,l,y}$	coefficient of the temporal power envelope
$\tau_{k,l}$	onset time
$Y\phi_{k,l}$	duration (note that Y is constant)
$\sigma_{k,l}$	diffusion of a harmonic component along the frequency axis

clear if we introduce a Q function given by

$$Q(\theta,\tilde{\theta}) = \sum_{c,k,l}\iint \overbrace{p(k,l|c,f,t,\theta)}^{\text{missing data pdf}}\overbrace{g^{(O)}(c,f,t)}^{\text{observed pdf}}\log\underbrace{p(k,l,c,f,t|\tilde{\theta})}_{\text{complete data pdf}}\,df\,dt + \sum_{c,k,l}\iint \underbrace{p(k,l,f,t)}_{\text{prior pdf}}\log\overbrace{p(k,l,c,f,t|\tilde{\theta})}^{\text{complete data pdf}}\,df\,dt,$$

$$p(k,l|c,f,t,\theta) = \frac{p(k,l,c,f,t|\theta)}{\sum_{k,l}p(k,l,c,f,t|\theta)}.$$

From this Q function, we can obtain a new cost function, J:

$$J = \sum_{c,k,l}\iint G_{k,l}(c,f,t)\log\frac{G_{k,l}(c,f,t)}{h_{k,l}(c,f,t)}\,df\,dt,$$

$$\text{where}\quad G_{k,l}(c,f,t) = \alpha m^{(O)}(k,l;c,f,t)g^{(O)}(c,f,t) + (1-\alpha)g_{k,l}^{(T)}(f,t).$$

J and J_0 are equivalent when they are minimized by optimizing $m^{(O)}$ and $h_{k,l}$.

4. INTEGRATED MODEL

The model, $h_{k,l}(c,f,t)$, is the integrated weighted-mixture model that consists of both harmonic-structure and inharmonic-structure tone models and is given by

$$h_{k,l}(c,f,t) = r_{k,l,c}\left(H_{k,l}(f,t) + I_{k,l}(f,t)\right),$$

where $H_{k,l}(f,t)$ denotes the harmonic-structure tone model for harmonic components, $I_{k,l}(f,t)$ denotes the inharmonic-structure tone model for inharmonic components, and $r_{k,l,c}$ is their relative amplitude in each channel.

4.1. Harmonic-structure tone model

The harmonic-structure tone model is given by

$$H_{k,l}(f,t) = \sum_{n=1}^{N}\sum_{y=0}^{Y-1} w_{k,l}F_{k,l,n}(f,t)E_{k,l,y}(t),$$

$$F_{k,l,n}(f,t) = \frac{v_{k,l,n}}{\sqrt{2\pi}\sigma_{k,l}}\exp\left(-\frac{(f-n\mu_{k,l}(t))^2}{2\sigma_{k,l}^2}\right),$$

$$\text{and}\quad E_{k,l,y}(f,t) = \frac{u_{k,l,y}}{\sqrt{2\pi}\phi_{k,l}}\exp\left(-\frac{(t-\tau_{k,l}-y\phi_{k,l})^2}{2\phi_{k,l}^2}\right),$$

where the parameters of this model are listed in Table 1. This model was designed by referring to the harmonic-temporal-structured clustering (HTC) source model [4].

In the original HTC model, $\mu_{k,l}(t)$ was defined as a polynomial function, but here, we define $\mu_{k,l}(t)$ as a more flexible nonparametric function. Because this function is not constrained, it might cause temporal discontinuities in the estimated F0 trajectory $\mu_{k,l}(t)$. To prevent these discontinuities, we introduce an original constraint given by

$$\beta_\mu\int\bar{\mu}_{k,l}(t)\log\frac{\bar{\mu}_{k,l}(t)}{\mu_{k,l}(t)}\,dt \quad\text{with}\quad \int\bar{\mu}_{k,l}(t)\,dt = \int\mu_{k,l}(t)\,dt.$$

This constraint makes $\mu_{k,l}(t)$ close to $\bar{\mu}_{k,l}(t)$, which is obtained by smoothing $\mu_{k,l}(t)$ with a Gaussian filter along the time axis.

4.2. Inharmonic-structure tone model

The inharmonic-structure tone model is represented as a nonparametric power spectrogram. As we pointed out in Section 1, input sound mixture is sometimes represented only by this very flexible inharmonic-structure models without any harmonic-structure models. To solve this problem, which is known as the problem of overtraining or overfitting due to too many parameters, we introduce an original constraint given by

$$\beta_{I2}\iint\left(\bar{I}_{k,l}(f,t)\log\frac{\bar{I}_{k,l}(f,t)}{I_{k,l}(f,t)} - \bar{I}_{k,l}(f,t) + I_{k,l}(f,t)\right)df\,dt.$$

This constraint has the effect of making $I_{k,l}(f,t)$ close to $\bar{I}_{k,l}(f,t)$, which is obtained by smoothing $I_{k,l}(f,t)$ with a Gaussian filter along the frequency axis.

4.3. Constraint of the intra-instrument consistency

The models $h_{k,l}(c,f,t)$ estimated for musical notes of the same instrument should have similar but different parameter values as discussed in Section 2. To maintain this intra-instrument consistency we introduce two constraints. The first constraint is for the harmonic-structure tone model:

$$\beta_v\sum_n\left(\bar{v}_{k,n}\log\frac{\bar{v}_{k,n}}{v_{k,l,n}} - \bar{v}_{k,n} + v_{k,l,n}\right),$$

where $\bar{v}_{k,n}$ is an intra-instrument average of $v_{k,l,n}$, which means that the relative amplitude of each harmonic component should be similar for a particular instrument. The second constraint is for the inharmonic-structure tone model:

$$\beta_{I1}\iint\left(\bar{I}_k(f,t)\log\frac{\bar{I}_k(f,t)}{I_{k,l}(f,t)} - \bar{I}_k(f,t) + I_{k,l}(f,t)\right)df\,dt,$$

where $\bar{I}_k(f,t)$ is an intra-instrument average of $I_{k,l}(f,t)$, which means that the power spectrogram itself should be similar for a particular instrument.

5. MODEL ADAPTATION

As described in Section 3, we can now minimize J_0, i.e., J, by optimizing the $h_{k,l}$ with the $m^{(O)}$ fixed. We assume that J is decomposed into $J_{k,l}$, given as $J = \sum_{k,l}J_{k,l}$. Given that λ_r, λ_v, and λ_u are Lagrange multipliers for $r_{k,l,c}$, $v_{k,l,n}$, and $u_{k,l,y}$, the update equations for each variable of the model, $h_{k,l}$, are derived by minimizing J, which consists of the following sub-cost function:

$$\begin{aligned}
J_{k,l} = &\sum_{c,n,y}\iint\left(G_{k,l,n,y}^{(H)}(c,f,t)\log\frac{G_{k,l,n,y}^{(H)}(c,f,t)}{r_{k,l,c}w_{k,l}F_{k,l,n}(f,t)E_{k,l,y}(t)} - G_{k,l,n,y}^{(H)}(c,f,t) + r_{k,l,c}w_{k,l}F_{k,l,n}(f,t)E_{k,l,y}(t)\right)df\,dt\\
&+\sum_{c}\iint\left(G_{k,l}^{(I)}(c,f,t)\log\frac{G_{k,l}^{(I)}(c,f,t)}{r_{k,l,c}I_{k,l}(f,t)} - G_{k,l}^{(I)}(c,f,t) + r_{k,l,c}I_{k,l}(f,t)\right)df\,dt\\
&+\lambda_r\left(\sum_c r_{k,l,c}-1\right)+\lambda_v\left(\sum_n v_{k,l,n}-1\right)+\lambda_u\left(\sum_y u_{k,l,y}-1\right)\\
&+\beta_v\sum_n\left(\bar{v}_{k,n}\log\frac{\bar{v}_{k,n}}{v_{k,l,n}} - \bar{v}_{k,n} + v_{k,l,n}\right)\\
&+\beta_\mu\int\left(\bar{\mu}_{k,l}(t)\log\frac{\bar{\mu}_{k,l}(t)}{\mu_{k,l}(t)} - \bar{\mu}_{k,l}(t) + \mu_{k,l}(t)\right)dt\\
&+\beta_{I1}\iint\left(\bar{I}_k(f,t)\log\frac{\bar{I}_k(f,t)}{I_{k,l}(f,t)} - \bar{I}_k(f,t) + I_{k,l}(f,t)\right)df\,dt\\
&+\beta_{I2}\iint\left(\bar{I}_{k,l}(f,t)\log\frac{\bar{I}_{k,l}(f,t)}{I_{k,l}(f,t)} - \bar{I}_{k,l}(f,t) + I_{k,l}(f,t)\right)df\,dt.
\end{aligned}$$

Table 2. Experimental conditions.

Frequency analysis	
sampling rate	44.1 kHz
STFT window	2048 points with a Gaussian window
Parameters	
# of partials: N	20
# of kernels in $E_{k,l,y}(t)$: Y	10
β_v	0.1
β_μ	0.01
ς_μ	0.1
β_{I1}	0.1
β_{I2}	0.1
ς_I	100
MIDI sound modules	
test data	YAMAHA MU2000
template sounds	Roland SD-90

Note that

$$\int \left(p(x) \log \frac{p(x)}{q(x;\theta)} - p(x) + q(x;\theta) \right) dx$$

and

$$\int p(x) \log \frac{p(x)}{q(x;\theta)}\, dx, \quad \text{with} \quad \int p(x)\, dx = \int q(x;\theta)\, dx$$

are equivalent when they are minimized about θ. The update equations are omitted because it would take a page to list them all.

6. EXPERIMENTAL RESULTS

As a preliminary experiment to evaluate our new approach and determine the effectiveness of the model adaptation, we tested our method on a small database consisting of five musical pieces so that we could measure the signal-to-noise ratio (SNR) between the original sounds and the separated sounds.

6.1. Experimental conditions

Although our method is intended to deal with CD recordings, this experiment uses audio signals recorded from a MIDI sound module for the following reasons.

1. Most music databases does not contain both SMFs synchronized with the musical pieces and master tracks, which are necessary for quantitative evaluation based on SNR, of the pieces with the separated signals.
2. Our separation method cannot deal with singing voices because representing singing voices by our models and generating templates of singing voices are difficult.

The template sounds were recorded from a different MIDI sound module made by another manufacturer. We used five SMFs from the RWC Music Database [13] (RWC-MDB-P-2001 Nos. 1, 2, 5, 8, and 10). Because of the computational time and memory limitation, the length of each musical signal was limited to 5 seconds. The details of the experimental conditions are listed in Table 2.

6.2. Experimental results

The results are listed in Table 3. Using the model adaptation, we increased the SNR from 5.53 to 5.73 dB. We think that separated audio signals with 8 dB SNRs are sufficient for some applications, such as our instrument equalizer. Musical piece No. 7 had the lowest SNR because this piece included some tracks which had very few musical notes. It was therefore difficult to model such notes because the constraints of the same musical instrument were inadequate. We think it is possible to improve the SNR if we use a longer excerpt of this particular musical piece.

Musical piece No. 1 had the highest SNR, but the SNR decreased after the model adaptation. Because this piece included many piano and acoustic guitar sounds, we think this fall-off was caused by the over-training of the models to those sounds. We have to introduce a new constraint to prevent this over-training problem.

Audio demonstrations, including the original sound mixtures and the sounds separated using our method, are available at the following URL. http://winnie.kuis.kyoto-u.ac.jp/˜itoyama/icassp2007/.

Table 3. Experimental results.

	SNR [dB]		
	w/o adapt.	1st adapt.	3rd adapt.
No. 1	9.25	9.15	9.11
No. 2	6.65	6.89	6.97
No. 5	5.40	5.45	5.47
No. 7	3.64	3.91	3.99
No. 10	5.74	6.21	6.30
Average	5.53	5.70	5.73

7. CONCLUSION

We have described a sound source separation method based on an integrated weighted-mixture model that represents both harmonic and inharmonic sounds. We implemented this method and tested it on a small databse to determine the effectiveness of our model adaption by using the EM algorithm. On the basis of this method, we have already implemented the instrument equalizer that enables a user to remix multiple separated tracks by changing the volume level of each track. Although this integrated model is flexible and powerful, we have not evaluated it on music signals, including vocals and CD recordings. We plan to apply our method to various sound sources and validate its effectiveness.

ACKNOWLEDGMENTS: This research was partially supported by the Ministry of Education, Culture, Sports, Science and Technology (MEXT), Grant-in-Aid for Scientific Research and Informatics Research Center for Development of Knowledge Society Infrastructure (COE program of MEXT, Japan).

8. REFERENCES

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

[2] T. Virtanen and A. Klapuri, "Separation of harmonic sounds using linear models for the overtone series," in *Proc. ICASSP*, 2002, vol. 2, pp. 1757–1760.

[3] M. Every and J. Szymanski, "A spectral-filtering approach to music signal separation," in *Proc. DAFx*, 2004, pp. 197–200.

[4] H. Kameoka, T. Nishimoto, and S. Sagayama, "Harmonic-temporal structured clustering via deterministic annealing EM algorithm for audio feature extraction," in *Proc. ISMIR*, 2005, pp. 115–122.

[5] H. Viste and G. Evangelista, "A method for separation of overlapping partials based on similarity of temporal envelopes in multichannel mixtures," *IEEE Transactions on Speech and Audio Processing*, vol. 14, no. 3, pp. 1051–1061, May 2006.

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

[7] D. Barry, D. Fitzgerald, E. Coyle, and B. Lawlor, "Drum source separation using percussive feature detection and spectral modulation," in *Proc. ISSC*, 2005, pp. 13–17.

[8] M. Goto, "A real-time music-scene-description system: Predominant-F0 Estimation for detecting melody and bass lines in real-world audio signals," *Speech Communication (ISCA Journal)*, vol. 43, no. 4, pp. 311–329, September 2004.

[9] M. Casey and A. Westner, "Separation of mixed audio sources by independent subspace analysis," in *Proc. ICMC*, 2000, pp. 154–161.

[10] P. Cano, A. Loscos, and J. Bonada, "Score-performance matching using HMMs," in *Proc. ICMC*, 1999, pp. 441–444.

[11] N. Adams, D. Marquez, and G. Wakefield, "Iterative deepening for melody alignment and retrieval," in *Proc. ISMIR*, 2005, pp. 199–206.

[12] Arshia Cont, "Realtime audio to score alignment for polyphonic music instruments using sparce non-negative constraints and hierarchical hmms," in *Proc. ICASSP*, 2006, vol. II, pp. 641–644.

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