

Regular Paper

Parameter Estimation for Harmonic and Inharmonic Models by Using Timbre Feature Distributions

KATSUTOSHI ITOYAMA,^{†1} MASATAKA GOTO,^{‡2}
 KAZUNORI KOMATANI,^{†1} TETSUYA OGATA^{†1}
 and HIROSHI G. OKUNO^{†1}

We describe an improved way of estimating parameters for an integrated weighted-mixture model consisting of both harmonic and inharmonic tone models. Our final goal is to build an *instrument equalizer* (music remixer) that enables a user to change the volume of parts of polyphonic sound mixtures. To realize the instrument equalizer, musical signals must be separated into each musical instrument part. We have developed a score-informed sound source separation method using the integrated model. A remaining but critical problem is to find a way to deal with timbre varieties caused by various performance styles and instrument bodies because our method used template sounds to represent their timbre. Template sounds are generated from a MIDI tone generator based on an aligned score. Difference of instrument bodies between mixed signals and template sounds causes timbre difference and decreases separation performance. To solve this problem, we train probabilistic distributions of timbre features using various sounds to reduce template dependency. By adding a new constraint of maximizing the likelihood of timbre features extracted from each tone model, we can estimate model parameters that express the timbre more accurately. Experimental results show that separation performance improved from 4.89 to 8.48 dB.

1. Introduction

Our goal is to build an instrument equalizer (remixer) that enables a user to change the volume of parts of polyphonic sound mixtures, such as stereo compact-disc recordings that consist of several instruments. Given separate audio tracks corresponding to different instrument parts, it is easy to build such an equalizer, but separating a mixture of these audio tracks into different tracks is difficult.

Although a number of sound source separation methods^{1)–5)} and automatic transcription methods^{6)–9)} have been studied, most of them still have difficulty dealing with music performed on both pitched instruments that have harmonic sounds and drums that have inharmonic sounds. For example, most separation methods for harmonic sounds^{2)–5)} cannot separate inharmonic sounds, while most separation methods for inharmonic sounds, such as drums¹⁰⁾, cannot separate harmonic ones. Although there are separation methods^{1),7)} that deal with both harmonic and inharmonic sounds in theory, they have practical difficulties in separating complex polyphonic sound mixtures like those in popular music because they need temporal and frequency sparseness of the input signals.

We therefore propose a sound source separation method¹¹⁾ that represents the input polyphonic audio signal as a mixture of both harmonic and inharmonic tone models that correspond to musical notes. We assume that a standard MIDI file (SMF) synchronized with the input audio signal is available as prior information^{*1}. By using an iterative algorithm for estimating the parameters of the harmonic and inharmonic models, the model parameters are initialized using template sounds recorded from a MIDI sound generator and are gradually improved so that they represent sounds in the input mixture. Once the model parameters are estimated, we can easily obtain the separated power spectrogram of each musical note. We found that the integration of harmonic and inharmonic models increased separation performance. However, this method did not model timbre varieties within each instrument since the template sounds are generated from a MIDI sound generator with single performance style. The instrument body and the performance style are usually different from that of the input sound mixtures. This difference decreases the separation performance.

We propose a new method that estimates the model parameters of the integrated model consisting of both harmonic and inharmonic models while considering timbre varieties caused by different instrument bodies and various performance styles. First, the timbre varieties of each instrument are modeled as a timbre feature distribution by using various training sound samples of that in-

^{†1} Kyoto University

^{‡2} National Institute of Advanced Industrial Science and Technology (AIST)

*1 We can assume several musical service providers, such as online music distributors, have a huge database which contains both, and our method would be useful with a business style which the providers do separation and clients use separated signals.

strument. Then the model parameters, i.e., separate sounds, are estimated so that the timbre features extracted from each separated sound have the maximum likelihood with its timbre feature distribution. Our method can thus be used to improve source separation performance using the varieties of timbre.

2. Sound Source Separation Using Integrated Models

In this section, we define our sound source separation problem and the integrated model.

The sound source separation problem is to decompose the input power spectrogram, $X(c, t, f)$, into the power spectrogram corresponding to each musical note, where c , t , and f are the channel (e.g., left and right), the time, and the frequency, respectively. We assume that $X(c, t, f)$ includes K musical instruments and the k -th instrument plays L_k musical notes. We use the tone model, $J(k, l, c, t, f)$, to represent the power spectrogram of the l -th musical note from the k -th musical instrument ((k, l) -th note), and the power spectrogram of a template sound, $Y(k, l, t, f)$, to initialize the parameters of $J(k, l, c, t, f)$. Each musical note of the SMF is played back on a MIDI sound generator in advance to record the corresponding template sound. $Y(k, l, t, f)$ is monaural because SMFs may not include accurate sound localization (channel) information. $Y(k, l, t, f)$ is normalized to satisfy the following relation:

$$\sum_c \iint X(c, t, f) dt df = C \sum_{k,l} \iint Y(k, l, t, f) dt df, \quad (1)$$

where C is the total number of channels.

We approximate the power spectrogram is additive. This approximation is valid when the sounds are harmonic and sparse. Note that the validity decreases if many instruments play simultaneously.

For this source separation, we define this integrated model, $J(k, l, c, t, f)$, as the sum of the harmonic-structure tone models, $H(k, l, t, f)$, and inharmonic-structure tone models, $I(k, l, t, f)$, multiplied by the whole amplitude of the model, $w_J(k, l)$, and the relative amplitude of each channel, $r(k, l, c)$:

$$J(k, l, c, t, f) = w_J(k, l) r(k, l, c) (H(k, l, t, f) + I(k, l, t, f)), \quad (2)$$

where $w_J(k, l)$ and $r(k, l, c)$ satisfy the following constraints:

Table 1 Parameters of integrated model.

Symbol	Description
$w_J(k, l)$	overall amplitude
$r(k, l, c)$	relative amplitude of each channel
$w_H(k, l), w_I(k, l)$	relative amplitude of harmonic and inharmonic tone models
$v_H(k, l, m, n)$	relative amplitude of n -th harmonic at time $m\phi_H(k, l)$
$\tau(k, l)$	onset time
$\phi_H(k, l)$	diffusion of a Gaussian distribution constructing power envelope of the harmonic tone model
$\omega_H(k, l, t)$	F0 trajectory
$\sigma_H(k, l)$	diffusion of a harmonic component along the frequency axis
$v_I(k, l, m, n)$	relative amplitude of n -th inharmonic frequency component at time $m\phi_I$
ϕ_I	diffusion of a Gaussian distribution constructing power envelope of the inharmonic tone model
$\omega_I(n)$	central frequency of the n -th inharmonic frequency component
$\sigma_I(n, f)$	diffusion of an inharmonic frequency component along the frequency axis

$$\sum_{k,l} w_J(k, l) = \frac{1}{C} \iint X(c, t, f) dt df \quad \text{and} \quad (3)$$

$$\forall k, l: \sum_c r(k, l, c) = C. \quad (4)$$

Our aim is to decompose the power spectrogram of each musical instrument sound into the harmonic and non-harmonic components, like a sinusoidal modelling decomposes an input signal into a sum of sinusoidals and residual parts.

All parameters of $J(k, l, c, t, f)$ are listed in **Table 1**. The harmonic model, $H(k, l, t, f)$, is defined as a constrained two-dimensional Gaussian mixture model and is designed by referring to the harmonic-temporal-structured clustering (HTC) source model¹²⁾ (see **Figs. 1** and **2**). The inharmonic model, $I(k, l, t, f)$, has a similar structure to the harmonic model. The inharmonic tone model has the same structure along the time axis as the harmonic tone model. Along the frequency axis, the inharmonic tone model has a structure in which the Gaussian kernels are located at equal intervals on the logarithmic frequency (see **Fig. 3**), to prevent the inharmonic model depriving from the harmonic model of the harmonic component when these models have similar shapes. The definitions of these models are as follows:

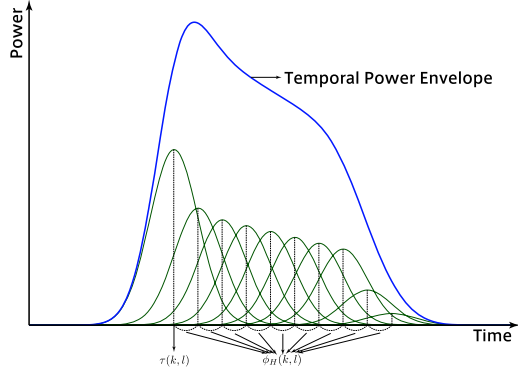


Fig. 1 Temporal power envelope of harmonic tone model.

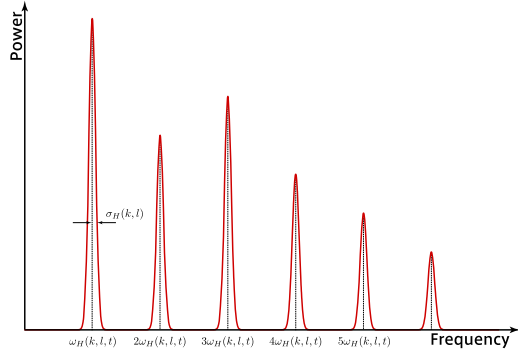


Fig. 2 Harmonic structure of harmonic tone model.

$$H(k, l, t, f) = w_H(k, l) \sum_{m=0}^{M_H} \sum_{n=1}^{N_H} H(k, l, m, n, t, f), \quad (5)$$

$$H(k, l, m, n, t, f) = \frac{v_H(k, l, m, n)}{2\pi\phi_H(k, l)\sigma_H(k, l)} \cdot \exp\left(-\frac{(t - (\tau(k, l) - m\phi_H(k, l)))^2}{2\phi_H(k, l)^2}\right) \exp\left(-\frac{(f - n\omega_H(k, l, t))^2}{2\sigma_H(k, l)^2}\right), \quad (6)$$

$$I(k, l, t, f) = w_I(k, l) \sum_{m=0}^{M_I} \sum_{n=1}^{N_I} I(k, l, m, n, t, f), \quad (7)$$

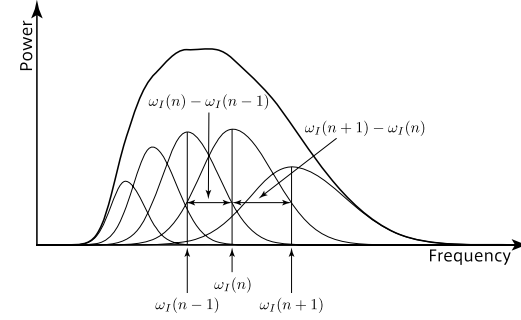


Fig. 3 Frequency structure of inharmonic tone model.

$$I(k, l, m, n, t, f) = \frac{v_I(k, l, m, n)}{2\pi\phi_I(\omega_I(n+1) - \omega_I(n-1))} \cdot \exp\left(-\frac{(t - (\tau(k, l) - m\phi_I))^2}{2\phi_I^2}\right) \exp\left(-\frac{(f - \omega_I(n))^2}{2\sigma_I(n, f)^2}\right), \quad (8)$$

where

$$\omega_I(n) = \omega_{Ia}((\omega_{Ib})^n - 1), \quad (9)$$

$$\sigma_I(n, f) = \begin{cases} \omega_I(n) - \omega_I(n-1) & (f \leq \omega_I(n)) \\ \omega_I(n+1) - \omega_I(n) & (f > \omega_I(n)) \end{cases}, \quad (10)$$

M_H and N_H are the number of Gaussian kernels representing the temporal power envelope and the harmonic components of the harmonic tone model, respectively, and M_I and N_I are the number of Gaussian kernels of the inharmonic tone model representing the same as above. $v_H(k, l, m, n)$, $v_I(k, l, m, n)$, $w_H(k, l)$, and $w_I(k, l)$ satisfy the following conditions:

$$\forall k, l: \sum_{m=0}^{M_H} \sum_{n=1}^{N_H} v_H(k, l, m, n) = 1, \quad (11)$$

$$\forall k, l: \sum_{m=0}^{M_I} \sum_{n=1}^{N_I} v_I(k, l, m, n) = 1, \quad \text{and} \quad (12)$$

$$\forall k, l: w_H(k, l) + w_I(k, l) = 1. \quad (13)$$

The goal of this separation is to decompose $X(c, t, f)$ into $J(k, l, c, t, f)$ by estimating a spectrogram distribution function, $\Delta_J(k, l, c, t, f)$, which satisfies

$$\forall k, l, c, t, f : 0 \leq \Delta_J(k, l, c, t, f) \leq 1, \quad \text{and} \quad (14)$$

$$\forall c, t, f : \sum_{k, l} \Delta_J(k, l, c, t, f) = 1. \quad (15)$$

With $\Delta_J(k, l, c, t, f)$, the separated power spectrogram, $X_J(k, l, c, t, f)$, is obtained as

$$X_J(k, l, c, t, f) = \Delta_J(k, l, c, t, f)X(c, t, f). \quad (16)$$

Furthermore, let $\Delta_H(k, l, m, n, t, f)$ and $\Delta_I(k, l, m, n, t, f)$ be spectrogram distribution functions which decompose $X_J(k, l, c, t, f)$ into each Gaussian distribution of the harmonic and inharmonic models, respectively. These functions satisfy

$$\forall k, l, m, n, t, f : 0 \leq \Delta_H(k, l, m, n, t, f) \leq 1, \quad (17)$$

$$\forall k, l, m, n, t, f : 0 \leq \Delta_I(k, l, m, n, t, f) \leq 1, \quad \text{and} \quad (18)$$

$$\forall k, l, t, f : \sum_{m, n} \Delta_H(k, l, m, n, t, f) + \sum_{m, n} \Delta_I(k, l, m, n, t, f) = 1. \quad (19)$$

To evaluate the ‘effectiveness’ of this separation, we can use a cost function defined as the Kullback-Leibler (KL) divergence from $X_J(k, l, c, t, f)$ to $J(k, l, c, t, f)$:

$$\sum_c \iint X_J(k, l, c, t, f) \log \frac{X_J(k, l, c, t, f)}{J(k, l, c, t, f)} dt df. \quad (20)$$

By minimizing the sum of the divergences over (k, l) pertaining to $\Delta_J(k, l, c, t, f)$, we obtain the spectrogram distribution function and model parameters (i.e., the most ‘effective’ decomposition).

By minimizing the divergence pertaining to each parameter of the integrated model, we obtain model parameters estimated from the distributed spectrogram. This parameter estimation is equivalent to a maximum likelihood estimation.

3. Timbre Varieties Representation Using Prior Distribution

In this section, we describe timbre varieties and timbre feature distributions for estimating parameters of the model.

3.1 Timbre Varieties within Each Instrument

Even within the same instrument, different instrument bodies have different timbres, although its timbral difference is smaller than the difference among different musical instruments. Moreover, in live performances, each musical note

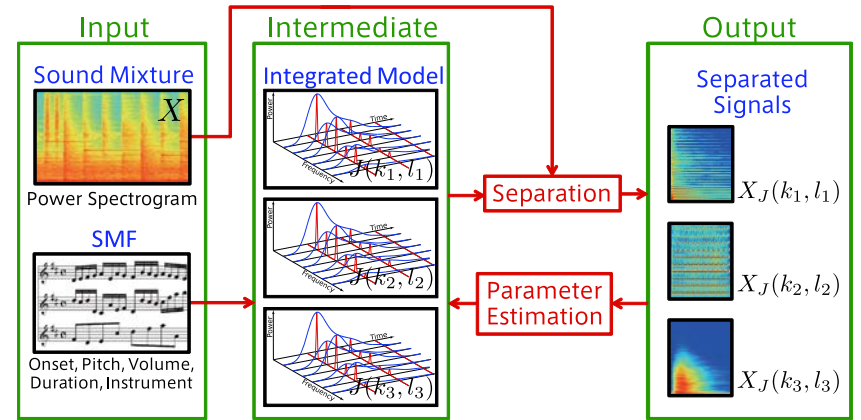


Fig. 4 Overview of iterating the separation and parameter estimation.

could have slightly different timbre according to the performance styles. Instead of preparing a set of many template sounds to represent such timbre varieties within each instrument, we represent them by using a probabilistic distribution.

We use parameters of the integrated model, $(w_H(k, l), w_I(k, l))$, $v_H(k, l, m, n)$, $v_I(k, l, m, n)$, to represent the timbre variety of instrument k by training Dirichlet distributions, which are known as the conjugate priors of these weight parameters. We defined three distributions for each instrument:

- (1) $p(w_H(k, l), w_I(k, l))$,
- (2) $p(v_H(k, l, 0, 1), \dots, v_H(k, l, M_H - 1, N_H))$ and
- (3) $p(v_I(k, l, 0, 1), \dots, v_I(k, l, M_I - 1, N_I))$.

The model parameters for training the prior distributions were extracted from the ‘‘RWC Music Database: Musical Instrument Sound’’¹³⁾ (i.e., the parameters are estimated without any prior distributions). The probability distribution functions of these prior distributions are described as follows:

$$p(w_H(k, l), w_I(k, l)) \propto w_H(k, l)^{\alpha_{w_H(k)}} w_I(k, l)^{\alpha_{w_I(k)} - 1}, \quad (21)$$

$$p(v_H(k, l, 0, 1), \dots, v_H(k, l, M_H - 1, N_H)) \propto \prod_{m, n} v_H(k, l, m, n)^{\alpha_{v_H(k, m, n)} - 1} \quad (22)$$

$$p(v_I(k, l, 0, 1), \dots, v_I(k, l, M_I - 1, N_I)) \propto \prod_{m,n} v_I(k, l, m, n)^{\alpha_{v_I}(k, m, n) - 1}, \quad (23)$$

where $\{\alpha_{w_H}(k), \alpha_{w_I}(k)\}$, $\{\alpha_{v_H}(k, m, n)\}$ and $\{\alpha_{v_I}(k, m, n)\}$ are the parameters of the prior distributions. We assume that the values of these parameters are more than 1.

Let $X_H(k, l, m, n, c, t, f)$ and $X_I(k, l, m, n, c, t, f)$ be the decomposed power:

$$X_H(k, l, m, n, c, t, f) = \Delta_H(k, l, m, n, t, f)X_J(k, l, c, t, f) \quad \text{and} \quad (24)$$

$$X_I(k, l, m, n, c, t, f) = \Delta_I(k, l, m, n, t, f)X_J(k, l, c, t, f). \quad (25)$$

By minimizing the cost function,

$$\begin{aligned} Q = & \sum_{c,m,n} \iint X_H(k, l, m, n, c, t, f) \\ & \cdot \log \frac{X_H(k, l, m, n, c, t, f)}{w_J(k, l)r(k, l, c)w_H(k, l)H(k, l, m, n, t, f)} dt df \\ & + \sum_{c,m,n} \iint X_I(k, l, m, n, c, t, f) \\ & \cdot \log \frac{X_I(k, l, m, n, c, t, f)}{w_J(k, l)r(k, l, c)w_I(k, l)I(k, l, m, n, t, f)} dt df \\ & - (\alpha_{w_H}(k) - 1) \log w_H(k, l) - (\alpha_{w_I}(k) - 1) \log w_I(k, l) \\ & - \sum_{m,n} (\alpha_{v_H}(k, m, n) - 1) \log v_H(k, l, m, n) \\ & - \sum_{m,n} (\alpha_{v_I}(k, m, n) - 1) \log v_I(k, l, m, n), \quad (26) \end{aligned}$$

where the latter three terms are additional costs by using the prior distribution, we obtain the parameters by taking into account the timbre varieties as shown in **Fig. 5**. This parameter estimation is equivalent to a maximum *A Posteriori* estimation. The parameter update equations are listed in the Appendix.

3.2 Previous Cost Function without Considering Timbre Feature Distributions

For comparison with our previous study¹¹⁾, we also tested the previous cost function¹¹⁾ in which we used template sounds instead of timbre feature distributions to evaluate the ‘goodness’ of the feature vector. Let $Y_H(k, l, m, n, t, f)$ and

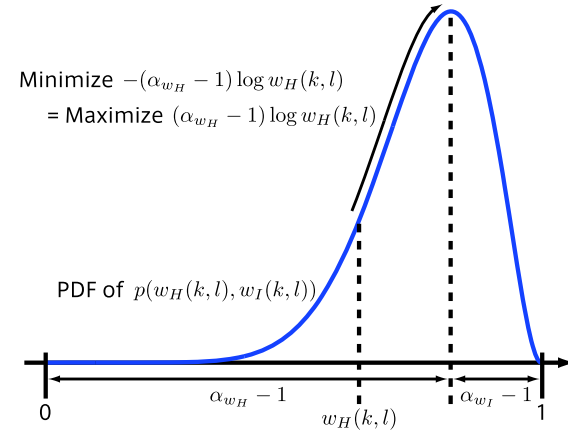


Fig. 5 Minimizing the additional costs.

$Y_I(k, l, m, n, t, f)$ be the decomposed template power:

$$Y_H(k, l, m, n, t, f) = \Delta_H(k, l, m, n, t, f)Y(k, l, t, f) \quad \text{and} \quad (27)$$

$$Y_I(k, l, m, n, t, f) = \Delta_I(k, l, m, n, t, f)Y(k, l, t, f). \quad (28)$$

The cost function, used in the previous study, can be obtained by replacing the negative log-likelihood (the terms about $\log w_H(k, l)$, $\log w_I(k, l)$, $\log v_H(k, l, m, n)$, and $\log v_I(k, l, m, n)$ in Eq. (26)) with the KL divergence from the power spectrogram of a template sound which is weighted by the relative amplitude of each channel, $r(k, l, c)Y(k, l, t, f)$, to $J(k, l, c, t, f)$:

$$\begin{aligned} Q = & \sum_{c,m,n} \iint X_H(k, l, m, n, c, t, f) \\ & \cdot \log \frac{X_H(k, l, m, n, c, t, f)}{w_J(k, l)r(k, l, c)w_H(k, l)H(k, l, m, n, t, f)} dt df \\ & + \sum_{c,m,n} \iint X_I(k, l, m, n, c, t, f) \\ & \cdot \log \frac{X_I(k, l, m, n, c, t, f)}{w_J(k, l)r(k, l, c)w_I(k, l)I(k, l, m, n, t, f)} dt df \\ & + \sum_{c,m,n} \iint r(k, l, c)Y_H(k, l, m, n, t, f) \end{aligned}$$

$$\begin{aligned}
& \cdot \log \frac{r(k, l, c)Y_H(k, l, m, n, t, f)}{w_J(k, l)r(k, l, c)w_H(k, l)H(k, l, m, n, t, f)} dt df \\
& + \sum_{c, m, n} \iint r(k, l, c)Y_I(k, l, m, n, t, f) \\
& \cdot \log \frac{r(k, l, c)Y_I(k, l, m, n, t, f)}{w_J(k, l)r(k, l, c)w_I(k, l)I(k, l, m, n, t, f)} dt df. \tag{29}
\end{aligned}$$

4. Experimental Evaluation

We conducted experiments to confirm whether the performance of the source separation using the prior distribution is better than the one using the template sounds. We separated sound mixtures which were generated by mixing musical instrument sounds in the ‘‘RWC Music Database: Musical Instrument Sound’’¹³⁾ according to the SMFs of the ‘‘RWC Music Database: Jazz Music’’ and ‘‘RWC Music Database: Classical Music’’¹⁴⁾ which were excerpted to be about 30 seconds. In this experiment, we compared the following two conditions:

- (1) using the log-likelihood of timbre feature distributions (proposed method, Section 3.1),
- (2) using the template sounds (previous method¹¹⁾, Section 3.2).

4.1 Experimental Conditions

We used 20 SMFs in total, which are listed in **Table 2**: ten SMFs are classical musical pieces and the other ten SMFs are jazz pieces. We prepared musical instrument sounds of 15 instruments listed in **Table 3** from the RWC Music Database: Musical Instrument Sounds¹³⁾ with two performance styles and three instrument bodies. We generated sound mixtures for the test (evaluation) data by mixing the instrument sounds corresponding to the notes in the SMFs. Since we used two performance-style sets and three instrument bodies, six sound mixtures were generated from a SMF.

The prior distributions were trained by using the rest of the instrument sounds. We assumed that $v_H(k, l, m, n)$ and $v_I(k, l, m, n)$ can be decomposed as follows:

$$\begin{aligned}
v_H(k, l, m, n) &= v_H(k, l, m)v_H(k, l, n) \quad \text{and} \\
v_I(k, l, m, n) &= v_I(k, l, m)v_I(k, l, n),
\end{aligned}$$

and we used prior distributions, $p(v_H(k, l, m))$, $p(v_H(k, l, n))$, $p(v_I(k, l, m))$ and $p(v_I(k, l, n))$, instead of $p(v_H(k, l, m, n))$ and $p(v_I(k, l, m, n))$.

Table 2 List of SMFs excerpted from RWC Music Database. Instruments are abbreviated, and are explained in Table 3.

Data Symbol	Instruments	Ave. # of sources
Classical No.2	VN, VL, VC, CB, TR, OB, FG, FL	6.23
Classical No.3	VN, VL, VC, CB, TR, OB, FG, CL, FL	6.51
Classical No.12	VN, VL, VC, CB, FL	4.23
Classical No.16	VN, VL, VC, CL	3.30
Classical No.17	VN, VL, VC, CL	3.76
Classical No.22	PF	4.33
Classical No.30	PF	4.94
Classical No.34	PF	5.96
Classical No.39	PF, VN	5.92
Classical No.40	PF, VN	7.54
Jazz No.1	PF	2.75
Jazz No.5	PF	6.92
Jazz No.8	EG	6.47
Jazz No.9	EG	3.23
Jazz No.16	PF, EB	3.55
Jazz No.17	PF, EB	5.19
Jazz No.23	PF, EB, TS	3.64
Jazz No.24	PF, EB, TS	6.28
Jazz No.27	PF, AG, EB, AS, TS, BS	11.71
Jazz No.28	PF, AG, EB, AS, TS, BS	5.46

The experimental procedure was as follows:

- (1) initialize the integrated model of each musical note using the corresponding template sound,
- (2) estimate all the model parameters from the input sound mixture, and
- (3) calculate the signal-to-noise ratio (SNR) for the evaluation.

SNR is defined as follow:

$$\text{SNR} = \frac{1}{C(T_1 - T_0)} \sum_c \int 10 \log_{10} \frac{X_{kl}^{(J)}(c, t)^2}{(X_{kl}^{(J)}(c, t) - Z_{kl}(c, t))^2} dt,$$

where

$$X_{kl}^{(J)}(c, t) = \int X_J(k, l, c, t, f) df \quad \text{and} \quad Z_{kl}(c, t) = \int Z(k, l, c, t, f) df, \tag{30}$$

T_0 and T_1 are the beginning and ending times of the input power spectrogram, $X(c, t, f)$, F_0 and F_1 are the beginning and ending frequencies, and $Z(k, l, c, t, f)$ is the ground-truth power spectrogram corresponding to the (k, l) -th note (i.e.,

Table 3 List of musical instruments. The instrument ID means the unique instrument number in the RWC Music Database: Musical Instrument Sounds¹³.

Inst. name (Abbr.)	Inst. ID	Perf. style set A (Abbr.)	Perf. style set B (Abbr.)
Pianoforte (PF)	No.1	Normal (NO)	Staccato (ST)
Electric Guitar (EG)	No.13	Legato/Pick (LP)	Vibrato/Pick (VP)
Electric Bass (EB)	No.14	Normal/Pick (PN)	Normal/Two-finger (TN)
Violin (VN)	No.15	Normal (NO)	Non-vibrato (NV)
Viola (VL)	No.16	Normal (NO)	Non-vibrato (NV)
Cello (VC)	No.17	Normal (NO)	Non-vibrato (NV)
Contrabass (CB)	No.18	Normal (NO)	Non-vibrato (NV)
Trumpet (TR)	No.21	Normal (NO)	Vibrato (VI)
Alto Sax (AS)	No.26	Normal (NO)	Vibrato (VI)
Tenor Sax (TS)	No.27	Normal (NO)	Vibrato (VI)
Baritone Sax (BS)	No.28	Normal (NO)	Vibrato (VI)
Oboe (OB)	No.29	Normal (NO)	Vibrato (VI)
Fagotto (FG)	No.30	Normal (NO)	Vibrato (VI)
Clarinet (CL)	No.31	Normal (NO)	Vibrato (VI)
Flute (FL)	No.33	Normal (NO)	Vibrato (VI)

Table 4 Experimental conditions.

Frequency Analysis	Sampling rate Analyzing method STFT window STFT shift	16 kHz STFT* 2048 points Gaussian 160 points (10 ms)
Constant Parameters	C M_H N_H M_I N_I ϕ_I ω_{Ia} ω_{Ib}	1 20 30 20 30 0.05 440.0 1.135
MIDI sound generator for template sounds		Roland SD-90

* Short-time Fourier Transform

the spectrogram of an actual sound before mixing). We have *original*, i.e., before mixing, source signals. If we obtain ‘completely’ separated signals, the SNRs of these signals must be positive infinity, or the SNRs will decrease as the separation performance becomes worse. Other experimental conditions are shown in **Table 4**.

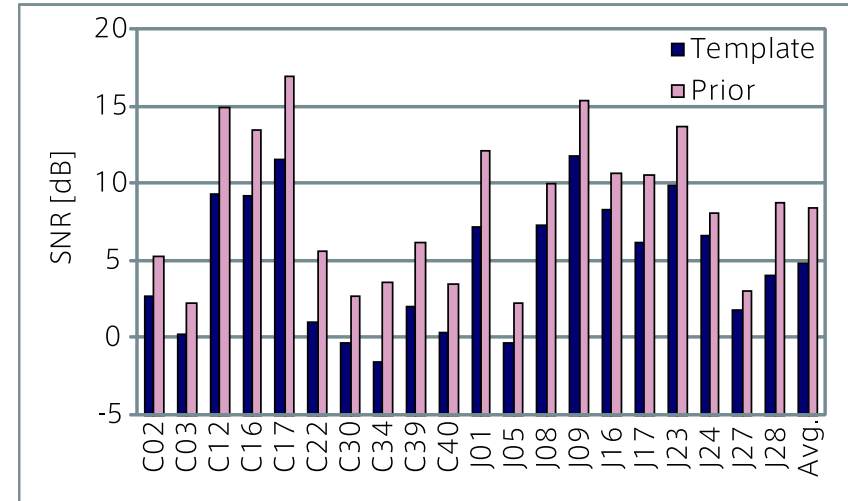


Fig. 6 SNRs of separated signals. [dB]

4.2 Experimental results

The average of SNRs of six sound mixtures for each musical piece is shown in **Fig. 6**, and **Fig. 7** shows the SNRs for each musical instrument and performance style. The SNRs improved from 4.89 to 8.48 dB in average by using the prior distributions. This result shows the robustness and effectiveness of our model parameter estimation method under the timbre difference between musical instrument sounds consisting of input sound mixtures and template sounds. Template sounds were generated from only one musical instrument body and performance style. These bodies and styles would be different from the ones of the input mixture signals and this difference decreased the separation performance.

The SNRs of pianoforte (PF) show a difference of more than 10 dB between the normal (NO) and the staccato (ST) styles, although the difference of other instruments between styles is at most 5 dB. Pianoforte sounds with the staccato style have long silence period because the duration of these sounds is shorter than each note in the test data. Noises in the silence period decrease the SNR even though the noises added to the separated signal is little.

The SNRs of the electric bass (EB) with the pick/normal (PN) style, contrabass

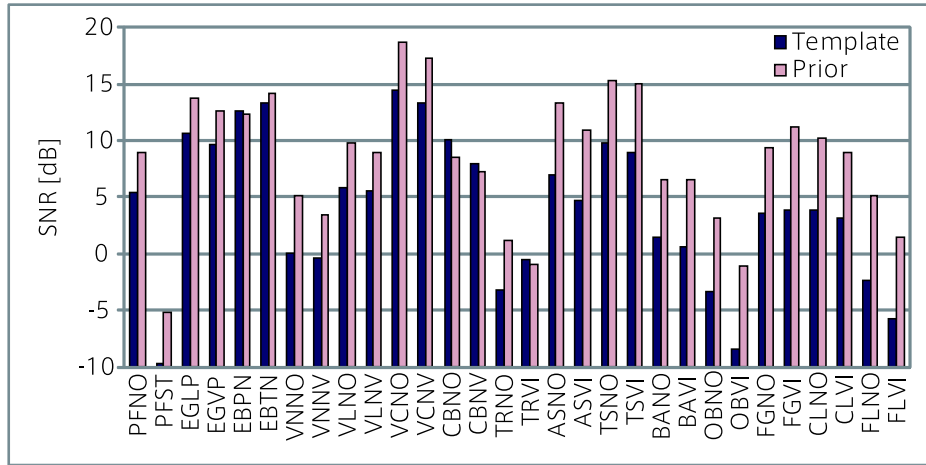


Fig. 7 SNRs of separated signals for each musical instrument.

(CB) with both styles, and trumpet (TR) with vibrato (VI) style decreased, as shown in Fig. 7. This decrease is considered to be caused by the following reasons:

- (1) the prior distributions with inappropriate parameter values,
- (2) the frequency resolution in low-frequency area.

In the future, reason (1) could be corrected by using an appropriate prior distribution, such as a mixture of the dirichlet distributions. This approach is effective in dealing with the timbre difference caused by performance styles. Reason (2) could be corrected by increasing the length of the Short-time Fourier Transform (STFT) window or using a nonlinear frequency analysis method, such as the wavelet transform.

4.3 Discussion

As shown in **Fig. 8**, there was a correlation between the SNR and the average number of notes for each musical piece. The Pearson product-moment correlation coefficient of these values is -0.59 . The average number of notes indicates the *difficulty* in separating the signal, and the average number can be used to evaluate the test data itself. **Fig. 9** shows the correlation of the averaged SNR for each frame of each musical note and the number of notes in the corresponding frame. The SNR in the frames in which the number of sources was less than 6 was

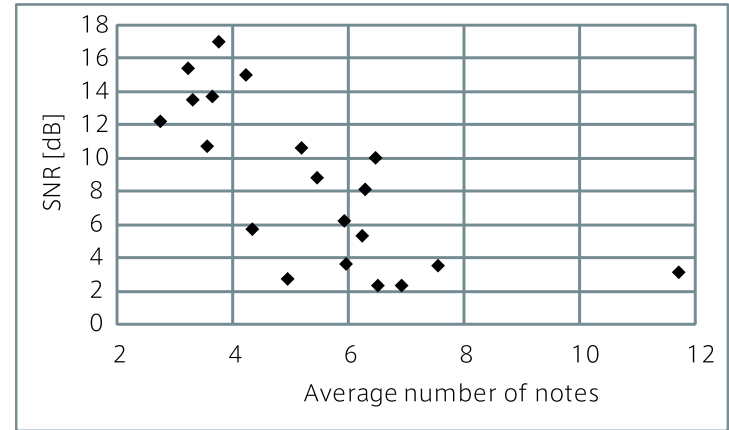


Fig. 8 Correlation between SNR and average number of notes for each musical piece.

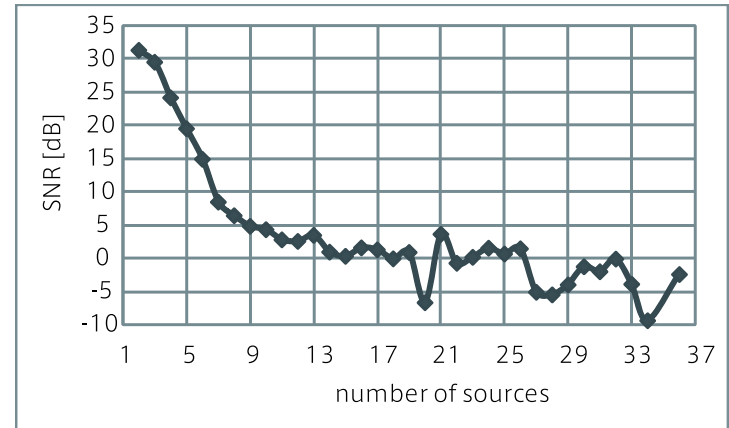


Fig. 9 Correlation between averaged SNR for each frame of each musical note and the number of notes performed in the corresponding frame.

more than 10 dB, and the SNR in the frames in which the number of sources was more than 9 was less than 5 dB. The validity of the additive approximation of the power spectrogram decreases as the number of sources increases, and this causes the separation performance decrease. These results mean additive

approximation is not effective when many instruments play simultaneously. To improve the performance of the source separation in these frames with a large number of sources, we will have to consider:

- restoration of the distorted signals, and
- decomposition of completely additive spectrogram (i.e., a complex spectrogram).

5. Conclusion

We described a new parameter estimation method for an integrated model by using the timbre feature distributions. We confirmed the following results:

- (1) our method increased the separation performance for most instruments,
- (2) in several musical instrument sounds which have very short duration or low frequency components, the separation performance decreased, and
- (3) the separation performance was affected to the validity of the additive approximation.

Our separation framework can be used as an instrument recognition method by regarding the prior distribution as a recognizer. Therefore, we plan to apply our method to the recognition problem by extending it to parallel processing of separation and recognition. Future work will also include the application of the separated signals to various music listening interfaces.

Acknowledgments This research was partially supported by the Ministry of Education, Science, Sports and Culture, Grant-in-Aid for Scientific Research of Priority Areas, Primordial Knowledge Model Core of Global COE program and CrestMuse Project.

References

- 1) Casey, M. and Westner, A.: Separation of Mixed Audio Sources by Independent Subspace Analysis, *Proc. ICMC*, pp.154–161 (2000).
- 2) Virtanen, T. and Klapuri, A.: Separation of Harmonic Sounds Using Linear Models for the Overtone Series, *Proc. ICASSP*, pp.1757–1760 (2002).
- 3) Every, M. and Szymanski, J.: A Spectral-filtering Approach to Music Signal Separation, *Proc. DAFX*, pp.197–200 (2004).
- 4) Woodruff, J., Pardo, B. and Dannenberg, R.: Remixing Stereo Music with Score-informed Source Separation, *Proc. ISMIR*, pp.314–319 (2006).
- 5) Viste, H. and Evangelista, G.: A Method for Separation of Overlapping Partial Based on Similarity of Temporal Envelopes in Multichannel Mixtures, *IEEE Trans. Audio, Speech and Lang. Process.*, Vol.14, No.3, pp.1051–1061 (2006).
- 6) Klapuri, A.: Multiple Fundamental Frequency Estimation based on Harmonicity and Spectral Smoothness, *IEEE Trans. Speech and Audio Process.*, Vol.11, No.6, pp.804–816 (2003).
- 7) Smaragdis, P. and Brown, J.C.: Non-negative Matrix Factorization for Polyphonic Music Transcription, *Proc. WASPAA*, pp.177–180 (2003).
- 8) Bertin, N., Badeau, R. and Richard, G.: Blind Signal Decompositions for Automatic Transcription of Polyphonic Music: NMF and K-SVD on the Benchmark, *Proc. ICASSP*, pp.65–68 (2007).
- 9) Ryyänen, M. and Klapuri, A.: Automatic Bass Line Transcription from Streaming Polyphonic Audio, *Proc. ICASSP*, pp.1437–1440 (2007).
- 10) Barry, D., Fitzgerald, D., Coyle, E. and Lawlor, B.: Drum Source Separation Using Percussive Feature Detection and Spectral Modulation, *Proc. ISSC*, pp.13–17 (2005).
- 11) Itoyama, K., Goto, M., Komatani, K., Ogata, T. and Okuno, H.: Integration and Adaptation of Harmonic and Inharmonic Models for Separating Polyphonic Musical Signals, *Proc. ICASSP*, pp.57–60 (2006).
- 12) Kameoka, H., Nishimoto, T. and Sagayama, S.: Harmonic-temporal Structured Clustering via Deterministic Annealing EM Algorithm for Audio Feature Extraction, *Proc. ISMIR*, pp.115–122 (2005).
- 13) Goto, M., Hashiguchi, H., Nishimura, T. and Oka, R.: RWC Music Database: Music Genre Database and Musical Instrument Sound Database, *Proc. ISMIR*, pp. 229–230 (2003).
- 14) Goto, M., Hashiguchi, H., Nishimura, T. and Oka, R.: RWC Music Database: Popular, Classical, and Jazz Music Databases, *Proc. ISMIR*, pp.287–288 (2002).

Appendix: Derivation of the parameter update equation

In this appendix, we describe the update equations of each parameter derived from the M-step of the EM algorithm. By differentiating the cost function for each parameter, the update equations were derived as follows:

$$w_J(k, l) = \frac{X_J(k, l)}{C}, \quad (31)$$

$$r(k, l, c) = \frac{C \iint X_J(k, l, c, t, f) dt df}{X_J(k, l)}, \quad (32)$$

$$w_H(k, l) = \frac{X_H(k, l) + (\alpha_{wH}(k) - 1)}{X_J(k, l) + (\alpha_{wH}(k) - 1) + (\alpha_{wI}(k) - 1)}, \quad (33)$$

$$w_I(k, l) = \frac{X_I(k, l) + (\alpha_{wI}(k) - 1)}{X_J(k, l) + (\alpha_{wH}(k) - 1) + (\alpha_{wI}(k) - 1)}, \quad (34)$$

$$v_H(k, l, m, n) = \frac{\sum_c \iint X_H(k, l, m, n, c, t, f) dt df + (\alpha_{vH}(k, m, n) - 1)}{X_H(k, l) + \sum_{m, n} (\alpha_{vH}(k, m, n) - 1)}, \quad (35)$$

$$v_I(k, l, m, n) = \frac{\sum_c \iint X_I(k, l, m, n, c, t, f) dt df + (\alpha_{vI}(k, m, n) - 1)}{X_I(k, l) + \sum_{m, n} (\alpha_{vI}(k, m, n) - 1)}, \quad (36)$$

$$\tau(k, l) = \frac{\sum_{c, m, n} \iint (t - m\phi_H(k, l)) X_H(k, l, m, n, c, t, f) dt df}{X_H(k, l)}, \quad (37)$$

$$\omega_H(k, l, t) = \frac{\sum_{c, m, n} \iint n f X_H(k, l, m, n, c, t, f) df}{\sum_{c, m, n} \iint n^2 X_H(k, l, m, n, c, t, f) df}, \quad (38)$$

$$\phi_H(k, l) = \frac{-a_{\phi H}(k, l) + \sqrt{a_{\phi H}(k, l)^2 + 4b_{\phi H}(k, l)X_H(k, l)}}{2X_H(k, l)}, \quad (39)$$

$$\sigma_H(k, l) = \sqrt{\frac{\sum_{c, m, n} \iint (f - n\omega_H(k, l, t))^2 X_H(k, l, m, n, c, t, f) dt df}{X_H(k, l)}}, \quad (40)$$

where

$$X_J(k, l) = \sum_c \iint X_J(k, l, c, t, f) dt df, \quad (41)$$

$$X_H(k, l) = \sum_{c, m, n} \iint X_H(k, l, m, n, c, t, f) dt df, \quad (42)$$

$$X_I(k, l) = \sum_{c, m, n} \iint X_I(k, l, m, n, c, t, f) dt df, \quad (43)$$

$$a_{\phi H}(k, l) = \sum_{c, m, n} \iint m(t - \tau(k, l)) X_H(k, l, m, n, c, t, f) dt df \quad \text{and} \quad (44)$$

$$b_{\phi H}(k, l) = \sum_{c, m, n} \iint (t - \tau(k, l))^2 X_H(k, l, m, n, c, t, f) dt df. \quad (45)$$

(Received May 12, 2008)

(Accepted April 6, 2009)

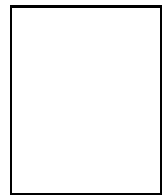
(Released January 1, 2009)



Katsutoshi Itoyama received the B.E. degree in 2006 and the M.S. degree in Informatics in 2008 from Kyoto University, Japan. He is currently a Ph.D. candidate in Informatics, Kyoto University. He is supported by the JSPS Research Fellowships for Young Scientists (DC1). His research interests include musical sound source separation, music listening interfaces, and music information retrieval. He received the 24th TAF Telecom Student Technology Award. He is a member of the IPSJ and IEEE.



Masataka Goto received the D.E. degree from Waseda University, Japan, in 1998. He is currently a Leader of Media Interaction Group, Information Technology Research Institute at the National Institute of Advanced Industrial Science and Technology (AIST). He serves concurrently as a Visiting Professor at the Institute of Statistical Mathematics and an Associate Professor (Cooperative Graduate School Program) at University of Tsukuba. He received 24 awards, including the Commendation for Science and Technology by the Minister of MEXT “Young Scientists’ Prize”, the DoCoMo Mobile Science Awards “Excellence Award in Fundamental Science”, and IPSJ Best Paper Award.



Kazunori Komatani received his B.E. degree in 1998, his M.S. degree in Informatics in 2000, and his Ph.D. in 2002, all from Kyoto University. He is currently an Assistant Professor of the Graduate School of Informatics, Kyoto University, Japan. From 2008 to 2009, he was a Visiting Scientist at Carnegie Mellon University, Pittsburgh, PA, USA. He has received several awards including the 2002 FIT Young Researcher Award and 2004 IPSJ Yamashita SIG Research Award, both from the Information Processing Society of Japan (IPSJ). His research interests center on spoken language processing, especially on spoken dialogue systems. He is a member of the IPSJ, Institute of Electronics, Information and Communication Engineers (IEICE), Association for Natural Language Processing (NLP), Japanese Society for Artificial Intelligence (JSAI), Association for Computational Linguistics (ACL), and International Speech Communication Association (ISCA).



Tetsuya Ogata received the B.S., M.S. and D.E. degrees in Mechanical Engineering in 1993, 1995, and 2000, respectively, from Waseda University. From 1999 to 2001, he was a Research Associate in Waseda University. From 2001 to 2003, he was a Research Scientist in the Brain Science Institute, RIKEN. Since 2003, he has been a Faculty Member in the Graduate School of Informatics, Kyoto University, where he is currently an Associate Professor. Since 2005, he has been a Visiting Associate Professor of the Humanoid Robotics Institute of Waseda University. His research interests "interaction emergence systems" including human-robot vocal-sound interaction, dynamics of human-robot mutual adaptation, and active sensing with robot systems. Dr. Ogata received the 2000 JSME Outstanding Paper Medal from the Japan Society of Mechanical Engineers, and the Best Paper Award of IEA/AIE-2005. He is a member of the IPSJ, JSAI, RSJ, HIS, SICE, and IEEE.



Hiroshi G. Okuno received the B.A. and Ph.D degrees from the University of Tokyo, Japan, in 1972 and 1996, respectively. He is currently a Professor of the Graduate School of Informatics, Kyoto University, Japan. He received various awards including the Best Paper Awards of JSAI. His research interests include computational auditory scene analysis, robot audition and music scene analysis.
