SIMULTANEOUS PROCESSING OF SOUND SOURCE SEPARATION AND MUSICAL INSTRUMENT IDENTIFICATION USING BAYESIAN SPECTRAL MODELING

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

† Graduate School of Informatics, Kyoto University, Japan
∗ JSPS Research Fellowship for Young Scientists (DC1)
‡ National Institute of Advanced Industrial Science and Technology (AIST), Japan

ABSTRACT

This paper presents a method of both separating audio mixtures into sound sources and identifying the musical instruments of the sources. A statistical tone model of the power spectrogram, called an integrated model, is defined and source separation and instrument identification are carried out on the basis of Bayesian inference. Since, the parameter distributions of the integrated model depend on each instrument, the instrument name is identified by selecting the one that has the maximum relative instrument weight. Experimental results showed correct instrument identification enables precise source separation even when many overtones overlap.

Index Terms— Source separation, instrument identification, Bayesian methods, spectrogram

1. INTRODUCTION

Musical instrument identification in complex musical audio mixtures and sound source separation of instrument sounds are challenging problems in musical audio processing. These problems have thus far been treated independently. For example, methods of musical instrument identification [1, 2] have been reported based on fundamental frequency (F0) estimates [3, 4, 5], tempo estimates and beat tracking [6, 7, 8]. Methods of sound source separation have also been reported for separating harmonic sounds [9, 10] and separating percussive ones [11, 12]. Although methods of blind source separation and source (talker) identification have been reported [13] for multi-channel audio signals recorded by using a microphone array, these methods cannot be applied to musical audio signals since most musical audio signals are monaural or stereo.

We believe that instrument identification and source separation rely on each other, i.e., accurate instrument identification should help source separation and high-quality source separation should simplify instrument identification. This paper reports a method of both separating audio mixtures and identifying instruments for each sound. The inputs are an audio mixture of instrument sounds, a number of mixed sounds, and the rough onset time and F0 of each sound, and the outputs are separated audio signals and the instrument name of each sound. We solved source separation as the decomposition of the input power spectrogram based on the responsibility for each instrument sound, and instrument identification as the selection of the spectral tone model based on maximum *A Posteriori* approximation. Since the distributions of the tone model parameters differ by instrument, we used prior distributions of the parameters, which were trained by using a musical instrument sound database.

2. BAYESIAN SPECTRAL MODELING

In this section, we define a stochastic tone model, which consists of harmonic and inharmonic tone models, introduce prior distributions of the tone model parameters, and describe source separation and instrument identification methods based on Bayesian inference. Let an observed power spectrogram, be a histogram on the two-dimensional area of time and frequency, (t, f). We assume that the spectrogram is obtained by using a short-time Fourier transform (STFT). Since the elements of a power spectrogram are not generally integers, we approximate them as integers by multiplying a sufficiently large number and rounding them. Let N be the number of samples on the histogram, $X = (x_1, \ldots, x_N)$ be a whole set of samples, and $x_n = (t_n, f_n)$ $(n = 1, \ldots, N)$ be each sample. We define separating sound sources as decomposing a power spectrogram of an audio mixture into a power spectrogram that corresponds to each sound.

Let J be the number of musical instrument sounds performed in the audio mixture and K be the number of candidate musical instruments. Onset time, duration, and pitch of each sound are given and the instrument which performed the sound is unknown. Our goal is both of estimating instruments which performed each sound, i.e., instrument identification, and decomposing the input power spectrogram to each sound, i.e., source separation.

2.1. Harmonic and Inharmonic Tone Models

Let $Y_j(t, f)$ be a spectral model which represents the power spectrogram of j-th instrument sound. Since the instrument which performed the sound is unknown, we define the power spectrogram model of an instrument sound as the sum of K models with weight parameter $b_{k|j}$:

$$Y_j(t, f) = \sum_{k=1}^{K} b_{k|j} Y_{k|j}(t, f)\,. \tag{1}$$

We also represent the power spectrogram of the audio mixture by the sum of J models with weight parameter a_j:

$$Y(t, f) = \sum_{j=1}^{J} a_j Y_j(t, f)\,. \tag{2}$$

To represent the power spectrogram of the k-th instrument of the j-th sound ((j, k)-th sound), we use a statistical tone model, called an integrated harmonic and inharmonic models [14]. All musical instrument sound consists of harmonic sounds generated from a vibration of strings and air column, and inharmonic (percussive) ones after musical instrument's excitation. Power spectrograms of various musical instruments, e.g., clarinet and marimba which have large harmonic and inharmonic energy, respectively, can be represented in the same structure by using the integrated models.

The integrated model is defined as the sum of harmonic structure model, $Y_{\mathrm{H}|j,k}(t, f)$, and inharmonic structure model, $Y_{\mathrm{I}|j,k}(t, f)$, with weight parameters $c_{\mathrm{H}|j,k}$ and $c_{\mathrm{I}|j,k}$:

$$Y_{k|j}(t, f) = c_{\mathrm{H}|j,k} Y_{\mathrm{H}|j,k}(t, f) + c_{\mathrm{I}|j,k} Y_{\mathrm{I}|j,k}(t, f)\,. \tag{3}$$

The harmonic-structure tone model is defined as the constrained two-dimensional Gaussian Mixture model (GMM), which is a product of two constrained one-dimensional GMMs. This model is designed by referring to the harmonic temporal structured clustering (HTC) source model [2]. The inharmonic-structure tone model is analogously defined as a constrained two-dimensional GMM. The temporal structures of these tone models are defined in similar form to the harmonic one, but the frequency structures are defined as different forms. These models are defined as:

$$Y_{\mathrm{H}|j,k}(t,f) = \sum_{l=0}^{L_{\mathrm{H}}-1} \sum_{m=1}^{M_{\mathrm{H}}} d_{l,m|j,k} Y_{l,m|j,k,\mathrm{H}}(t,f)\,, \tag{4}$$

$$Y_{l,m|j,k,\mathrm{H}}(t,f) = \mathcal{N}\big(t;\tau_j + l\rho, \rho^2\big)\,\mathcal{N}\big(f;m\omega_j,\sigma_j^2\big)\,, \tag{5}$$

$$Y_{\mathrm{I}|j,k}(t,f) = \sum_{l=0}^{L_{\mathrm{I}}-1} \sum_{m=1}^{M_{\mathrm{I}}} e_{l,m|j,k} Y_{l,m|j,k,\mathrm{I}}(t,f)\,, \text{ and} \tag{6}$$

$$Y_{l,m|j,k,\mathrm{I}}(t,f) = \mathcal{N}\big(t;\tau_j + l\varrho, \varrho^2\big)\,\mathcal{M}(f;m,\lambda,\kappa)\,, \tag{7}$$

where

$$\mathcal{N}\big(x;\mu,\sigma^2\big) = \frac{1}{\sqrt{2\pi}\sigma}\exp\!\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \text{ and} \tag{8}$$

$$\mathcal{M}(x;\mu,\beta,\kappa) = \frac{\beta}{\sqrt{2\pi}(x+\kappa)}\exp\!\left(-\frac{(\beta\log(x/\kappa+1)-\mu)^2}{2}\right). \tag{9}$$

L_{H} and M_{H} are the number of Gaussian kernels that represent the temporal and frequency parts of the harmonic model, and L_{I} and M_{I} are the number of Gaussians that respectively represent those of the inharmonic model. The parameters are described as follows:

- $\boldsymbol{a} = (a_1,\ldots,a_J)$. This parameter represents the relative weight (volume) of the j-th note. The $\boldsymbol{a}$ is in the standard $(J-1)$-simplex, i.e., $\boldsymbol{a}$ satisfies $0 \le a_j \le 1$ $(j = 1,\ldots,J)$ and $\sum_j a_j = 1$.
- $\boldsymbol{b}_j = (b_{1|j},\ldots,b_{K|j})$. This parameter represents the relative weight of the k-th instrument of the j-th sound. Each $\boldsymbol{b}_j$ is in the $(K-1)$-simplex.
- $\boldsymbol{c}_{j,k} = (c_{\mathrm{H}|j,k}, c_{\mathrm{I}|j,k})$. This parameter represents the relative weight of the harmonic and inharmonic components. Each $\boldsymbol{c}_{j,k}$ is in the standard 1-simplex.
- $\boldsymbol{d}_{j,k} = (d_{0,1|j,k},\ldots,d_{L_{\mathrm{H}}-1,M_{\mathrm{H}}|j,k})$. This parameter represents the time-frequency energy distribution of the harmonic component. Each $\boldsymbol{d}_{j,k}$ is in the standard $(L_{\mathrm{H}}\times M_{\mathrm{H}}-1)$-simplex.
- $\boldsymbol{e}_{j,k} = (e_{0,1|j,k},\ldots,e_{L_{\mathrm{I}}-1,M_{\mathrm{I}}|j,k})$. This parameter represents the time-frequency energy distribution of the inharmonic component. Each $\boldsymbol{e}_{j,k}$ is in the standard $(L_{\mathrm{I}}\times M_{\mathrm{I}}-1)$-simplex.
- τ_j. This parameter represents the onset time of the j-th sound.
- ρ and ϱ. The $L_{\mathrm{H}}\rho$ and $L_{\mathrm{I}}\varrho$ mean the duration of the harmonic component and inharmonic component, respectively.
- ω_j. This parameter represents the fundamental frequency.
- σ_j. This parameter represents the deviation in energy distribution along the frequency axis.
- λ and κ. These parameters are coefficients that determine the arrangement of the Gaussian kernels for the frequency structure of the inharmonic model[1].

[1] If λ and κ are set to 1127 and 700, respectively, $\lambda\log(f/\kappa+1)$ is equivalent to the mel scale of f Hz.

Since the integrated model is defined as a hierarchical weighted mixture of Gaussian distributions, we introduce latent variables, which specify the element distribution which generate sample x_n:

- $\boldsymbol{z}^n = (z_1^n,\ldots,z_J^n)$. $z_j^n = 1$ means sample x_n is generated from the j-th sound.
- $\boldsymbol{z}_j^n = (z_{1|j}^n,\ldots,z_{K|j}^n)$. $z_{k|j}^n = 1$ means sample x_n is generated from the (j,k)-th sound.
- $\boldsymbol{z}_{j,k}^n = (z_{\mathrm{H}|j,k}^n, z_{\mathrm{I}|j,k}^n)$. $z_{\mathrm{H}|j,k}^n = 1$ and $z_{\mathrm{I}|j,k}^n = 1$ means sample x_n is generated from the harmonic and inharmonic components of the (j,k)-th sound, respectively.
- $\boldsymbol{z}_{j,k,\mathrm{H}}^n = (z_{0,1|j,k,\mathrm{H}}^n,\ldots,z_{L_{\mathrm{H}}-1,M_{\mathrm{H}}|j,k,\mathrm{H}}^n)$. $z_{l,m|j,k,\mathrm{H}}^n = 1$ means sample x_n is generated from the component whose temporary and frequency indices are l and m ((l,m)-th component) of the (j,k)-th sound's harmonic component.
- $\boldsymbol{z}_{j,k,\mathrm{I}}^n = (z_{0,1|j,k,\mathrm{I}}^n,\ldots,z_{L_{\mathrm{I}}-1,M_{\mathrm{I}}|j,k,\mathrm{I}}^n)$. $z_{l,m|j,k,\mathrm{I}}^n = 1$ means sample x_n is generated from the (l,m)-th component of the (j,k)-th sound's inharmonic component.

$\boldsymbol{z}^n$ has a 1-of-J representation, i.e., one of z_j^n is 1 and the others are 0, thus $\boldsymbol{z}^n$ satisfies $z_j^n \in \{0,1\}$ and $\sum_j z_j^n = 1$. Other latent variables have the same representation.

2.2. Prior distribution

We introduce prior distributions to prevent the model parameters from deviating in source separation and instrument identification. For example, the energy distribution of the inharmonic component generally converges just after sound excitation and decreases with time, so usually $e_{l,m|j,k} > e_{l',m|j,k}$ $(l < l')$. Since acoustic features, e.g., relative amplitude of the harmonic components, are different for each musical instrument, we use different prior distributions of the model parameters for each instrument. Prior distributions are trained by estimating the model parameters for isolated musical instrument sounds from a sound database with noninformative priors and averaging them. Let θ be a whole set of model parameters; the prior distributions are described as:

$$p(\theta) = p(\boldsymbol{a})\prod_j p(\boldsymbol{c}_{j,k})\,p(\boldsymbol{d}_{j,k})\,p(\boldsymbol{e}_{j,k})\,p(\tau_j)\,p(\omega_j,\sigma_j)\,, \tag{10}$$

$$p(\boldsymbol{a}) = \mathcal{D}(\boldsymbol{a};\boldsymbol{\alpha}) \quad (\boldsymbol{\alpha} = (\alpha_1,\ldots,\alpha_J)), \tag{11}$$

$$p(\boldsymbol{b}_j) = \mathcal{D}(\boldsymbol{b}_j;\boldsymbol{\beta}) \quad (\boldsymbol{\beta} = (\beta_1,\ldots,\beta_K)), \tag{12}$$

$$p(\boldsymbol{c}_{j,k}) = \mathcal{D}(\boldsymbol{c}_{j,k};\boldsymbol{\gamma_j}) \quad (\boldsymbol{\gamma}_j = (\gamma_{j,\mathrm{H}},\gamma_{j,\mathrm{I}})), \tag{13}$$

$$p(\boldsymbol{d}_{j,k}) = \mathcal{D}(\boldsymbol{d}_{j,k};\boldsymbol{\delta_j}) \quad (\boldsymbol{\delta}_j = (\delta_{j,0,1},\ldots\delta_{j,L_{\mathrm{H}}-1,M_{\mathrm{H}}})), \tag{14}$$

$$p(\boldsymbol{e}_{j,k}) = \mathcal{D}(\boldsymbol{e}_{j,k};\boldsymbol{\varepsilon_j}) \quad (\boldsymbol{\varepsilon}_j = (\varepsilon_{j,0,1},\ldots\varepsilon_{j,L_{\mathrm{I}}-1,M_{\mathrm{I}}})), \tag{15}$$

$$p(\tau_j) = \mathcal{N}\big(\tau_j;\nu_k,\xi_k^{-1}\big)\,, \text{ and} \tag{16}$$

$$p(\omega_j,\sigma_j) = \mathcal{N}\big(\omega_j;\varphi_k,\sigma_j^{-2}\chi_k\big)\,\mathcal{G}\big(\sigma_j^{-2};\eta_k,\zeta_k\big)\,. \tag{17}$$

Prior distributions are defined as a conjugate prior of the corresponding parameters. Parameters without prior distributions, ρ, ϱ, λ, and κ, are treated as constants. The $\mathcal{D}(\cdot)$ and $\mathcal{G}(\cdot)$ mean Dirichlet and gamma distributions. The probabilistic density functions of these distributions are given as follows except for normalizing factors:

$$\mathcal{D}(x_1,\ldots,x_N;\phi_1,\ldots,\phi_N) \propto \prod_{n=1}^{N} {x_n}^{\phi_n-1}, \tag{18}$$

$$\mathcal{G}(x;\eta,\zeta) \propto x^{\eta-1}\exp(-\zeta x)\,. \tag{19}$$

Let Z be a whole set of latent variables; a probabilistic model with the latent variables is described as:

$$p(X,Z,\theta) = \prod_{n=1}^{N} p\big(x_n,\boldsymbol{z}^n,\boldsymbol{z}_j^n,\boldsymbol{z}_{j,k}^n,\boldsymbol{z}_{j,k,\mathrm{H}}^n,\boldsymbol{z}_{j,k,\mathrm{I}}^n|\theta\big)\,p(\theta)\,, \tag{20}$$

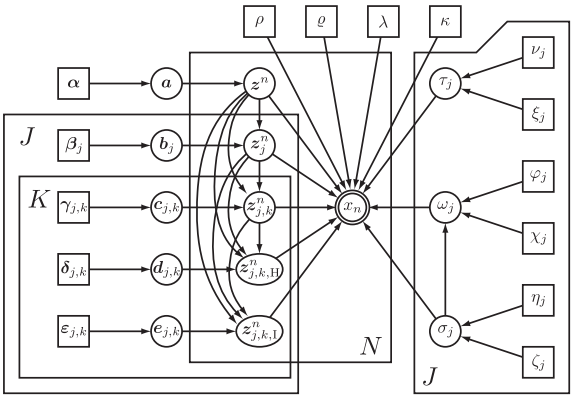

Fig. 1. Graphical model of integrated model.

$$p\big(x_n, \boldsymbol{z}^n, \boldsymbol{z}^n_j, \boldsymbol{z}^n_{j,k}, \boldsymbol{z}^n_{j,k,\mathrm{H}}, \boldsymbol{z}^n_{j,k,\mathrm{I}} | \theta\big) = \prod_{j=1}^{J} \big(a_j p\big(x_n, \boldsymbol{z}^n_j, \boldsymbol{z}^n_{j,k}, \boldsymbol{z}^n_{j,k,\mathrm{H}}, \boldsymbol{z}^n_{j,k,\mathrm{I}} | z^n_j = 1, \theta\big)\big)^{z^n_j}, \tag{21}$$

$$p\big(x_n, \boldsymbol{z}^n_j, \boldsymbol{z}^n_{j,k}, \boldsymbol{z}^n_{j,k,\mathrm{H}}, \boldsymbol{z}^n_{j,k,\mathrm{I}} | z^n_j = 1, \theta\big) = \prod_{k=1}^{K} \big(b_{k|j} p\big(x_n, \boldsymbol{z}^n_{j,k}, \boldsymbol{z}^n_{j,k,\mathrm{H}}, \boldsymbol{z}^n_{j,k,\mathrm{I}} | z^n_{k|j} = 1, \theta\big)\big)^{z^n_{k|j}}, \tag{22}$$

$$p\big(x_n, \boldsymbol{z}^n_{j,k}, \boldsymbol{z}^n_{j,k,\mathrm{H}}, \boldsymbol{z}^n_{j,k,\mathrm{I}} | z^n_{k|j} = 1, \theta\big) = \big(c_{\mathrm{H}|j,k} p\big(x_n, \boldsymbol{z}^n_{j,k,\mathrm{H}} | z^n_{\mathrm{H}|j,k} = 1, \theta\big)\big)^{z^n_{\mathrm{H}|j,k}} \times \big(c_{\mathrm{I}|j,k} p\big(x_n, \boldsymbol{z}^n_{j,k,\mathrm{I}} | z^n_{\mathrm{I}|j,k} = 1, \theta\big)\big)^{z^n_{\mathrm{I}|j,k}}, \tag{23}$$

$$p\big(x_n, \boldsymbol{z}^n_{j,k,\mathrm{H}} | z^n_{\mathrm{H}|j,k} = 1, \theta\big) = \prod_{l=0}^{L_\mathrm{H}-1} \prod_{m=1}^{M_\mathrm{H}} \big(d_{l,m|j,k} Y_{l,m|j,k,\mathrm{H}}(t,f)\big)^{z_{l,m|j,k,\mathrm{H}}}, \text{ and} \tag{24}$$

$$p\big(x_n, \boldsymbol{z}^n_{j,k,\mathrm{I}} | z^n_{\mathrm{I}|j,k} = 1, \theta\big) = \prod_{l=0}^{L_\mathrm{I}-1} \prod_{m=1}^{M_\mathrm{I}} \big(e_{l,m|j,k} Y_{l,m|j,k,\mathrm{I}}(t,f)\big)^{z_{l,m|j,k,\mathrm{I}}}. \tag{25}$$

A graphical model of the observation is given in Fig. 1.

2.3. Bayesian inference

As we described, source separation is defined as the decomposition of the input power spectrogram. Decomposed power spectrogram of the j-th sound, $X_j(t,f)$, is obtained by multiplying the responsibility, i.e., the expectation value of latent variable z^n_j, with the observed power spectrogram:

$$X_j(t,f) = \sum_{x_n \in \{x | x \in X, x=(t,f)\}} p\big(z^n_j = 1 | X\big). \tag{26}$$

Separated audio signals are obtained by an inverse STFT of the decomposed spectrograms.

Instrument identification is performed by model selection based on Bayesian inference. The instruments are estimated by using model selection based on maximum *A Posteriori* approximation:

$$\text{(Instrument of } j\text{-th note)} = \arg\max_k \big\langle b_{k|j} \big\rangle_{p(\boldsymbol{b}_j | X)}, \tag{27}$$

where $\langle x \rangle_{f(x)}$ means the expectation value of the random variable x with the density function $f(x)$, i.e., $\langle x \rangle_{f(x)} = \int x f(x)\, dx$.

The posterior probabilities in above equations are defined as:

$$p\big(z^n_j = 1 | X\big) = \frac{p\big(X, z^n_j = 1, Z_{-\boldsymbol{z}^n}, \theta\big)\, dZ_{-\boldsymbol{z}^n}\, d\theta}{p(X, Z, \theta)\, dZ\, d\theta} \quad \text{and} \tag{28}$$

$$p(\boldsymbol{b}_j | X) = \frac{p(X, Z, \theta)\, dZ\, d\theta_{-\boldsymbol{b}_j}}{p(X, Z, \theta)\, dZ\, d\theta}, \tag{29}$$

where $Z_{-\boldsymbol{z}^n}$ and $\theta_{-\boldsymbol{b}_j}$ are the whole set of the latent variables and model parameters except for $\boldsymbol{z}^n$ and $\boldsymbol{b}_j$, respectively, but it is impractical to directly calculate that integral since the parameter space is extremely huge. Therefore, we adopt a method of variational inference for estimating the posterior probability of the latent variables and model parameters.

Let $q(Z, \theta)$ be a test distribution that approximates the posterior distribution, $p(Z, \theta | X)$. We assume that the test distribution can be factorized as:

$$q(Z, \theta) = q(Z)\, q(\theta), \tag{30}$$

$$q(Z) = \prod_{n=1}^{N} q(\boldsymbol{z}^n) \prod_{j=1}^{J} q(\boldsymbol{z}^n_j) \prod_{k=1}^{K} q(\boldsymbol{z}^n_{j,k})\, q(\boldsymbol{z}^n_{j,k,\mathrm{H}})\, q(\boldsymbol{z}^n_{j,k,\mathrm{I}}), \text{ and} \tag{31}$$

$$q(\theta) = q(\boldsymbol{a}) \prod_{j=1}^{J} q(\boldsymbol{b}_j) \left(\prod_{k=1}^{K} q(\boldsymbol{c}_{j,k})\, q(\boldsymbol{d}_{j,k})\, q(\boldsymbol{e}_{j,k}) \right) q(\tau_j)\, q(\omega_j, \sigma_j) \tag{32}$$

An objective function for estimating the optimal $q(z, \theta)$ is defined as:

$$\mathcal{F}[q] = \int \sum_{z \in Z} q(z, \theta) \log \frac{p(X, z, \theta)}{q(z, \theta)}\, d\theta, \tag{33}$$

where $\mathcal{F}[q]$ is a functional that depends on function q. The q that maximizes $\mathcal{F}[q]$ most approximates the posterior distribution, $p(Z, \theta | X)$, under the factorization assumption.

To calculate an optimal test distribution that maximizes the objective function, we solved an Euler-Lagrange equation. All update equations have been omitted since it would take a page to list them.

3. EXPERIMENTAL EVALUATION

We conducted an experiment to evaluate the efficiency of our source separation and instrument identification methods. Given audio mixtures that consisted of two or three musical instrument sounds excerpted from the *RWC Music Database: Musical Instrument Sound* [15], the audio mixtures were separated into sources and instruments were estimated. As shown in Table 1, eight musical instruments were excerpted from the database and sounds were divided into subsets for 10-fold cross validation. The prior distribution of each instrument was created by averaging the model parameters estimated from the training data (nine subsets). Audio mixtures were produced from the combination of the instrument sounds for each data subset except pairs consisting of the same instrument sounds. The constant parameters in the integrated models were set as listed in Table 2. The performance of instrument identification and source separation were respectively evaluated by using the accuracy rate and log spectral distance defined as:

$$\sqrt{\sum_{t=0}^{T} \sum_{f=0}^{F} \left| 20 \log_{10} \frac{X_{\mathrm{org}}(t,f)}{X_{\mathrm{sep}}(t,f)} \right|^2} \Big/ TF \tag{34}$$

Table 3 summarizes the accuracy rate of instrument identification and log spectral distance for the source instruments. The fagotto (FG), violin (VN), and clarinet (CL) have a high accuracy rate for identification and short log spectral distances. This suggests that correct instrument identification help to improve source separation.

Table 1. Musical Instruments.

Inst. name (Abbr.)	# of tones
Acoustic piano (PF)	1584
Violin (VN)	2304
Trumpet (TR)	1964
Alto sax (AS)	891
Clarinet (CL)	1080
Fagotto (FG)	1079
Marimba (MB)	909
Vibraphone (VI)	1332

Table 2. Constant values.

Parameter	Value
ρ_{H}	0.1
ρ_{I}	0.1
λ	1.134
κ	440.0
L_{H}	100
M_{H}	30
L_{I}	100
M_{I}	30

Table 3. Experimental results for instrument identification and source separation, which show averaged log spectral distances in instrument sounds. Bold characters mean top two numbers.

Inst.	Acc. rate [%]		Log spect. dist. ($\times 10^{-2}$)	
name	2 sounds	3 sounds	2 sounds	3 sounds
PF	63.4	28.0	3.25	3.88
VN	**87.8**	76.5	2.43	3.31
TR	79.6	61.5	2.78	3.40
AS	39.1	12.7	3.29	3.97
CL	85.1	**79.4**	**1.59**	**2.12**
FG	**91.7**	**85.1**	**1.83**	**2.39**
MB	48.6	28.2	5.25	5.34
VI	67.6	53.9	6.43	5.86
Avg.	72.1	54.8	3.12	3.65

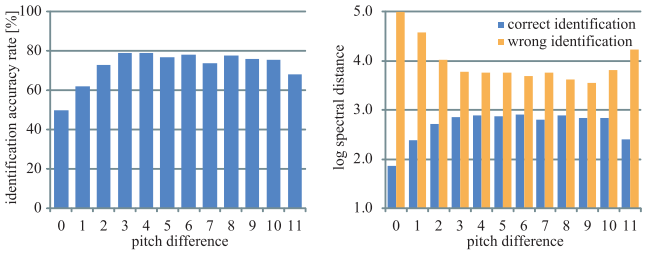

Fig. 2. Relationship between pitch (MIDI note number) difference in two instrument sounds to accuracy rate of instrument identification (left) and between the difference to log spectral distances of separated sounds whose instruments are correctly or incorrectly estimated (right).

It is easier to decompose audio mixture of two sounds than mix of three sounds and decreasing the number of sounds increases the accuracy rate of identification on average. This suggests precise source separation increases the accuracy of instrument identification. The marimba (MB) and vibraphone (VI) have larger spectral distances than the other instruments. These instrument sounds have percussive properties and are sensitive to the diffusion of onset time. The distances can decrease by accurately estimating the onset time.

Fig. 2 shows the relationship between the pitch difference in two instrument sounds when two sounds are mixed to the accuracy of identification and the log spectral distance of separated sounds. The pitch difference is based on the difference of MIDI pitch numbers. The spectral distances are shown in cases of correct and incorrect instrument identification. When pitch differences are 0 (unison), 1, 2, and 11, the Gaussian distribution for F0 overlaps with other harmonics and this overlap decreases the accuracy of identification. When the differences are 5 (perfect fourth) and 7 (perfect fifth), although the F0-Gaussian does not have any overlap, other many harmonics have overlap and the overlap slightly decreases the identification accuracy. This suggests that the overlap of overtones degrades the accuracy of source separation. Spectral distances also degrade when pitch differences are 0, 1, and 11 when instruments are identified incorrectly. However, when instruments are identified correctly, spectral distances did not increase when many overtones overlapped. This suggests that correct instrument identification enables precise source separation even when many overtones overlap.

4. CONCLUSION

We reported a method of simultaneously processing sound source separation and musical instrument identification using Bayesian spectral modeling. We defined the integrated harmonic and inharmonic tone models, decomposed the observed power spectrogram by using the expectation value of the latent variable, and identified the instrument for each sound in the audio mixture by selecting the instrument based on maximum *A Posteriori* approximation. The experimental results revealed that the accuracy of instrument identification and source separation rely on each other and correct instrument identification enables precise source separation even when many overtones overlap.

Acknowledgements: 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 JST CrestMuse Project.

5. REFERENCES

[1] T. Kitahara, *Computational Musical Instrument Recognition and Its Application to Content-based Music Information Retrieval*, Ph.D. thesis, Kyoto University, 2007.

[2] H. Kameoka *et al*., "A multipitch analyzer based on harmonic temporal structured clustering," *IEEE Trans. Audio, Speech and Lang. Process.*, vol. 15, no. 3, pp. 982–994, 2007.

[3] K. Kashino, *Computational Auditory Scene Analysis for Music Signals*, Ph.D. thesis, University of Tokyo, 1994.

[4] M. Goto, "A real-time music-scene-analysis system: Predominant-F0 estimation for detecting melody and bass lines in real-world audio signals," *Speech Communication*, vol. 43, no. 4, pp. 311–329, 2004.

[5] A. Klapuri, "Multipitch analysis of polyphonic music and speech signals using an auditory model," *IEEE Trans. Audio, Speech and Lang. Process.*, vol. 16, no. 2, pp. 255–266, 2008.

[6] M. Goto, "An audio-based real-time beat tracking system for music with or without drum-sounds," *J. New Music Res.*, vol. 30, no. 2, pp. 159–171, 2001.

[7] S. Dixon, "Automatic extraction of tempo and beat from expressive performances," *J. New Music Res.*, vol. 30, no. 1, pp. 39–58, 2001.

[8] M. E. P. Davies and M. D. Plumbley, "Context-dependent beat tracking of musical audio," *IEEE Trans. Audio, Speech and Lang. Process.*, vol. 15, no. 3, pp. 1009–1020, 2007.

[9] J. Woodruff *et al*., "Remixing stereo music with score-informed source separation," in *ISMIR2006*, pp. 314–319.

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

[11] M. A. Casey and A. Westner, "Separation of mixed audio sources by independent subspace analysis," in *ICMC2000*, pp. 154–161.

[12] K. Yoshii *et al*., "Drum sound recognition for polyphonic audio signals by adaptation and matching of spectrogram templates with harmonic structure suppression," *IEEE Trans. Audio, Speech and Lang. Process.*, vol. 15, no. 1, pp. 333–345, 2007.

[13] H. Saruwatari *et al*., "Blind source separation combining independent component analysis and beamforming," *EURASIP Journal on Applied Signal Processing*, vol. 2003, no. 11, pp. 1135–1146, 2003.

[14] K. Itoyama *et al*., "Integration and adaptation of harmonic and inharmonic models for separating polyphonic musical signals," in *ICASSP2006*, pp. 57–60.

[15] M. Goto *et al*., "RWC music database: Music genre database and musical instrument sound database," in *ISMIR2003*, pp. 229–230.