Regular Paper

Robust Multipitch Analyzer against Initialization based on Latent Harmonic Allocation using Overtone Corpus

Daichi Sakaue[1,a) Katsutoshi Itoyama[1] Tetsuya Ogata[1,2]
Hiroshi G. Okuno[1]

Received: July 2, 2012, Accepted: January 11, 2013

Abstract: We present a Bayesian analysis method that estimates the harmonic structure of musical instruments in music signals on the basis of psychoacoustic evidence. Since the main objective of multipitch analysis is joint estimation of the fundamental frequencies and their harmonic structures, the performance of harmonic structure estimation significantly affects fundamental frequency estimation accuracy. Many methods have been proposed for estimating the harmonic structure accurately, but no method has been proposed that satisfies all these requirements: robust against initialization, optimization-free, and psychoacoustically appropriate and thus easy to develop further. Our method satisfies these requirements by explicitly incorporating Terhardt's virtual pitch theory within a Bayesian framework. It does this by automatically learning the valid weight range of the harmonic components using a MIDI synthesizer. The bounds are termed "overtone corpus." Modeling demonstrated that the proposed overtone corpus method can stably estimate the harmonic structure of 40 musical pieces for a wide variety of initial settings.

Keywords: multipitch estimation, harmonic clustering, overtone estimation, musical instrument sounds

1. Introduction

Popular music is usually performed by people playing multiple instruments, for example, piano, guitar, bass, and drums, and one or more people singing [1], [2]. Multipitch analysis is used to estimate the simultaneous pitches (melodies, bass lines, and chords), at each moment in a musical performance. Trained musicians can do this quite accurately, and even untrained listeners can do it to a certain extent [3]. On the other hand, this is still an unsolved problem in music signal processing [4], [5], [6], [7], [8], [9], [10], [11]. This is because music signals contain many fluctuations, including vibrato, time-varying timber, and tempo variation. To make matters worse, most commercially-available audio files are in monaural or stereo format. Despite the obstacles, multipitch analysis is an important research area because the automatic extraction of pitch patterns benefits a wide range of applications, including sound source separation [12], musical signal manipulation [13], [14], [15], musical instrument identification [16], and musical chord recognition [17], [18].

Two approaches have been taken to solving this problem: auditory [8], [19] and statistical [5], [6], [7]. Auditory-approach-based methods try to imitate the human hearing system because humans are good at recognizing multiple pitches sounding simultaneously. This approach is good to a certain extent, but the estimation accuracy is easily saturated. Our knowledge of the human auditory system is limited, which prevents us from perfectly imitating it. Indeed, the recognition result of the human auditory system is not equal to the direct output of the cochlea. The system has a complex mechanism for tracking the sources of the sounds we hear.

By contrast, statistical-approach-based methods try to estimate the relationship between the pitch pattern and other musical aspects, including music structure [20], musical instrument [21], harmonic structure [7], chord [22], and onset [20]. Since these aspects strongly depend on each other, their joint estimation improves the accuracy of multipitch estimation. Bayesian probabilistic models [5], [6], [7], [20], [21], [22] are widely used because they are suitable for representing the probabilistic relationships [23]. In this article, we focus on latent harmonic allocation (LHA) [7], a promising Bayesian multipitch analysis method that estimates the most likely combination of latent variables, including fundamental frequency, pitch activity, and harmonic structure.

Successful estimation of LHA requires precise initialization of the fundamental frequencies (F0s) and the volumes of the sound sources, because the model contains many inappropriate optima. Numerous techniques have been developed for solving this optimization problem. They include careful initialization [6], Gibbs sampling [24], collapsed estimation [7], [25], deterministic annealing [26], [27], [28], and prior distribution optimization [6], [7], [20], [22]. These methods, except the prior-based ones, focus on the search for better optimal points. They do not, however, guarantee that the global optimum corresponds to the most suitable answer to the problem. The prior-based methods directly modify the optimal solution, so they are perceptually more appropriate.

The aim of prior distribution optimization is to avoid invalid es-

[1] Graduate School of Informatics, Kyoto University, Kyoto 606–8501, Japan
[2] Currently, Faculty of Science and Engineering, Waseda University
[a)] dsakaue@kuis.kyoto-u.ac.jp

© 2013 Information Processing Society of Japan

timation results. For example, earlier methods [5], [6] manually set the prior distribution parameters, known as *hyperparameters*, to prefer an exponentially decaying harmonic structure. However, there is an inherent problem with this approach: there is no universal parameter for estimating the multiple pitches of all musical genres and all musical instruments. To make matters worse, the search for an optimal parameter requires very expensive algorithms like cross-validation.

Recently introduced nonparametric methods [7], [20], [22], [29] have been widely adopted because they optimize the model parameters automatically. While this approach is successful to a certain extent, the estimation accuracy easily saturates because these methods only estimate the posterior distribution that maximizes the model likelihood. Also, further improvement is difficult because they are purely mathematical, making it difficult to imitate the recognition process of humans such as multipitch analysis.

To solve the optimization problem, we introduce a new construction of harmonic structure. Our method is prior-based and free from initialization and hyperparameter optimization. The obtained model is suitable for multipitch analysis. For each harmonic structure model, the division of total sound volume into M harmonic components is represented as a point on an $(M-1)$-simplex. We divide the simplex into two regions, one corresponding to valid harmonic structure and the other corresponding to invalid structure. Further, the former region is approximated as a convex hull based on psychoacoustic evidence [30]. We constructed a newly designed probabilistic model that enforces all the harmonic structure parameters contained in the convex hull and successfully derived a variational Bayesian update.

2. Overtone Structure Modeling

2.1 Spectrogram Modeling

To obtain the multiple pitch activities of a musical piece, we need to analyze the prominent frequency components at any moment of the piece. For this purpose, we use the constant Q transform [31]. We use X_{df} to denote the amplitude of the constant Q wavelet spectrogram obtained from the input signal, d to denote the time frame index, and f to denote the log-frequency bin index. The log-frequency scale is defined, for the sake of simplicity, as

$$f_{\log} = 1200(\log_2 f_{\text{linear}} - \log_2 440) + 5700. \tag{1}$$

A common way of analyzing a spectrogram is to represent each time frame spectrum as a linear combination of K basis spectra:

$$X_{df} \approx \sum_{k=1}^{K} U_{dk} H_{fk}, \tag{2}$$

where U_{dk} represents the mixing coefficients and H_{fk} represents the spectrum of the k-th basis. This idea has been used quite extensively in music analysis [5], [6], [7], [10], [11], [12], [21], [28], [29]. In many multipitch analysis methods, H_{fk} is further decomposed into a series of harmonic component spectra:

$$H_{fk} = \sum_{m=1}^{M} \tau_{km} H_{km}(x_f), \tag{3}$$

where τ_{km} represents the relative weight of m-th overtone component, H_{km} represents the energy distribution function of the component over the log-frequency axis, and x_f denotes the log-frequency of the f-th frequency bin. To simplify the discussion, τ_{km} and H_{km} are assumed to satisfy $\sum_m \tau_{km} = 1$ and $\int H_{km}(x)dx = 1$. M is used to denote the number of harmonic components considered in the model.

There are several ways of modeling the shape of harmonic components. They include using a normal distribution [5], [6], [7], using a sinc function [11], and using a nonparametric [*1] spectrum with binary mask [32]. We use a normal distribution:

$$H_{km}(x) = \mathcal{N}(x|\mu_k + o_m, \lambda_k^{-1}), \tag{4}$$

$$o_m = 1200 \log_2 m, \tag{5}$$

where $\mathcal{N}$ denotes a normal distribution, μ_k denotes the fundamental frequency, λ_k denotes the precision of the distribution, and o_m denotes the relative position of the m-th overtone component on the log-frequency axis. This kind of spectrum modeling using a normal distribution was developed by Goto and termed "predominant-F0 estimation (PreFEst)" [5]. It was followed by harmonic temporal clustering (HTC) [6] and latent harmonic allocation (LHA) [7].

Generally, only the wavelet spectrogram X_{df} is observed; all the other parameters are estimated. They include U_{dk}, μ_k, λ_k, and τ_{km}. The most difficult part of such methods is estimating the overtone structure τ_{km}. We will discuss this problem in detail in the next section. We categorize these methods as "harmonic clustering."

2.2 Previous Methods

The quality of multipitch estimation depends on the accuracy of the harmonic structure estimation. This is illustrated by the following situation. Imagine that the estimation result for a basis function is $\mu_k = 440$ Hz while $\tau_k = [\tau_{k1}, \cdots, \tau_{kM}] = [0, 0, 1, 0, \cdots, 0]$. The estimation result is not reliable because the parameter is unrealistic. When we hear a harmonic sound with that parameter, it is heard as 1,320 Hz, that is, the frequency of only a salient component. There is thus a difference between the estimation result and our recognition.

This conflict occurs because we have not restricted the relative overtone weight τ_k so that the parameter does not become an unrealistic value. To obtain a more accurate result, we should reflect prior knowledge about the harmonic structure in the estimation method. Two main techniques have been proposed for doing this.

The first one uses an optimized conjugate prior distribution of the relative overtone weights, so that it prefers typical harmonic structures. PreFEst and HTC use this technique. It works well to a certain extent, but further development is difficult because the appropriate conjugate prior distribution cannot be determined automatically nor universally. Since a Bayesian framework does not provide a statistically meaningful way of training hyperparameters, they must be optimized manually or be updated using

[*1] Here, the term *nonparametric* does not mean using an infinite mixture of components as in Dirichlet Process.

© 2013 Information Processing Society of Japan

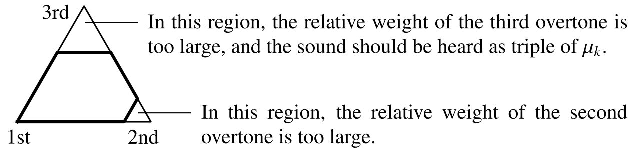
3rd
In this region, the relative weight of the third overtone is too large, and the sound should be heard as triple of μ_k.
In this region, the relative weight of the second overtone is too large.
1st
2nd

Fig. 1 Illustration of appropriate and inappropriate regions of a harmonic sound. Only the first three components are shown due to high-dimensionality.

a costly method like cross-validation. Moreover, the universality problem is also difficult to solve. A good prior distribution should reflect the combination of the musical instruments used to play the target piece and the distribution of their pitch. Since the distribution varies significantly from piece to piece, their universal distribution is unlikely to be considered. Although using infinite LHA [7] is a possible approach to this problem, its methodology is indirect because it tries to solve the problem of auditory recognition by using a purely mathematical approach. As a result, discussion and further development of the method is difficult.

Vincent et al. [11] proposed another approach to this problem. They focused on the fact that the perceived fundamental frequency is the greatest common divisor of the overtone component frequencies. To make the model fundamental frequency and the perceived one equal, they force each harmonic component in the model to always appear with its adjacent harmonic components. This modification to harmonic clustering also works well to a certain extent, but further extension is difficult because the technique does not always guarantee correspondence between the model fundamental frequency and the perceived one.

Given these considerations, we developed a method that explicitly forces the model fundamental frequency and the perceived one, also known as *virtual pitch* [30], to correspond.

2.3 Overtone Corpus

In contrast to previous methods, our overtone corpus (OC) method is a more direct method that enforces desirable behavior in harmonic structure estimation. As illustrated in **Fig. 1**, each overtone weight vector τ_k is represented as a point on an $(M-1)$-simplex. There are some inappropriate regions where the weight of the upper harmonic component is too large. As a result, the perceived fundamental frequency is a multiple of μ_k.

Our method avoids this situation by restricting the overtone weight to one existing in a convex hull, which excludes inappropriate overtone structures. The vertices of the convex hull are determined by a collection of single notes of musical instruments. The reference signals are recorded using a musical instrument digital interface (MIDI) synthesizer. Each harmonic structure is represented as a nonnegative linear combination of J templates, where J is the number of vertices. **Figure 2** illustrates this method. Let $\tau^0_j = [\tau^0_{j1}, \cdots, \tau^0_{jM}]$ be the j-th harmonic template and $\eta_k = [\eta_{k1}, \cdots, \eta_{kJ}]$ be the mixing coefficients of the templates. We model the probabilistic energy distribution function of k-th harmonic sound as

$$p_k(x|\eta_k, \mu_k, \lambda_k) = \sum_{j=1}^{J} \eta_{kj} \sum_{m=1}^{M} \tau^0_{jm} \mathcal{N}(x|\mu_k + o_m, \lambda_k^{-1}). \qquad (6)$$

Among the parameters, τ^0_{jm} are calculated in advance and are not updated during estimation. The total weight of each overtone is represented as

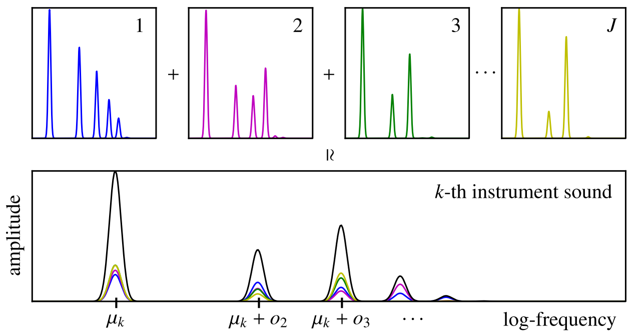

Fig. 2 Illustration of proposed overtone corpus method. Each harmonic structure is designed as a summation of J reference harmonic structures.

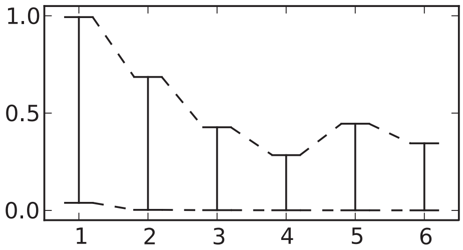

Fig. 3 Upper and lower bounds of relative weight of harmonic partials of reference signals.

$$\tau_{km} = \sum_{j=1}^{J} \eta_{kj} \tau^0_{jm}. \qquad (7)$$

Due to the characteristics of a convex hull, any of its interior points can be represented by this method, and the other points cannot be represented. As a result, the harmonic structure is forced to be appropriate.

This model introduces upper and lower bounds on each overtone weight. Let $\tau_m^{(\min)}$ be the smallest value of the m-th component weight among J templates and $\tau_m^{(\max)}$ be the largest one:

$$\tau_m^{(\min)} = \min_j \tau^0_{jm}, \qquad \tau_m^{(\max)} = \max_j \tau^0_{jm}. \qquad (8)$$

It is obvious that the value τ_{km} is between $\tau_m^{(\min)}$ and $\tau_m^{(\max)}$ because of the non-negativity of the coefficients. The limits of the component weights obtained in our experiments are shown in **Fig. 3**.

This method of restricting the mixing coefficients to ones in a precalculated convex hull was previously and independently proposed as latent variable decomposition [33]. Our contribution is to set appropriate criteria for determining the convex hull that is supported by the psychoacoustic evidence of pitch perception.

The most costly part of the computation is the calculation of the responsibility, the expected value of a latent variable, which will be described in the next section. The total computation time is quite large and nearly proportional to the number of template vectors. The number should be reduced without degrading performance substantially. To do this, we focus on the fact that any interior point of the convex can also be represented as a weighted average of its vertices. This reduction can be strictly done by using multidimensional Delaunay triangulation, but this does not guarantee a reduction in the number of points. Therefore, we use

© 2013 Information Processing Society of Japan

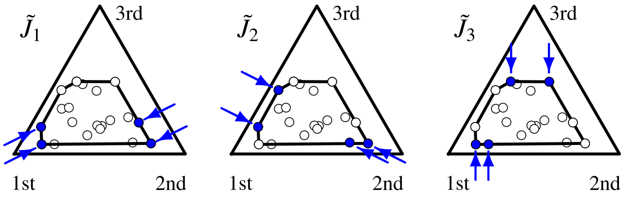

Fig. 4 Illustration of vertex reduction with $J = 20, I = 2$, and $M = 3$. Vertices $\tilde{J}_1, \tilde{J}_2$, and $\tilde{J}_3$ are selected from both ends of each component axis. Solid line indicates convex hull, $\tilde{J}$.

another technique that limits the maximum number of vertices in the reduced set. The reduced set $\tilde{J}$ is obtained using

$$\hat{J}_m = \operatorname*{argsort}_j \tau^0_{jm} = [\hat{j}_{m1}, \cdots, \hat{j}_{mJ}], \tag{9}$$

$$\tilde{J}_m = \bigcup_{i=1}^{I} \left\{\hat{j}_{mi}, \hat{j}_{m,J-i+1}\right\}, \tag{10}$$

$$\tilde{J} = \bigcup_{m=1}^{M} \tilde{J}_m, \tag{11}$$

where I determines the approximation accuracy of the convex and $\tilde{J}_m$ is the set of vector indices that contain I indices at both ends of the m-th axis. This method efficiently reduces the number of vertices to less than $2IM$. The reduction procedure is illustrated in **Fig. 4**.

3. Latent Harmonic Allocation Combined with Overtone Corpus

In this section, we construct a Bayesian framework that combines the original LHA and our overtone corpus. Let D be the number of time frames and F be the number of frequency bins. To fit the observed spectrogram to the variational Bayesian (VB) framework, we interpret spectrogram X_{df} as a histogram of a large number of independently observed particles. Therefore, we assume that the particles in the d-th frame and the f-th frequency bin are observed X_{df} times. To do this, X_{df} is multiplied by a large scaling factor and then quantized to an integer value. In the following, $X = [X_1, \cdots, X_D]$ denotes the set of all observed particles, $X_d = [x_{d1}, \cdots, x_{dN_d}]$ denotes the set of particles in the d-th frame, and x_{dn} denotes the independently observed frequency value. Here, N_d is the number of particles observed in the d-th frame. For each observation x_{dn}, we introduce a latent variable z_{dn}. This is a KJM-dimensional vector. $z_{dnkjm} = 1$ indicates that observation x_{dn} is produced by the m-th overtone of the j-th template of the k-th harmonic sound. Further, α_0, β_0, γ_0, δ_0, m_0, and w_0 denote the hyperparameters of the model. The likelihoods of the proposed model are stated as

$$p(X|Z, \mu, \lambda) = \prod_{dnkjm} \mathcal{N}(x_{dn}|\mu_k + o_m, \lambda_k^{-1})^{z_{dnkjm}}, \tag{12}$$

$$p(Z|\pi, \eta) = \prod_{dnkjm} \left(\pi_{dk}\eta_{kj}\tau^0_{jm}\right)^{z_{dnkjm}}, \tag{13}$$

and the prior probabilities are stated as

$$p(\pi) = \prod_{d=1}^{D} \mathrm{Dir}(\pi_d|\alpha_0) \propto \prod_{d=1}^{D}\prod_{k=1}^{K} \pi_{dk}^{\alpha^0_k - 1}, \tag{14}$$

$$p(\eta) = \prod_{k=1}^{K} \mathrm{Dir}(\eta_k|\beta_0) \propto \prod_{k=1}^{K}\prod_{j=1}^{J} \eta_{kj}^{\beta^0_j - 1}, \tag{15}$$

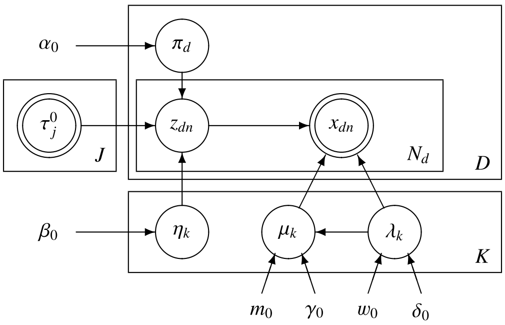

Fig. 5 Graphical model of proposed method. Single solid lines indicate latent variables, and double solid lines indicate observed variables.

$$p(\mu, \lambda) = \prod_{k=1}^{K} \mathcal{N}(\mu_k|m_0, (\gamma_0\lambda_k)^{-1})\mathcal{W}(\lambda_k|w_0, \delta_0), \tag{16}$$

where Dir denotes a Dirichlet distribution and $\mathcal{W}$ denotes a Wishart distribution. **Figure 5** shows a graphical model of our method.

The latent variables of this model are Z, π, η, μ, and λ. The goal of estimation is to obtain their joint posterior distribution, $p(Z, \pi, \eta, \mu, \lambda|X)$. This is an intractable problem, so we use variational Bayesian approximation that decomposes it as $p(Z, \pi, \eta, \mu, \lambda|X) \simeq q(Z)q(\pi, \eta, \mu, \lambda)$. We update $q(Z)$ and $q(\pi, \eta, \mu, \lambda)$ iteratively to search for a local optimum of approximation.

3.1 VB-E Step

In the VB-E step, we calculate $\rho_{dnkjm} = \mathbb{E}[z_{dnkjm}]$ using the temporal estimation of π, η, μ, and λ:

$$\begin{aligned}\log q^*(Z) &= \mathbb{E}_{\pi,\eta,\mu,\lambda}\left[\log p(X, Z, \pi, \eta, \mu, \lambda)\right] + \text{const.}\\ &= \sum_{dnkjm} z_{dnkjm} \log \rho_{dnkjm} + \text{const.},\end{aligned} \tag{17}$$

where ρ_{dnkjm} is calculated as

$$\begin{aligned}\log \tilde{\rho}_{dnkjm} &= \mathbb{E}\left[\log \pi_{dk}\right] + \mathbb{E}\left[\log \eta_{kj}\right] + \log \tau^0_{jm}\\ &\quad + \mathbb{E}\left[\log \mathcal{N}\left(x_{dn}|\mu_k + o_m, \lambda_k^{-1}\right)\right],\end{aligned} \tag{18}$$

$$\rho_{dnkjm} = \frac{\tilde{\rho}_{dnkjm}}{\sum_{kjm} \tilde{\rho}_{dnkjm}}. \tag{19}$$

3.2 VB-M Step

In the VB-M step, we calculate the variational posterior distribution of π, η, μ, and λ. Since all the prior distributions are conjugate, their variational posterior probability $q(\pi, \eta, \mu, \lambda)$ is decomposed as

$$q(\pi, \eta, \mu, \lambda) = \prod_{d=1}^{D} q(\pi_d) \prod_{k=1}^{K} \left\{q(\eta_k)q(\mu_k, \lambda_k)\right\}, \tag{20}$$

where

$$q(\pi_d) = \mathrm{Dir}(\pi_d|\alpha_d), \qquad q(\eta_k) = \mathrm{Dir}(\eta_k|\beta_k), \tag{21}$$

$$q(\mu_k, \lambda_k) = \mathcal{N}(\mu_k|m_k, (\gamma_k\lambda_k)^{-1})\mathcal{W}(\lambda_k|w_k, \delta_k). \tag{22}$$

The decomposed posterior distributions are described with the hyperparameters α_{dk}, β_{kj}, γ_k, δ_k, m_k, and w_k. These values are calculated as

© 2013 Information Processing Society of Japan

Table 1 Parameters of proposed method.

Model parameters	
Symbol	Description
x_f	log-frequency of f-th frequency bin
o_m	relative position of m-th overtone component
Observed and latent variables	
Symbol	Description
x_{dn}	log-frequency of independent observation
z_{dnkjm}	indicator of class allocation
μ_k, λ_k	log-frequency and precision of harmonic components
π_{dk}	relative weight of k-th sound
η_{kj}	relative weight of j-th template
τ^0_{jm}	overtone amplitude ratio of j-th template
Prior and posterior hyperparameters	
Symbol	Description
ρ_{dnkjm}	temporal posterior estimation of z_{dnkjm}
α^0_k, α_{dk}	prior and posterior hyperparameters of π_{dk}
β^0_j, β_{kj}	those of η_{kj}
m_0, m_k	those of μ_k
γ_0, γ_k	those of μ_k
w_0, w_k	those of λ_k
δ_0, δ_k	those of λ_k

$$\alpha_{dk} = \alpha^0_k + N_{dk}, \qquad \beta_{kj} = \beta^0_j + N_{kj}, \tag{23}$$

$$\gamma_k = \gamma_0 + N_k, \qquad \delta_k = \delta_0 + N_k, \tag{24}$$

$$m_k = \frac{\gamma_0 m_0 + \sum_{fm} N_{fkm}(x_f - o_m)}{\gamma_0 + N_k}, \tag{25}$$

$$w_k^{-1} = w_0^{-1} + \gamma_0 m_0^2 + \sum_{fm} N_{fkm}(x_f - o_m)^2 - \gamma_k m_k^2. \tag{26}$$

N_k, N_{dk}, N_{kj}, and N_{fkm} are called sufficient statistics and are calculated as

$$N_k = \sum_{dnjm} \rho_{dnkjm}, \qquad N_{dk} = \sum_{njm} \rho_{dnkjm}, \tag{27}$$

$$N_{kj} = \sum_{dnm} \rho_{dnkjm}, \qquad N_{fkm} = \sum_{dj} \sum_{x_{dn}=x_f} \rho_{dnkjm}, . \tag{28}$$

The main symbols of the model are described in **Table 1**.

3.3 Implementation Issues

Since the calculation of the proposed method is heavy, we optimized implementation. First, we omit the calculation of the responsibility at the tail of Gaussian distribution. To do this, the calculation is performed only in $m_k+o_m-W_k \le x_f \le m_k+o_m+W_k$ for the m-th harmonic partial of the k-th harmonic sound. The width W_k is determined as the maximum of the following two values:

$$W_k = \max(W'_k, 200 \text{ [cents]}), \tag{29}$$

$$W'_k = \frac{3}{\sqrt{\mathbb{E}[\lambda_k]}}. \tag{30}$$

The calculation of LHA is performed as its definition because the computation time is relatively short. Second, we do not retain all the values of ρ_{dnkjm} in memory; instead we retain only sufficient statistics. The calculated responsibilities are immediately summed up to the statistics. This optimization greatly reduces the space complexity. Third, the calculation is parallelized using OpenMP [34].

4. Evaluation

To evaluate the robustness of the proposed model, we conducted multipitch estimation experiments using 40 musical pieces with three initialization conditions.

4.1 Corpus Construction

We recorded 80 General MIDI (GM) instrument sounds using a MIDI synthesizer (Roland SD-80). They were recorded at A4 (440 Hz) for one second and then transformed into wavelet spectrograms using Gabor wavelets. Instruments 81 to 128 were omitted for simplicity because most of them are artificial sounds, which would complicate the discussion of pitch validity. Template overtone structures were filtered using the following three criteria.

4.1.1 Harmonicity

First, we chose instruments that contain more than 50% of their energy on the harmonic partials. More specifically, the instruments that satisfied

$$\sum_{d=1}^{D} \sum_{m=1}^{M} \sum_{a_m < x_f < b_m} Y^{(j)}_{df} \ge \sum_{d=1}^{D} \sum_{f=1}^{F} \frac{Y^{(j)}_{df}}{2} \tag{31}$$

were selected. Here,

$$a_m = f^{(\log)}_m - 100 \text{ [cents]}, \tag{32}$$

$$b_m = f^{(\log)}_m + 100 \text{ [cents]}, \tag{33}$$

$Y^{(j)}_{df}$ is the wavelet spectrogram of the j-th instrument sound, and $f^{(\log)}_m$ is the log-frequency of the m-th harmonic partial.

4.1.2 Pitch Validity

Second, we chose instruments with valid pitch [35]. The validity was measured using subharmonic summation [30]. The measurement was done using a pitch salience p_f:

$$p_f = \sum_{d=1}^{D} \sum_{m=1}^{M} (0.84)^m Y^{(j)}_{d,f+g_m}, \tag{34}$$

where g_m is the offset of the m-th harmonic partial. The instruments that did not satisfy 5,650 [cents] $\le \operatorname{argmax} p_f \le$ 5,750 [cents] were omitted. Here, 5,700 cents is the log-frequency of 440 Hz.

4.1.3 Filtering

Next, we integrated the spectrograms over time and each overtone frequency band, $\tilde{f}_m \le x_f < \tilde{f}_{m+1}$, to create the overtone weights:

$$\tilde{f}_m = \left(m - \frac{1}{2}\right) \times f_0, \qquad \tau^0_{jm} \propto \sum_{d=1}^{D} \sum_{\tilde{f}^{(\log)}_m \le x_f < \tilde{f}^{(\log)}_{m+1}} Y^{(j)}_{df}, \tag{35}$$

where f_0 is the fundamental frequency and $\tilde{f}^{(\log)}_m$ is the corresponding log-frequency of $\tilde{f}_m$.

The obtained vertices, τ^0_{jm}, were reduced using the criteria explained in Section 2.3. We set $I = 2$ and then the size of the corpus, J, became 14. All the template candidates and the selected ones are represented in **Fig. 6**. These graphs indicate that the upper harmonic partials generally had less energy than the lower ones. This algorithm is more objective than the prior corpus selection algorithm [36]. Following the procedure, we automatically obtain a convex hull corresponding to appropriate harmonic structures.

© 2013 Information Processing Society of Japan

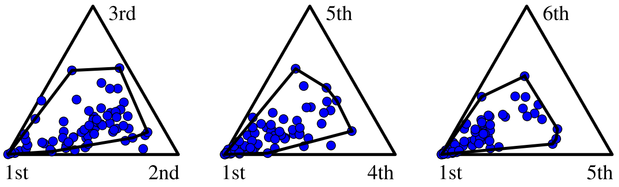

Fig. 6 All overtone template candidates and selected ones. Illustrations show projection of overtone templates to different axes. Solid line indicates convex hull used in multipitch estimation.

4.2 Estimation Target

From the RWC Music Database [37], we used five piano solo pieces (RM-J001 to RM-J005), five guitar solo pieces (RM-J006 to RM-J010), ten jazz duo pieces (RM-J011 to RM-J020), ten jazz pieces played with three or more players (RM-J021 to RM-J030), and ten classical chamber pieces (RM-C012 to RM-C021) to compare the performance of the proposed method with that of LHA. They were recorded using another MIDI synthesizer (Yamaha MOTIF-XS) to generate audio signals. All the drum tracks were muted, and the number of players excluded the drum player. The recorded signals were truncated to the first 32 seconds to reduce the computational time. They were transformed into wavelet spectrograms using Gabor wavelets with a time resolution of 16 [ms], frequency bins from 30 to 3,000 [Hz], and frequency resolution of 12 [cents]. This was done using constant-Q transform [31], and the Q-factor was set to 0.2.

4.3 Experimental Setting

In the experiment, three different initializations of the model were evaluated: random, linear, and exponential. The first initializes the responsibility parameters, and the other two initialize the other parameters to start estimation. For the random initialization, we initialized ρ_{dnkm} or ρ_{dnkjm} by using a uniform distribution. Here, ρ_{dnkm} is the responsibility parameter in the original LHA. This initialization setting tests model stability against initialization because this is substantially the worst case. This is because it uses no prior knowledge about model parameters. For the latter two, the model fundamental frequencies m_k were initialized from 33 Hz (C1) to 2,093 Hz (C7), which reflected the scale of equal temperament. Their standard deviations, $\sigma_k = (w_k \delta_k)^{-1/2}$, were initialized as 50 [cents]. For the linear initialization, the overtone weights were initialized as uniform; for the exponential one, they were initialized as decaying exponentially. The relative weight of each harmonic structure of each time frame, π_{dk}, was initialized proportionally to the sum of the amplitudes of the nearest frequency bins of its overtones. For example, the exponential initialization of LHA was done using

$$\alpha_{dk} \propto \sum_{m=1}^{M} 2^{-m} X_{d f_{km}}, \qquad \beta_{km} \propto 2^{-m}, \tag{36}$$

$$\gamma_k = \sum_{d=1}^{D} \alpha_{dk}, \delta_k = \sum_{d=1}^{D} \alpha_{dk}, w_k^{-1} = \delta_k \left(50\,[\text{cents}]\right)^2. \tag{37}$$

For the initialization of α_{dk} and β_{km}, their total sums were set identical to the number of observation particles to imitate the update equation of the standard VB-M step. Since the proposed method represents the overtone component weight τ_{km} as a summation of other parameters, it is impossible to initialize them directly. Instead, we optimized β_{kj} by using EUC-NMF [38] so that the total overtone weight $\sum_j \beta_{kj} \tau_{jm}$ approximated the desired one. The iteration of EUC-NMF was truncated at 100 iterations.

Table 2 Calculated F-measures: *rand* stands for random initialization, *linear* stands for linear initialization, and *exp* stands for exponential initialization.

	LHA			OC-LHA (proposed)		
Music type	rand	linear	exp	rand	linear	exp
Piano solo	0.339	0.535	0.586	0.563	**0.626**	0.584
Guitar solo	0.137	0.514	**0.745**	0.659	0.736	0.710
Jazz (duo)	0.228	0.532	**0.559**	0.484	0.555	0.542
Jazz (trio~)	0.258	0.478	**0.559**	0.474	0.542	0.531
Chamber	0.247	0.374	0.496	0.464	**0.539**	0.509

For the evaluation, all the prior distributions were set as non-informative. That is, $\alpha_0, \beta_0, \delta_0$, and w_0 were set to unity, γ_0 was set to 10^{-3}, and m_0 was set to zero. Model complexities K, J, and M were 73, 14, and 6, respectively. The number of overtones M was determined in accordance with the setting of HTC [6]. The estimation was truncated at 1,000 iterations for the random initialization, and 100 iterations for the linear and exponential ones. These truncation points were determined experimentally on the basis of estimation accuracy saturation.

After the iterations, we calculated the pitch activity by using the posterior hyperparameters. Let r be the threshold. For each basis of each time frame, those satisfying $N_{dk} \geq r \max_{dk} N_{dk}$ were interpreted as being heard. We tried r between 0 and 1 in steps of 0.01 for each piece, initialization, and method to achieve fair comparison between all settings. Note that this procedure evaluates potential performance, not actual performance, because the optimization of the threshold itself is a problem remaining to be solved. Afterwards, the resultant temporal activity of μ_k was considered as that of the nearest note number.

We used $D \times 128$ binary matrix representation in the performance evaluation. We used the F-measure for comparison, that is, the harmonic mean of precision and recall. Let N be the number of true entries in the estimated matrix, C be the number of entries in the ground truth matrix, and R be the number of correct true entries in the estimated matrix. The F-measure was calculated using $F = 2R/(N + C)$. By definition, $F = 1$ means perfect estimation and $F = 0$ means failure, so the larger the F-measure, the better the performance.

4.4 Results

4.4.1 F-measure

Table 2 shows the overall performance. Since LHA with exponential initialization and the proposed method with linear initialization had similar performances, they both worked properly once the appropriate initialization had been completed. Comparison of the linear and exponential columns shows the proposed method was more robust against initialization. This is because the performance degradation between the optimal and suboptimal settings was smaller with our method for all five musical genres. The initialization sensitivity of LHA appears in the random column because LHA could not estimate the appropriate overtone structure in that case. By contrast, our method was considerably more robust against initialization.

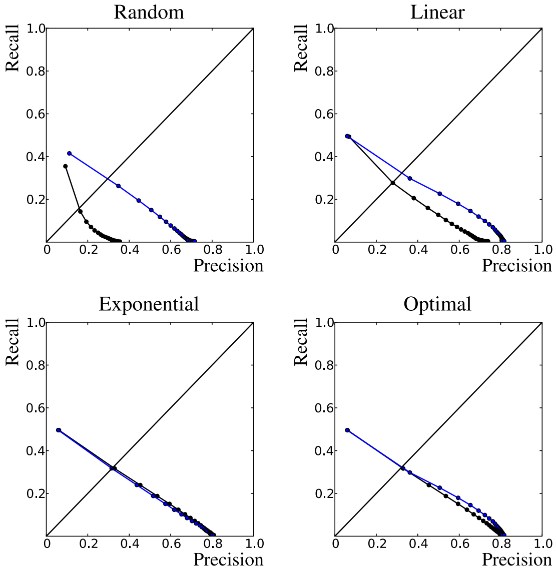

Fig. 7 Precision-recall curve of three initialization settings for previous and proposed method. Black and blue lines show results of previous and proposed methods, respectively. Optimal performance shows result of exponential initialization for previous method and that of linear one for proposed method.

4.4.2 Precision-recall Curve

The precision-recall curves of the previous and proposed methods are shown in **Fig. 7**. The plotted values are the average over the 40 target musical pieces. During plotting, the threshold of estimation r varied from 0 to 0.5. These graphs indicate the estimation stability of the proposed method against the initialization settings.

4.4.3 Example

The ground truth and estimated pitch salience results of the previous and proposed methods obtained using the random initialization are shown in **Fig. 8**. The proposed method captured the pitch salience more accurately. This indicates that the proposed method is able to extract the pitch salience automatically even from random initialization.

Figure 9 shows an example result for overtone weights. LHA tended to estimate the source models using only a second overtone or only a third overtone while the proposed method did not estimate the source models using incorrect overtone weights.

4.5 Effect of Convex Vertex Reduction

We briefly evaluated the effect of the vertex reduction described in Section 2.3. The experiment was limited in scale due to the time complexity.

Figure 10 shows the F-measure calculated for five musical pieces using four different corpus sizes against random initialization. The corpus size (7, 14, 21, and 70) corresponds to approximation accuracy I (1, 2, 3, and 35, respectively). $I \geq 35$ means no filtering since the original corpus size was 70 (Eq. (10)). Three musical pieces slightly favored more precise corpus construction while the result for the other two did not change against the corpus size. Therefore, an increase in the corpus size may slightly improve the overall performance of the method although it increases the computational time.

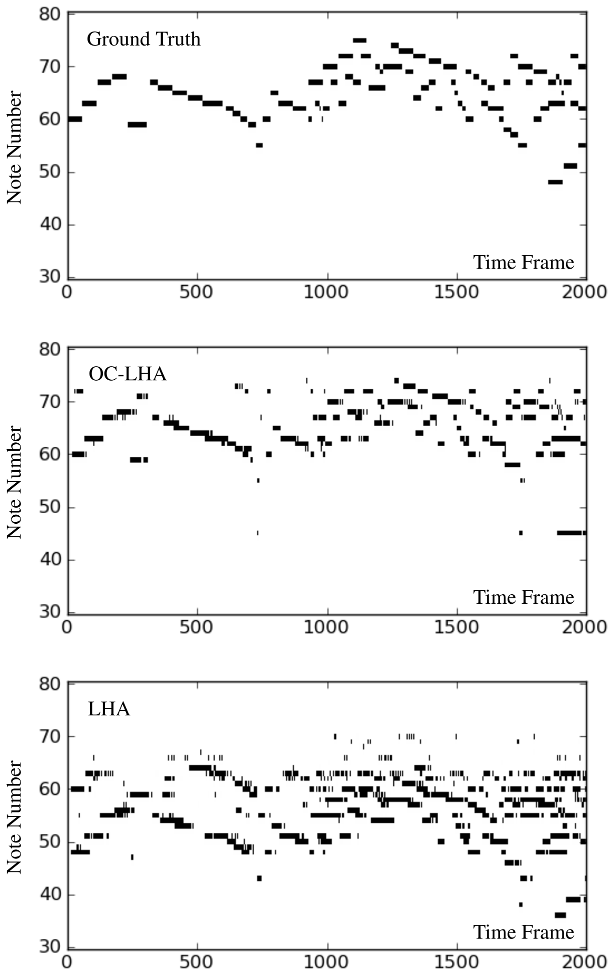

Fig. 8 Ground truth and estimated pitch salience for musical piece *RM-C012*. Illustrations show ground truth (top), result of proposed method with random initialization (middle), and that of LHA (bottom).

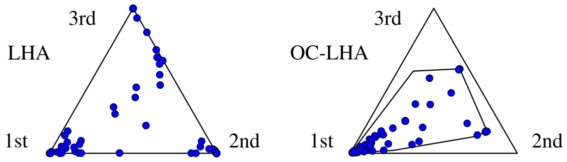

Fig. 9 Estimated component weights of first three harmonic components of 73 sound source models obtained for random initialization with musical piece *RM-C012*. Convex hull projected to 2-simplex is shown as solid lines. Corresponding F-measures are 0.206 for LHA and 0.593 for OC-LHA.

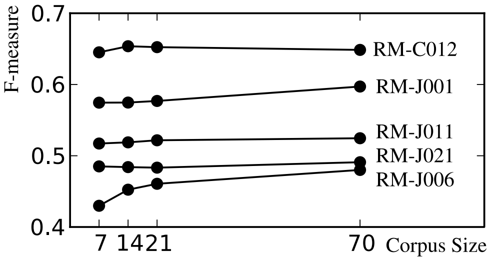

Fig. 10 Number of reduced vertices and performance change. Each solid line indicates performance change between different corpus reduction levels.

5. Discussion

5.1 Properties of Overtone Corpus

For the model construction and evaluation, we made a non-

© 2013 Information Processing Society of Japan

trivial assumption: each interior point of the convex hull corresponds to a valid harmonic structure. This assumption is justified as follows. Suppose we have J valid harmonic structures in the sense that model fundamental frequency and perceived virtual pitch are equal. In this case, any interior point of the convex hull, which is spanned by the template structures, corresponds to a linear combination of template sounds with their sinusoidal component phases being identical. Therefore, the perceived pitch of the sound is almost certainly equal to that of the template sounds. More precisely, there is a convex region corresponding to a valid harmonic structure, and the convex hull is an approximation of it.

Other important properties of this method indicate possible research directions. First, we simply divided the overall region into valid and invalid subsets as if there is a universal division that is true for all people. There should also be a region between the valid and invalid regions, where people cannot perceive a definite pitch. A precise investigation of this point may improve our knowledge of the human auditory system and multipitch analysis performance.

The second point is the effect of audio equalization. In most commercially-available audio files, the harmonic structures are often distorted from the original ones because some frequency bands are reduced or boosted. The estimation of the proposed method fails when the post-processed overtone structure is represented as the external point of the convex hull. Since it is unknown what kind of equalization has been used to the original signal, a possible way of solving the problem is the use of additional prior knowledge from different viewpoints of music signal. For example, audio equalization is generally used to avoid the collision of two or more instrument sounds. Conversely, we may improve the performance of multipitch estimation by detecting the existence of other instruments.

Third, the template and model spectra are averaged over time. This assumption should be relaxed in future research, because the harmonic structures of musical instruments significantly change over time. Since the optimistic modeling of spectrum variation introduces a difficult problem of local optimum, we should develop a perceptually appropriate construction of spectrum variation.

We implemented a multipitch analyzer supported by Terhardt's virtual pitch theory [35]. In addition to his research, the theory of virtual pitch has been studied extensively [30], [35], [39], [40], [41]. In particular, it is reported that the valid pitch region slightly differs from pitch to pitch [30]. The knowledge thereby gained should be applied to machine-learning-based multipitch analysis.

5.2 Relationship with Conventional Method

Our method includes the original LHA as a special case whenever the harmonic templates are set to $\tau_{jm} = \omega_{jm}$, where ω_{jm} is the Dirac delta function. In that case, the update equation is written as

$$\log \tilde{\rho}_{dnkjm} = \begin{cases} \mathbb{E}\left[\log \pi_{dk}\right] + \mathbb{E}\left[\log \eta_{kj}\right] & \\ \quad +\mathbb{E}\left[\log \mathcal{N}\left(x_{dn}|\mu_k + o_m, \lambda_k^{-1}\right)\right] & (j = m) \\ -\infty & (j \neq m). \end{cases} \tag{38}$$

As a result, the update equation for the responsibility is

$$\rho_{dnkjm} = \omega_{jm} \frac{\tilde{\rho}_{dnkjm}}{\sum_{km} \tilde{\rho}_{dnkjm}}, \tag{39}$$

which is equivalent to the one obtained for the conventional method.

6. Conclusion

Our proposed Bayesian method for expressing harmonic structures is based on Terhardt's virtual pitch theory. The proposed method is robust against initialization, optimization-free, and psychoacoustically appropriate so that it is useful for a wide range of further developments of Bayesian multipitch analysis. The appropriate harmonic structure region is automatically learned by using a MIDI synthesizer. Evaluation showed that the proposed method stably estimates the harmonic structure for a wide variety of initial settings. We are planning to apply the obtained robustness characteristics to complex Bayesian estimation frameworks that jointly estimate multiple musical features.

Acknowledgments This research was partially supported by KAKENHI (S) No.24220006 and (B) No.24700168.

References

[1] Piston, W.: *Harmony*, 5th edition, W.W. Norton & Company, Inc. (1987).

[2] Sebesky, D.: *The Contemporary Arranger*, Alfred Music Publishing (1994).

[3] Bregman, A.S.: *Auditory Scene Analysis*, A Bradford Book (1994).

[4] Klapuri, A.: Automatic Music Transcription as We Know it Today, *Journal of New Music Research*, Vol.33, No.3, pp.269–282 (2004).

[5] Goto, M.: 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).

[6] Kameoka, H., Nishimoto, T. and Sagayama, S.: A Multipitch Analyzer Based on Harmonic Temporal Structured Clustering, *IEEE Trans. Audio, Speech, and Language Processing*, Vol.15, No.3, pp.982–994 (2007).

[7] Yoshii, K. and Goto, M.: A Nonparametric Bayesian Multipitch Analyzer based on Infinite Latent Harmonic Allocation, *IEEE Trans. Audio, Speech, and Language Processing*, Vol.20, No.3, pp.717–730 (2012).

[8] Klapuri, A.: Multipitch Analysis of Polyphonic Music and Speech Signals Using an Auditory Model, *IEEE Trans. Audio, Speech, and Language Processing*, Vol.16, No.2, pp.255–266 (2008).

[9] Emiya, V., Badeau, R. and David, B.: Multipitch Estimation of Piano Sounds Using a New Probabilistic Spectral Smoothness Principle, *IEEE Trans. Audio, Speech, and Language Processing*, Vol.18, No.6, pp.1643–1654 (2010).

[10] Smaragdis, P. and Brown, J.C.: Non-Negative Matrix Factorization for Polyphonic Music Transcription, *Proc. WASPAA*, pp.177–180 (2003).

[11] Vincent, E., Bertin, N. and Badeau, R.: Adaptive Harmonic Spectral Decomposition for Multiple Pitch Estimation, *IEEE Trans. Audio, Speech, and Language Processing*, Vol.18, No.3, pp.528–537 (2010).

[12] Itoyama, K., Goto, M., Komatani, K., Ogata, T. and Okuno, H.G.: Simultaneous Processing of Sound Source Separation and Musical Instrument Identification Using Bayesian Spectral Modeling, *Proc. ICASSP*, pp.3816–3819 (2011).

[13] Yasuraoka, N., Abe, T., Itoyama, K., Takahashi, T., Ogata, T. and Okuno, H.G.: Changing Timbre and Phrase in Existing Musical Performances as You Like: Manipulations of Single Part Using Harmonic and Inharmonic Models, *Proc. ACM Multimedia*, pp.203–212 (2009).

[14] Yasuraoka, N., Kameoka, H., Yoshioka, T. and Okuno, H.G.: I-Divergence-Based Dereverberation Method with Auxiliary Function Approach, *Proc. ICASSP*, pp.369–372 (2011).

[15] Itoyama, K., Goto, M., Komatani, K., Ogata, T. and Okuno, H.G.: Query-by-Example Music Information Retrieval by Score-Informed Source Separation and Remixing Technologies, *EURASIP Journal on Advances in Signal Processing*, Vol.2010, pp.1–14 (2010).

[16] Kitahara, T., Goto, M., Komatani, K., Ogata, T. and Okuno, H.G.: Instrument Identification in Polyphonic Music: Feature Weighting to Minimize Influence of Sound Overlaps, *EURASIP Journal on Advances in Signal Processing*, Vol.2007, pp.1–15 (2007).

© 2013 Information Processing Society of Japan

[17] Ueda, Y., Uchiyama, Y., Nishimoto, T., Ono, N. and Sagayama, S.: HMM-Based Approach for Automatic Chord Detection Using Refined Acoustic Features, *Proc. ICASSP*, pp.5518–5521 (2010).

[18] Barbancho, A.M., Klapuri, A., Tardón, L.J. and Barbancho, I.: Automatic Transcription of Guitar Chords and Fingering From Audio, *IEEE Trans. Audio, Speech, and Language Processing*, Vol.20, No.3, pp.915–921 (2012).

[19] Wu, M., Wang, D. and Brown, G.J.: A Multipitch Tracking Algorithm for Noisy Speech, *IEEE Trans. Speech and Audio Processing*, Vol.11, No.3, pp.229–241 (2003).

[20] Nakano, M., Ohishi, Y., Kameoka, H., Mukai, R. and Kashino, K.: Bayesian Nonparametric Music Parser, *Proc. ICASSP*, pp.461–464 (2012).

[21] Miyamoto, K., Kameoka, H., Nishimoto, T., Ono, N. and Sagayama, S.: Harmonic-Temporal-Timbral Clustering (HTTC) for the Analysis of Multi-Instrument Polyphonic Music Signals, *Proc. ICASSP*, pp.113–116 (2008).

[22] Yoshii, K. and Goto, M.: Unsupervised Music Understanding Based on Nonparametric Bayesian Models, *Proc. ICASSP*, pp.5353–5356 (2012).

[23] Winn, J. and Bishop, C.M.: Variational Message Passing, *Journal of Machine Learning Research*, Vol.6, pp.661–694 (2005).

[24] Porteous, I., Asuncion, A., Newman, D., Ihler, A., Smyth, P. and Welling, M.: Fast Collapsed Gibbs Sampling for Latent Dirichlet Allocation, *Proc. ACM SIGKDD*, pp.569–577 (2008).

[25] Teh, Y.W., Kurihara, K. and Welling, M.: Collapsed Variational Inference for HDP, *Proc. NIPS* (2008).

[26] Ueda, N. and Nakano, R.: Deterministic Annealing EM Algorithm, *Neural Networks*, Vol.11, pp.271–282 (1998).

[27] Rose, K.: Deterministic Annealing for Clustering, Compression, Classification, Regression, and Related Optimization Problems, *Proc. IEEE*, Vol.86, No.11, pp.2210–2239 (1998).

[28] Nakano, M., Le Roux, J., Kameoka, H., Nakamura, T., Ono, N. and Sagayama, S.: Bayesian Nonparametric Spectrogram Modeling Based on Infinite Factorial Infinite Hidden Markov Model, *Proc. WASPAA* (2011).

[29] Hoffman, M.D., Blei, D.M. and Cook, P.R.: Bayesian Nonparametric Matrix Factorization for Recorded Music, *Proc. ICML* (2010).

[30] Hermes, D.J.: Measurement of Pitch by Subharmonic Summation, *Journal of the Acoustical Society of America*, Vol.83, No.1, pp.257–264 (1988).

[31] Brown, J.C.: Calculation of a Constant Q Spectral Transform, *Journal of the Acoustical Society of America*, Vol.89, No.1, pp.425–434 (1991).

[32] Raczyński, S.A., Ono, N. and Sagayama, S.: Multipitch Analysis with Harmonic Nonnegative Matrix Approximation, *Proc. ISMIR*, pp.381–386 (2007).

[33] Raj, B. and Smaragdis, P.: Latent Variable Decomposition of Spectrograms for Single Channel Speaker Separation, *Proc. WASPAA*, pp.17–20 (2005).

[34] Mattson, T.G., Sanders, B.A. and Massingill, B.L.: *Patterns for Parallel Programming*, Addison-Wesley Professional, 5th edition (2009).

[35] Terhardt, E.: Pitch, Consonance, and Harmony, *Journal of the Acoustical Society of America*, Vol.55, No.5, pp.1061–1069 (1974).

[36] Sakaue, D., Itoyama, K., Ogata, T. and Okuno, H.G.: Initialization-Robust Multipitch Analyzer Based on Latent Harmonic Allocation Using Overtone Corpus, *Proc. ICASSP*, pp.425–428 (2012).

[37] Goto, M., Hashiguchi, H., Nishimura, T. and Oka, R.: RWC Music Database: Popular, Classical, and Jazz Music Databases, *Proc. ISMIR*, pp.287–288 (2002).

[38] Nakano, M., Kameoka, H., Le Roux, J., Kitano, Y., Ono, N. and Sagayama, S.: Convergence-Guaranteed Multiplicative Algorithms for Nonnegative Matrix Factorization with Beta-Divergence, *Proc. MLSP*, pp.283–288 (2010).

[39] Goldstein, J.L.: An Optimum Processor Theory for the Central Formation of the Pitch of Complex Tones, *Journal of the Acoustical Society of America*, Vol.54, No.6, pp.1496–1516 (1973).

[40] Terhardt, E., Stoll, G. and Seewann, M.: Algorithm for Extraction of Pitch and Pitch Salience from Complex Tonal Signals, *Journal of the Acoustical Society of America*, Vol.71, No.3, pp.679–688 (1982).

[41] Dai, H.: On the Relative Influence of Individual Harmonics on Pitch Judgement, *Journal of the Acoustical Society of America*, Vol.107, No.2, pp.953–959 (2000).

Appendix

The expected values used in this article were calculated using

$$\mathbb{E}\left[\log \pi_{dk}\right] = \psi(\alpha_{dk}) - \psi\left(\sum_{k=1}^{K} \alpha_{dk}\right), \tag{A.1}$$

$$\mathbb{E}\left[\log \eta_{kj}\right] = \psi(\beta_{kj}) - \psi\left(\sum_{j=1}^{J} \beta_{kj}\right), \tag{A.2}$$

$$\mathbb{E}\left[\log \mathcal{N}(x_{dn}|\mu_k + o_m, \lambda_k^{-1})\right] = \frac{\mathbb{E}\left[\log \lambda_k\right] - \log 2\pi - \mathbb{E}\left[\lambda_k(x_{dn} - \mu_k - o_m)^2\right]}{2}, \tag{A.3}$$

$$\mathbb{E}\left[\log \lambda_k\right] = \psi\left(\frac{\delta_k}{2}\right) + \log(2w_k), \tag{A.4}$$

$$\mathbb{E}\left[\lambda_k(x_{dn} - \mu_k - o_m)^2\right] = \beta_k^{-1} + w_k\delta_k(x_{dn} - \mu_k - o_m)^2, \tag{A.5}$$

where ψ is the digamma function.

Daichi Sakaue received his B.E. degree in Science in 2011 from Kyoto University, Japan. He is currently an M.S. candidate in Informatics, Kyoto University. His research interests include music information retrieval, especially multipitch analysis, music structure analysis, and musical sound source separation. He is a member of IPSJ, ASJ, and IEEE.

Katsutoshi Itoyama received his B.E. degree in 2006, M.S. degree in Informatics in 2008, and Ph.D. degree in Informatics in 2011 all from Kyoto University. He is currently an Assistant Professor of the Graduate School of Informatics, Kyoto University, Japan. His research interests include musical sound source separation, music listening interfaces, and music information retrieval. He recieved the 24th TAF Telecom Student Technology Award and the IPSJ Digital Courier Funai Young Researcher Encouragement Award. He is a member of IPSJ, ASJ, and IEEE.

Tetsuya Ogata received his 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. From 2003 to 2012, he was an Associate Professor in the Graduate School of Informatics, Kyoto University. Since 2012, he has been a Professor of the Faculty of Science and Engineering, Waseda University. Since 2009, he has been a JST (Japan Science and Technology Agency) PRESTO Researcher (5 years). His research interests include human-robot interaction, dynamics of human-robot mutual adaptation and inter-sensory translation in robot systems.

Hiroshi G. Okuno received his B.A. and Ph.D. degrees from the University of Tokyo in 1972 and 1996, respectively. He worked for NTT, JST, and Tokyo University of Science. He is currently a Professor of the Graduate School of Informatics, Kyoto University. He was a Visiting Scholar at Stanford University from 1986 to 1988. He is currently engaged in computational auditory scene analysis, music information processing and robot audition. He received various awards including the 1990 Best Paper Award of JSAI, the Best Paper Award of IEA/AIE-2001, 2005 and 2010, IEEE/RSJ IROS-2001 and the 2006 Best Paper Nomination Finalist, and NTF Award for Entertainment Robots and Systems in 2010. He co-edited "Computational Auditory Scene Analysis" (Lawrence Erlbaum Associates, 1998), "Advanced Lisp Technology" (Taylor and Francis, 2002), and "New Trends in Applied Artificial Intelligence (IEA/AIE)" (Springer, 2007). He is an IEEE Fellow and a member of AAAI, ACM, ASJ, and 5 Japanese societies.

© 2013 Information Processing Society of Japan