A Nonparametric Bayesian Multipitch Analyzer Based on Infinite Latent Harmonic Allocation

Kazuyoshi Yoshii, *Member, IEEE*, and Masataka Goto

***Abstract*—The statistical multipitch analyzer described in this paper estimates multiple fundamental frequencies (F0s) in polyphonic music audio signals produced by pitched instruments. It is based on hierarchical nonparametric Bayesian models that can deal with uncertainty of unknown random variables such as model complexities (e.g., the number of F0s and the number of harmonic partials), model parameters (e.g., the values of F0s and the relative weights of harmonic partials), and hyperparameters (i.e., prior knowledge on complexities and parameters). Using these models, we propose a statistical method called infinite latent harmonic allocation (iLHA). To avoid model-complexity control, we allow the observed spectra to contain an unbounded number of sound sources (F0s), each of which is allowed to contain an unbounded number of harmonic partials. More specifically, to model a set of time-sliced spectra, we formulated nested infinite Gaussian mixture models based on hierarchical and generalized Dirichlet processes. To avoid manual tuning of influential hyperparameters, we put noninformative hyperprior distributions on them in a hierarchical manner. For efficient Bayesian inference, we used a modern technique called collapsed variational Bayes. In comparative experiments using audio recordings of piano and guitar solo performances, iLHA yielded promising results and we found that there would be room for improvement based on modeling of temporal continuity and spectral smoothness.**

***Index Terms*—Bayesian nonparametrics, Dirichlet process, infinite latent harmonic allocation (iLHA), multipitch analysis.**

I. Introduction

UNCERTAINTY is inherent in music analysis. A musical piece about which we have little prior knowledge can often be interpreted in various ways. One might, for example, have various degrees of belief in different possible interpretations of tempo and semantic structures, and when we try to transcribe the music we hear in an audio recording, we often find difficult to identify the notes with absolute confidence. Even if in the end we need to determine which interpretation or transcription is the most reasonable, during the analysis it is important to keep all possibilities open with various degrees of belief. We should therefore take an approach that can evaluate, propagate, and integrate the uncertainties of interdependent musical elements or musical notes.

A natural way to manage uncertainty is to take a Bayesian approach and use Bayesian probabilities to indicate degrees of belief. For example, suppose we have a distorted die. If the probabilities of getting the numbers $1, 2, 3, \cdots, 6$ (called *parameters*) are known, we can evaluate the likelihood for a set of numbers (called *observed data*) obtained by casting the die many times. Note that the true values of the parameters do not vary stochastically. When the parameters are unknown, a probabilistic distribution is used as a means of representing how strongly possible values are believed to be the true values. Such degrees of belief vary according to the amount of observed data. Before we get observed data, prior distributions tend to be widely spread. The more data we get, the sharper the peaks of posterior distributions become. That is, the degree of belief on a certain possibility increases. The objective of Bayesian inference is to calculate posterior distributions of unknown variables by formulating probabilistic models defined by likelihood functions and prior distributions.

A critical problem in the conventional Bayesian approach is that we have to specify the complexity of the probabilistic models in advance (*complexity* means the number of mixtures in Gaussian mixture models (GMMs) and the number of states in hidden Markov models (HMMs)). If model complexities are unknown, both the uncertainty of model complexities and that of model parameters should be dealt with appropriately. The conventional approach, however, forces us to train many models of different complexities independently and then select one according to some criteria. Such fine-comb model selection, or model-complexity control, is often impractical, especially in the optimization of combinatorial-complexity models.

A nonparametric Bayesian approach avoiding the model selection problem has recently attracted a lot of attention [1]. Here the term “nonparametric” means that the size of a parameter space (complexity) is not fixed and in theory an infinite number of parameters (infinite complexity) are considered. If an infinite amount of observed data were available, an infinite number of parameters would be needed to represent variety of the data. Actually, however, only a limited number of parameters are needed because the amount of observed data is limited. The effective complexities of nonparametric models can be automatically adjusted according to observed data. Such nonparametric models are essentially different from conventional parametric models. In a single nonparametric model, an infinite number of parametric models with different complexities are stochastically overlapped.

In this paper, we propose a nonparametric Bayesian method for multipitch analysis, which is the basis of music transcription and music information retrieval (MIR). The method is called infinite latent harmonic allocation (iLHA), and our goal is to esti-

Manuscript received November 19, 2010; revised March 29, 2011; accepted July 26, 2011. Date of publication August 15, 2011; date of current version January 11, 2012. This work was supported in part by CREST, JST, and KAKENHI 20800084. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Daniel Ellis.

The authors are with the National Institute of Advanced Industrial Science and Technology (AIST), Tsukuba 305-8568, Japan (e-mail: k.yoshii@aist.go.jp; m.goto@aist.go.jp).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASL.2011.2164530

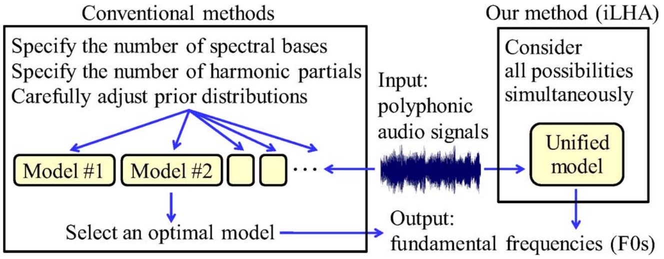

Fig. 1. Advantage of our method: We are not required to specify the number of spectral bases and the number of harmonic partials in advance. In addition, we do not have to adjust hyperparameters carefully.

mate multiple fundamental frequencies (F0s) from polyphonic audio signals. Instead of determining the values of F0s (parameters) and the number of them (complexity) uniquely, our method estimates a joint posterior distribution of all unknown variables when amplitude spectra of musical audio signals are given as observed data. We formulate nested infinite GMMs for observed spectra by using nonparametric priors called Dirichlet processes (DPs). These models can be obtained by taking the limit of the nested finite GMMs proposed by Goto [2] and Kameoka *et al.* [3] as the number of mixtures goes to infinity. More specifically, each spectral strip is allowed to contain an unbounded number of sound sources (harmonic structures), each of which is allowed to contain an unbounded number of harmonic partials. An important problem is that the parameters of the DPs (called *hyperparameters*) should be given appropriately because they affect the effective number of mixtures.

To avoid hyperparameter tuning, our models are formulated in a hierarchical Bayesian manner by putting prior distributions (called *hyperprior distributions*) on influential hyperparameters. Conventionally, we need to specify the hyperparameters of Dirichlet prior distributions on the relative weights of harmonic partials [2], [3]. Although these hyperparameters strongly impact the accuracy of F0 estimation, it is difficult to optimize them by hand. We instead put noninformative hyperprior distributions on the hyperparameters of DP priors of the infinite number of F0s and harmonic partials. This is reasonable because we have little knowledge of the hyperparameters. As shown in Fig. 1, we can completely automate iLHA by leveraging natural Bayesian treatment of parameters, complexities, and influential hyperparameters.

The reminder is organized as follows. Section II introduces related work. Section III compares parametric models of conventional methods and nonparametric models of our method. Section IV describes a finite version of our method (LHA) and Section V explains our method (iLHA). Section VI reports our experiments. Section VII concludes this paper.

II. Related Work

Many researchers have applied probabilistic models to multipitch analysis. Goto [2] proposed a probabilistic model for a single-frame amplitude spectrum (spectral strip) that contains multiple harmonic structures (see Section III) and used it to estimate the F0s of melody and bass lines from polyphonic audio signals. Kameoka *et al.* [3] estimated multiple F0s by using a similar model for grouping frequency components into multiple sound sources. Kameoka *et al.* [4] extended the model by capturing the temporal continuity of harmonic structures. Raphael [5] formulated a HMM based on a large number of chord hypotheses. Cemgil *et al.* [6] used a dynamic Bayesian network (DBN) to represent the sound generation process, i.e., to associate a music-score level with an audio-signal level. Raczyński *et al.* [7] also used a DBN to model temporal dependencies between musical notes. Emiya *et al.* [8] proposed a probabilistic model that jointly represents spectral envelopes and harmonic partials.

Recently, nonnegative matrix factorization (NMF) [9] has been considered to be promising. It regards time–frequency spectra as a nonnegative matrix and decomposes it into the product of two nonnegative matrices, one corresponding to a set of spectral bases and the other corresponding to a set of temporal activations. Smaragdis *et al.* [10] pioneered the use of NMF for music transcription. Virtanen *et al.* [11] and Peeling *et al.* [12] proposed Bayesian extensions of NMF. Raczyński *et al.* [13] and FitzGerald *et al.* [14] proposed harmonicity constraints for spectral bases, and Bertin *et al.* [15] further introduced smoothness constraints for temporal activations. Vincent *et al.* [16] proposed a method of training spectral bases from audio signals of isolated tones and adapting them to target polyphonic audio signals. Cont [17] developed NMF with sparsity constraints for real-time pitch tracking. Several variants of NMF—such as the complex NMF proposed by Kameoka *et al.* [18], the Itakura–Saito (IS) divergence NMF proposed by Févotte *et al.* [19], and the gamma process NMF proposed by Hoffman *et al.* [20]—have been applied to spectrogram decomposition, but F0s have not been estimated from the spectral bases thus obtained.

Many other approaches have been also proposed (see [21] for a review). For example, Marolt [22] and Klapuri [23] proposed auditory-model-based methods that use a peripheral hearing model. Computationally efficient approaches based on harmonic sums [24] and correlograms [25] have also been investigated. Pertusa and Iñesta [26] proposed a spectral-peak clustering method. Bello *et al.* [27] tackled grouping of frequency components by using a heuristic set of rules.

There have been attempt to estimate F0 contours of melody lines (vocal parts) from polyphonic audio signals. Dressler [28] used instantaneous frequency estimation, sinusoidal extraction, psychoacoustics, and auditory stream segregation. Ryynänen and Klapuri [29] formulated a HMM based on acoustic and musicological modeling, and Durrieu *et al.* [30] proposed a statistical method of extracting the main melody by using source/filter models. Poliner *et al.* [31] have reported a comparative evaluation of several approaches.

Most methods mentioned above can achieve good results if the number of sound sources and/or manual parameters are appropriately specified. However, it is difficult to always bring out the full potential of these methods in practice.

III. Probabilistic Models

Our method is based on nonparametric Bayesian extension of conventional finite mixture models proposed by Goto [2] and Kameoka *et al.* [3]. Here we explain the conventional models for observed spectra and then derive our infinite mixture models by extending the conventional models.

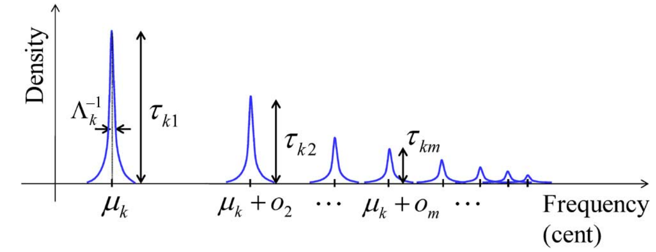

Fig. 2. Gaussian mixture model for the kth basis (single basis). Each Gaussian corresponds to a harmonic partial, and the mixing weights represent the relative strengths of M harmonic partials.

A. Notations

Suppose that given polyphonic audio signals contain K bases, each of which consists of M harmonic partials located at integral multiples of the F0 on a linear frequency scale. Each *basis* can be associated with multiple *sounds* of different temporal positions if these sounds are derived from the same pitch of the same instrument. We transform the audio signals into wavelet spectra. Let D be the number of frames. Note that K and M are finite integers that in conventional methods are specified in advance. Our method considers that K and M go to infinity.

B. Conventional Finite Models and MAP Estimation

Probabilistic models can evaluate how likely observed data is to be generated by using a limited number of parameters. Therefore, estimation of multiple F0s corresponds directly to finding model parameters that give the highest probability to the generation of the observed data (called *model training*).

Goto [2] first proposed probabilistic models of harmonic structures by regarding an amplitude spectrum (a spectral strip of a single frame) as a probability density function. As shown in Fig. 2, the amplitude distribution of basis $k (1 \leq k \leq K)$ can be modeled by a harmonic GMM as follows:

$$\mathcal{M}_k(\boldsymbol{x}) = \sum_{m=1}^{M} \tau_{km} \mathcal{N}\left(\boldsymbol{x} \middle| \boldsymbol{\mu}_k + \boldsymbol{o}_m, \boldsymbol{\Lambda}_k^{-1}\right) \tag{1}$$

where $\boldsymbol{x}$ is a one-dimensional vector indicating a logarithmic frequency [cents].[1] The Gaussian parameters (mean $\boldsymbol{\mu}_k$ and variance $\boldsymbol{\Lambda}_k^{-1}$) represent the F0 of basis k and the degree of energy spread around the F0. τ_{km} is the relative strength of the mth harmonic partial $(1 \leq m \leq M)$ in basis k. We set $\boldsymbol{o}_m$ to $[1200 \log_2 m]$. This means that M Gaussians are located to have the harmonic relationship on the logarithmic frequency scale. One might think that the value of $\boldsymbol{\Lambda}_k^{-1}$ can be precomputed because the basis sound consists of M sinusoidal signals (see Appendix I in [4]). This is true if these sinusoidal signals are stationary, but frequency-modulated sounds (e.g., vibrato) result in a larger value of $\boldsymbol{\Lambda}_k^{-1}$ because of the uncertainty principle of time–frequency resolution.

As shown in Fig. 3, the spectral strip of frame d is modeled by mixing K harmonic GMMs as follows:

$$\mathcal{M}_d(\boldsymbol{x}) = \sum_{k=1}^{K} \pi_{dk} \mathcal{M}_k(\boldsymbol{x}) \tag{2}$$

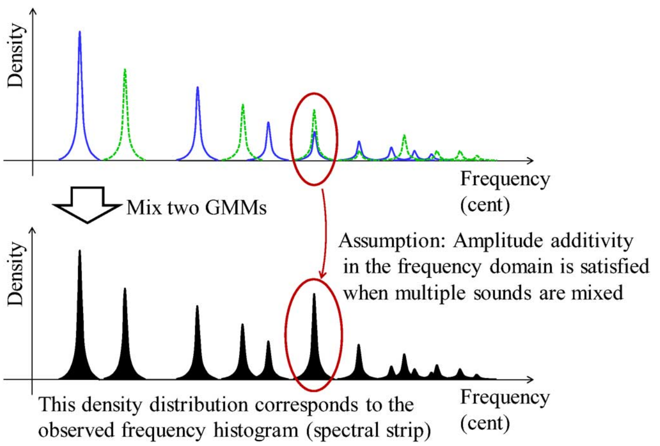
Density
Frequency (cent)
Mix two GMMs
Assumption: Amplitude additivity in the frequency domain is satisfied when multiple sounds are mixed
Density
This density distribution corresponds to the observed frequency histogram (spectral strip)
Frequency (cent)

Fig. 3. Nested Gaussian mixture model for mixed multiple bases. It is obtained by mixing multiple Gaussian mixture models in a weighted manner under the assumption of amplitude additivity.

TABLE I
MULTIPITCH ANALYSIS METHODS

	N. of bases	N. of partials	Temporal modeling
PreFEst [2]	Fixed	Fixed	None
HC [3]	Selected	Fixed	None
HTC [4]	Fixed	Fixed	Continuous
NMF [10]	Fixed	Not used	Exchangeable
iLHA	∞	∞	Exchangeable

where π_{dk} is a relative strength of basis k in frame d. Consequently, the polyphonic spectral strip can be represented by means of a nested finite GMM.

Several methods that have been proposed for parameter estimation are listed in Table I. Goto [2] proposed a predominant-F0 estimation method (PreFEst) that estimates only relative strengths $\boldsymbol{\tau}$ and $\boldsymbol{\pi}$ by allocating many GMMs ($\boldsymbol{\mu}$ and $\boldsymbol{\Lambda}$ are fixed) to cover the entire frequency range as F0 candidates. Kameoka *et al.* [3] proposed harmonic clustering (HC), which estimates all the parameters and selects the optimal number of bases by using the Bayesian information criterion. Although these methods yielded the promising results, they analyze the spectral strips of different frames independently. Kameoka *et al.* [4] therefore proposed harmonic-temporal-structured clustering (HTC), which captures the temporal continuity of spectral bases. Because all the above methods use a maximum *a posteriori* (MAP) estimation strategy to train the finite models, a prior distribution of relative strengths $\boldsymbol{\tau}$ has a large effect on the accuracy of F0 estimation.

C. Our Infinite Models and Bayesian Inference

We would like to discuss the limit of (1) and (2) as K and M diverge to infinity. There is a reason that taking the infinite limit is reasonable even though there are a finite number of discrete pitches (e.g., the standard piano has 88 keys). The F0s and spectral shapes of many instruments (strings, woodwinds, brasses, etc.) vary infinitely according to playing styles (vibrato, marcato, legato, staccato, etc.), and it is difficult to capture these variations when using a parametric model of fixed complexity.

Although there are theoretically infinite number of mixing weights $\{\pi_{d1}, \pi_{d2}, \ldots, \pi_{d_\infty}\}$ and $\{\tau_{k1}, \tau_{k2}, \ldots, \tau_{k_\infty}\}$, in the

[1] Linear frequency f_h in hertz can be converted to logarithmic frequency f_c in cents as follows: $f_c = 1200 \log_2(f_h/(440(2^{(3/12)-5})))$.

finite amount of observed data in practice there are a finite number of bases and a finite number of harmonic partials. Most of mixing weights must therefore be almost equal to zero. In other words, only a limited number of bases and a limited number of harmonic partials are allowed to become active. To realize such "sparse" GMMs, we put nonparametric prior distributions on mixing weights as sparsity constraints. We developed a method of Bayesian inference called iLHA to train the nested infinite GMMs (see Section V).

1) Definition of Observed Data: In the context of Bayesian inference we need to explicitly define the observed data from the statistical viewpoint. More specifically, we regard each spectral strip as a histogram of observed frequencies as in [32]. If a spectral strip at frame $d(1 \leq d \leq D)$ has amplitude a at frequency f, we assume that frequency f was observed $\lfloor \omega a \rfloor$ times in frame d, where ω is a scaling factor of wavelet spectra. In other words, we suppose there are countable frequency "particles" (sound quanta), each corresponding to an independent and identically distributed (i.i.d.) observation. Note that there is a nontrivial issue in determining the value of ω (see Section III-C3). Assuming that amplitudes are additive, we can consider each observation to be generated from one of M partials in one of K bases.

Let the total observations over all D frames be represented by $\boldsymbol{X} = \{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_D\}$, where $\boldsymbol{X}_d$ is a set of observed frequencies $\boldsymbol{X}_d = \{\boldsymbol{x}_{d1}, \ldots, \boldsymbol{x}_{dN_d}\}$ in frame d. N_d is the number of frequency observations (i.e., the sum of spectral amplitudes over all frequency bins in frame d) and $\boldsymbol{x}_{dn}(1 \leq n \leq N_d)$ is a one-dimensional vector that represents an observed frequency. We let $N = \sum_d N_d$ be the total number of observations over all frames.

Let the total latent variables corresponding to $\boldsymbol{X}$ be similarly represented by $\boldsymbol{Z} = \{\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_D\}$, where $\boldsymbol{Z}_d = \{\boldsymbol{z}_{d1}, \ldots, \boldsymbol{z}_{dN_d}\}$. $\boldsymbol{z}_{dn}$ is a KM-dimensional vector in which only one entry, z_{dnkm}, takes a value of 1 and the others take values of 0 when frequency $\boldsymbol{x}_{dn}$ is generated from partial $m(1 \leq m \leq M)$ of basis $k(1 \leq k \leq K)$.

2) Positioning of Our Method: Our method can be viewed as an extension of a well-known topic modeling method called latent Dirichlet allocation (LDA) [33]. LDA was developed as a Bayesian extension of probabilistic latent semantic analysis (pLSA) [34] in the field of natural language processing. In LDA, each document is represented as a weighted mixture of multiple topics that are shared over all documents contained in observed data. Our method similarly represents frames as weighted mixtures of bases. An important difference between our method and LDA, however, is that iLHA represents each basis as a continuous distribution (a GMM) on the frequency space while LDA represents each topic as a discrete distribution over words (a set of unigram probabilities).

Another relevant extension of pLSA is probabilistic latent component analysis (PLCA) [35]. PLCA has been applied to source separation by assuming the time–frequency spectrogram to be a two-dimensional histogram of sound quanta. A major difference between our method and PLCA is that iLHA is based on a continuous distribution on the frequency space at each frame while PLCA is based on a two-dimensional discrete distribution on the space of frame-frequency pairs.

Our method is also similar to the standard NMF [10] based on temporal exchangeability of spectral strips (see Table I). Our method simultaneously trains GMMs of all frames contained in the observed spectra. In other words, if we permute a temporal sequence of spectral strips, the same results would be obtained. Although such temporal modeling is not appropriate for music, it is known to work well in practice.

As discussed above, we fuse the topic modeling framework into the NMF-style decomposition. This is reasonable because any (local) maximum-likelihood solution of pLSA is proven to be a solution of NMF that uses Kullback–Leibler (KL) divergence as a cost function [36]. In addition, we propose a nonparametric Bayesian extension.

3) Limitations of Our Method: The amplitude quantization and i.i.d. assumption are not justified in a physical sense. The amplitudes at the integral multiples of a F0 are correlated to each other when they were generated from a single harmonic sound. Besides this, there is arbitrariness in determining the total number of observations N (the scaling factor ω multiplied to raw wavelet spectra). The larger ω is, the more observations we have, resulting in a more compact posterior distribution because of reduced uncertainty. This criticism can be applied not only to topic models like [32], [35] but also to probabilistic models of NMF with KL divergence. This NMF assumes the value of amplitude to follow a Poisson distribution that is defined over nonnegative integers and has no scale parameter. Note that another NMF with IS divergence [19] does not have such a problem because it assumes the value of power (squared amplitude) to follow an exponential distribution that is defined over nonnegative real numbers and has a scale parameter. Therefore, NMF with IS divergence is scale invariant.

This is more problematic in the context of "nonparametric" Bayesian inference because the larger number of observations allows iLHA to activate more relatively small mixture components (i.e., bases and harmonic partials). We therefore need to perform a thresholding process according to the value of ω after training the weights of bases. In our experiments, the accuracy of multipitch analysis little varied if we changed the value of ω (see Section VI).

Another limitation is that our method represents harmonic sounds in an oversimplified manner. We assume that harmonic sounds consist only of several sinusoidal signals corresponding to harmonic partials. Actually, however, measurable noisy components are widely distributed along the frequency axis even if the target musical pieces are played only by pitched instruments. iLHA is thus forced to use too many harmonic GMMs to represent such noisy components. This is another reason that we need the thresholding process in the end.

IV. Latent Harmonic Allocation

This section explains LHA, the finite version of iLHA, as a preliminary step to deriving iLHA. LHA deals with nested finite GMMs described in Section III in a Bayesian manner. First, we mathematically represent the LHA model by putting prior distributions on unknown variables. Then, we explain a training method of estimating posterior distributions.

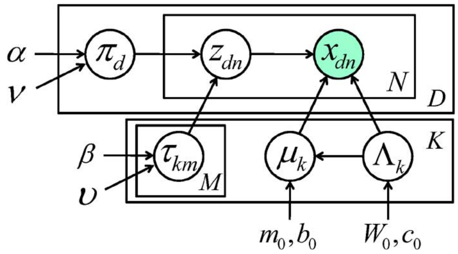

Fig. 4. Graphical representation of nested finite Gaussian mixture models for LHA. First, finite sets of mixing weights, $\boldsymbol{\pi}$ and $\boldsymbol{\tau}$, are stochastically generated according to Dirichlet prior distributions. At the same time, KM Gaussian distributions are stochastically generated according to a Gaussian–Wishart prior distribution. Then one of M harmonic partials in one of K bases is stochastically selected as a latent variable $\boldsymbol{z}_{dn}$ according to multinomial distributions defined by $\boldsymbol{\pi}$ and $\boldsymbol{\tau}$. Finally, frequency $\boldsymbol{x}_{dn}$ is stochastically generated according to a Gaussian distribution specified by $\boldsymbol{z}_{dn}$.

A. Model Formulation

Fig. 4 shows a graphical representation of the LHA model. The full joint distribution is given by

$$p(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\pi}, \boldsymbol{\tau}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) = p(\boldsymbol{X}|\boldsymbol{Z}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) p(\boldsymbol{Z}|\boldsymbol{\pi}, \boldsymbol{\tau}) p(\boldsymbol{\pi}) p(\boldsymbol{\tau}) p(\boldsymbol{\mu}, \boldsymbol{\Lambda}) \tag{3}$$

where the first two terms are likelihood functions and the other three terms are prior distributions. The likelihood functions are defined as

$$p(\boldsymbol{X}|\boldsymbol{Z}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) = \prod_{dnkm} \mathcal{N}\left(\boldsymbol{x}_{dn}|\boldsymbol{\mu}_k + \boldsymbol{o}_m, \boldsymbol{\Lambda}_k^{-1}\right)^{z_{dnkm}} \tag{4}$$

$$p(\boldsymbol{Z}|\boldsymbol{\pi}, \boldsymbol{\tau}) = \prod_{dnkm} (\pi_{dk}\tau_{km})^{z_{dnkm}} \tag{5}$$

Then, we introduce conjugate priors as follows:

$$p(\boldsymbol{\pi}) = \prod_{d=1}^{D} \mathrm{Dir}(\boldsymbol{\pi}_d|\alpha\boldsymbol{\nu}) = \prod_{d=1}^{D} C(\alpha\boldsymbol{\nu}) \prod_{k=1}^{K} \pi_{dk}^{\alpha\nu_k - 1} \tag{6}$$

$$p(\boldsymbol{\tau}) = \prod_{k=1}^{K} \mathrm{Dir}(\boldsymbol{\tau}_k|\beta\boldsymbol{\upsilon}) = \prod_{k=1}^{K} C(\beta\boldsymbol{\upsilon}) \prod_{m=1}^{M} \tau_{km}^{\beta\upsilon_m - 1} \tag{7}$$

$$p(\boldsymbol{\mu}, \boldsymbol{\Lambda}) = \prod_{k=1}^{K} \mathcal{N}\left(\boldsymbol{\mu}_k|\boldsymbol{m}_0, (b_0\boldsymbol{\Lambda}_k)^{-1}\right) \mathcal{W}(\boldsymbol{\Lambda}_k|\boldsymbol{W}_0, c_0) \tag{8}$$

where $p(\boldsymbol{\pi})$ and $p(\boldsymbol{\tau})$ are products of Dirichlet distributions and $p(\boldsymbol{\mu}, \boldsymbol{\Lambda})$ is a product of Gaussian–Wishart distributions. $C(\alpha\boldsymbol{\nu})$ and $C(\beta\boldsymbol{\upsilon})$ are normalization factors, and $\alpha\boldsymbol{\nu}$ and $\beta\boldsymbol{\upsilon}$ are hyperparameters. We let $\boldsymbol{\nu}$ and $\boldsymbol{\upsilon}$ sum to unity, respectively. α and β are often called concentration parameters. $\boldsymbol{m}_0$, b_0, $\boldsymbol{W}_0$, and c_0 are also hyperparameters: $\boldsymbol{m}_0$ is a Gaussian mean, b_0 is a scaling factor of the precision matrix, $\boldsymbol{W}_0$ is a scale matrix, and c_0 is a degree of freedom.

B. Variational Bayesian Inference

The goal of Bayesian inference is to compute a true posterior distribution of all unknown variables: $p(\boldsymbol{Z}, \boldsymbol{\pi}, \boldsymbol{\tau}, \boldsymbol{\mu}, \boldsymbol{\Lambda}|\boldsymbol{X})$. Because analytical calculation of the posterior distribution is intractable, we use an approximation technique, called variational Bayes (VB) [37], that limits the posterior distribution to an analytical form and optimizes it iteratively in a deterministic way. Another possible technique is Markov chain Monte Carlo (MCMC) [38], which sequentially generates samples (the concrete values of unknown variables) from the true posterior distribution in a stochastic way by constructing a Markov chain that has the target distribution as its equilibrium distribution. It is generally difficult, however, to tell whether or not a Markov chain has reached a stationary distribution from which we can get samples within an acceptable error.

In the VB framework, we introduce a variational posterior distribution $q(\boldsymbol{Z}, \boldsymbol{\pi}, \boldsymbol{\tau}, \boldsymbol{\mu}, \boldsymbol{\Lambda})$ and make it close to the true posterior $p(\boldsymbol{Z}, \boldsymbol{\pi}, \boldsymbol{\tau}, \boldsymbol{\mu}, \boldsymbol{\Lambda}|\boldsymbol{X})$ iteratively. Here we assume that the variational distribution can be factorized as

$$q(\boldsymbol{Z}, \boldsymbol{\pi}, \boldsymbol{\tau}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) = q(\boldsymbol{Z}) q(\boldsymbol{\pi}, \boldsymbol{\tau}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) \tag{9}$$

To optimize $q(\boldsymbol{Z}, \boldsymbol{\pi}, \boldsymbol{\tau}, \boldsymbol{\mu}, \boldsymbol{\Lambda})$, we use a variational version of the expectation–maximization (EM) algorithm [37]. We iterate VB-E and VB-M steps alternately until a variational lower bound of evidence $p(\boldsymbol{X})$ converges as follows:

$$q^*(\boldsymbol{Z}) \propto \exp(\mathbb{E}_{\boldsymbol{\pi},\boldsymbol{\tau},\boldsymbol{\mu},\boldsymbol{\Lambda}}[\log p(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\pi}, \boldsymbol{\tau}, \boldsymbol{\mu}, \boldsymbol{\Lambda})]) \tag{10}$$

$$q^*(\boldsymbol{\pi}, \boldsymbol{\tau}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) \propto \exp(\mathbb{E}_{\boldsymbol{Z}}[\log p(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\pi}, \boldsymbol{\tau}, \boldsymbol{\mu}, \boldsymbol{\Lambda})]). \tag{11}$$

C. Variational Posterior Distributions

We derive the formulas for updating variational posterior distributions according to (10) and (11).

1) VB-E Step: An optimal variational posterior distribution of latent variables $\boldsymbol{Z}$ can be computed as follows:

$$\begin{aligned}\log q^*(\boldsymbol{Z}) &= \mathbb{E}_{\boldsymbol{\pi},\boldsymbol{\tau},\boldsymbol{\mu},\boldsymbol{\Lambda}}\left[\log p(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\pi}, \boldsymbol{\tau}, \boldsymbol{\mu}, \boldsymbol{\Lambda})\right] + \text{const.} \\ &= \mathbb{E}_{\boldsymbol{\mu},\boldsymbol{\Lambda}}\left[\log p(\boldsymbol{X}|\boldsymbol{Z}, \boldsymbol{\mu}, \boldsymbol{\Lambda})\right] + \mathbb{E}_{\boldsymbol{\pi},\boldsymbol{\tau}}\left[\log p(\boldsymbol{Z}|\boldsymbol{\pi}, \boldsymbol{\tau})\right] \\ &\quad + \text{const.} \\ &= \sum_{dnkm} z_{dnkm} \log \rho_{dnkm} + \text{const.}\end{aligned} \tag{12}$$

where ρ_{dnkm} is defined as

$$\begin{aligned}\log \rho_{dnkm} &= \mathbb{E}_{\boldsymbol{\pi}_d}[\log \pi_{dk}] + \mathbb{E}_{\boldsymbol{\tau}_k}[\log \tau_{km}] \\ &\quad + \mathbb{E}_{\boldsymbol{\mu}_k,\boldsymbol{\Lambda}_k}\left[\log \mathcal{N}\left(\boldsymbol{x}_{dnm}|\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k^{-1}\right)\right]\end{aligned} \tag{13}$$

where $\boldsymbol{x}_{dnm} = \boldsymbol{x}_{dn} - \boldsymbol{o}_m$. Consequently, $q^*(\boldsymbol{Z})$ is obtained as multinomial distributions given by

$$q^*(\boldsymbol{Z}) = \prod_{dnkm} \gamma_{dnkm}^{z_{dnkm}} \tag{14}$$

where $\gamma_{dnkm} = \rho_{dnkm} / \sum_{km} \rho_{dnkm}$ is called a responsibility that indicates how likely it is that observed frequency $\boldsymbol{x}_{dn}$ is generated from harmonic partial m of basis k. Let n_{dkm} be the number of frequencies that were generated from harmonic partial m of basis k in frame d. This number and its expected value can be calculated as follows:

$$n_{dkm} = \sum_{n} z_{dnkm} \quad \mathbb{E}[n_{dkm}] = \sum_{n} \gamma_{dnkm} \tag{15}$$

For convenience in executing the VB-M step, we calculate several sufficient statistics as follows:

$$\mathbb{S}_k[1] \equiv \sum_{dnm} \gamma_{dnkm} \tag{16}$$

$$\mathbb{S}_k[\boldsymbol{x}] \equiv \sum_{dnm} \gamma_{dnkm}\boldsymbol{x}_{dnm} \tag{17}$$

$$\mathbb{S}_k[\boldsymbol{x}\boldsymbol{x}^T] \equiv \sum_{dnm} \gamma_{dnkm}\boldsymbol{x}_{dnm}\boldsymbol{x}_{dnm}^T. \tag{18}$$

2) VB-M Step: Similarly, an optimal variational posterior distribution of parameters $\boldsymbol{\pi}, \boldsymbol{\tau}, \boldsymbol{\mu}, \boldsymbol{\Lambda}$ is given by

$$\begin{aligned}\log q^*(\boldsymbol{\pi}, \boldsymbol{\tau}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) &= \log p(\boldsymbol{\pi})p(\boldsymbol{\tau}) + \mathbb{E}_{\boldsymbol{z}}\left[\log p(\boldsymbol{Z}|\boldsymbol{\pi}, \boldsymbol{\tau})\right] \\ &\quad + \log p(\boldsymbol{\mu}, \boldsymbol{\Lambda}) + \mathbb{E}_{\boldsymbol{z}}\left[\log p(\boldsymbol{X}|\boldsymbol{Z}, \boldsymbol{\mu}, \boldsymbol{\Lambda})\right] + \text{const.}\end{aligned} \tag{19}$$

This distribution can be factorized into the product of posterior distributions of respective parameters as follows:

$$q^*(\boldsymbol{\pi}, \boldsymbol{\tau}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) = \prod_{d=1}^{D} q^*(\boldsymbol{\pi}_d) \prod_{k=1}^{K} q^*(\boldsymbol{\tau}_k) \prod_{k=1}^{K} q^*(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k) \tag{20}$$

Since our model is based on the conjugate prior distributions, each posterior distribution has the same form of the corresponding prior distribution as follows:

$$q^*(\boldsymbol{\pi}_d) = \mathrm{Dir}(\boldsymbol{\pi}_d|\boldsymbol{\alpha}_d) \tag{21}$$

$$q^*(\boldsymbol{\tau}_k) = \mathrm{Dir}(\boldsymbol{\tau}_k|\boldsymbol{\beta}_k) \tag{22}$$

$$q^*(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k) = \mathcal{N}\left(\boldsymbol{\mu}_k|\boldsymbol{m}_k, (b_k\boldsymbol{\Lambda}_k)^{-1}\right)\mathcal{W}(\boldsymbol{\Lambda}_k|\boldsymbol{W}_k, c_k) \tag{23}$$

where the variational parameters are given by

$$\alpha_{dk} = \alpha\nu_k + \mathbb{E}[n_{dk\cdot}] \tag{24}$$

$$\beta_{km} = \beta\upsilon_m + \mathbb{E}[n_{\cdot km}] \tag{25}$$

$$b_k = b_0 + \mathbb{S}_k[1] \tag{26}$$

$$c_k = c_0 + \mathbb{S}_k[1] \tag{27}$$

$$\boldsymbol{m}_k = \frac{b_0\boldsymbol{m}_0 + \mathbb{S}_k[\boldsymbol{x}]}{b_0 + \mathbb{S}_k[1]} = \frac{b_0\boldsymbol{m}_0 + \mathbb{S}_k[\boldsymbol{x}]}{b_k} \tag{28}$$

$$\boldsymbol{W}_k^{-1} = \boldsymbol{W}_0^{-1} + b_0\boldsymbol{m}_0\boldsymbol{m}_0^T + \mathbb{S}_k[\boldsymbol{x}\boldsymbol{x}^T] - b_k\boldsymbol{m}_k\boldsymbol{m}_k^T \tag{29}$$

where we introduced a dot notation for improved readability. We let dot "$\cdot$" denote the sum over that index. For convenience in the subsequent sections, we also introduce notations using comparison operators ($>$ and $\geq$). For example, we write

$$n_{dk\cdot} = \sum_{m'} n_{dkm'} \quad n_{dk>m} = \sum_{m'>m} n_{dkm'}. \tag{30}$$

The three terms of (13) can therefore be calculated as follows:

$$\mathbb{E}_{\boldsymbol{\pi}_d}[\log \pi_{dk}] = \psi(\alpha_{dk}) - \psi\left(\sum_{k=1}^{K} \alpha_{dk}\right) \tag{31}$$

$$\mathbb{E}_{\boldsymbol{\tau}_k}[\log \tau_{km}] = \psi(\beta_{km}) - \psi\left(\sum_{m=1}^{M} \beta_{km}\right) \tag{32}$$

$$\begin{aligned}&\mathbb{E}_{\boldsymbol{\mu},\boldsymbol{\Lambda}}\left[\log\mathcal{N}\left(\boldsymbol{x}_{dnm}|\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k^{-1}\right)\right] \\ &\quad = -\frac{1}{2}\log(2\pi) + \frac{1}{2}\mathbb{E}_{\boldsymbol{\Lambda}_k}\left[\log|\boldsymbol{\Lambda}_k|\right] \\ &\quad\quad - \frac{1}{2}c_k(\boldsymbol{x}_{dnm} - \boldsymbol{m}_k)^T\boldsymbol{W}_k(\boldsymbol{x}_{dnm} - \boldsymbol{m}_k) - \frac{1}{2b_k}\end{aligned} \tag{33}$$

where ψ is the digamma function, which is defined as the logarithmic derivative of the gamma function.

D. Variational Lower Bound

To judge convergence, we examine the increase of the variational lower bound. Its maximization is inextricably linked with minimization of the KL divergence between the true and variational posteriors. Let $\mathcal{L}$ be the lower bound given by

$$\begin{aligned}\mathcal{L} &= \mathbb{E}\left[\log p(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\pi}, \boldsymbol{\tau}, \boldsymbol{\mu}, \boldsymbol{\Lambda})\right] - \mathbb{E}\left[q(\boldsymbol{Z}, \boldsymbol{\pi}, \boldsymbol{\tau}, \boldsymbol{\mu}, \boldsymbol{\Lambda})\right] \\ &= \mathbb{E}\left[\log p(\boldsymbol{X}|\boldsymbol{Z}, \boldsymbol{\mu}, \boldsymbol{\Lambda})\right] + \mathbb{E}\left[\log p(\boldsymbol{Z}|\boldsymbol{\pi}, \boldsymbol{\tau})\right] - \mathbb{E}\left[\log q(\boldsymbol{Z})\right] \\ &\quad + \mathbb{E}\left[\log p(\boldsymbol{\pi})\right] + \mathbb{E}\left[\log p(\boldsymbol{\tau})\right] + \mathbb{E}\left[\log p(\boldsymbol{\mu}, \boldsymbol{\Lambda})\right] \\ &\quad - \mathbb{E}\left[\log q(\boldsymbol{\pi})\right] - \mathbb{E}\left[\log q(\boldsymbol{\tau})\right] - \mathbb{E}\left[\log q\left(\boldsymbol{\mu}, \boldsymbol{\Lambda}\right))\right].\end{aligned} \tag{34}$$

The calculation of these terms is described in Appendix I.

V. Infinite Latent Harmonic Allocation

Our goal is to formulate and train nested *infinite* GMMs without model selection and hyperparameter tuning. To do this, we consider the limit of the nested *finite* GMMs described in Section IV as both K and M approach infinity. In addition, we put noninformative hyperprior distributions on influential hyperparameters in a hierarchical Bayesian manner and then calculate the posterior distributions of those hyperparameters. As a result of Bayesian inference, likely values are given large posterior probabilistic densities and unlikely values are given small densities. Such *informative* posterior distributions naturally emerge from *noninformative* prior distributions as polyphonic spectra are observed. That is, uncertainty is decreased by getting additional information. In the end we estimate F0s by taking MAP values of the posterior distributions.

A. Mathematical Preparation

We explain the Dirichlet process (DP) and the hierarchical Dirichlet process (HDP), which can be used as nonparametric Bayesian priors in our infinite models. In this section, mathematical symbols are defined according to the custom. Therefore, the definition is valid only in this section.

1) Dirichlet Process: The DP and its extensions play important roles in the theory of Bayesian nonparametrics [39]. Formally introduced by Ferguson [40] in 1973, in the past 10 years it has often been used as a building block of infinite mixture models.

A formal definition of the DP is that its marginal distributions must be Dirichlet distributed [40]. Let α be a positive real number and G_0 be a distribution over a sample space Θ. We say a random distribution G over Θ is DP distributed with concentration parameter α and base measure G_0 if

$$\begin{aligned}&(G(A_1), G(A_2), \ldots, G(A_K)) \\ &\qquad \sim \mathrm{Dir}\left(\alpha G_0(A_1), \alpha G_0(A_2), \ldots, \alpha G_0(A_K)\right)\end{aligned} \tag{35}$$

for any finite measurable partition $\{A_1, A_2, \ldots, A_K\}$ of Θ. The DP is thus a distribution over distributions. This is written

$$G \sim \mathrm{DP}(\alpha, G_0). \tag{36}$$

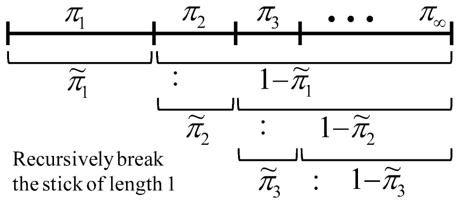
π_1
π_2
π_3
• • •
π_∞
$\tilde{\pi}_1$: $1-\tilde{\pi}_1$
$\tilde{\pi}_2$: $1-\tilde{\pi}_2$
Recursively break
the stick of length 1
$\tilde{\pi}_3$: $1-\tilde{\pi}_3$

Fig. 5. Stick-breaking construction of the Dirichlet process. Starting with a stick of length 1, we break it at $\pi'_1 \sim \mathrm{Beta}(1, \alpha)$ and assign π_1 to be the length of the stick we just broke off. We obtain the infinite number of mixing weights, $\{\pi_2, \pi_3, \ldots, \pi_\infty\}$, by breaking the remaining portion recursively.

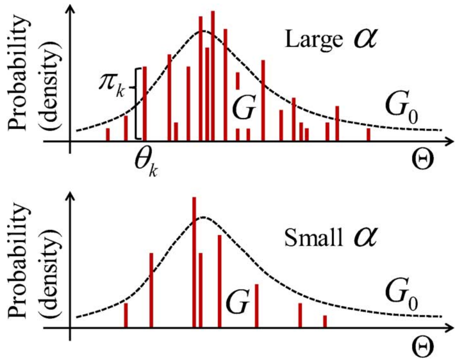

Fig. 6. Discretization property of the Dirichlet process. G becomes an infinite-dimensional discrete distribution when G_0 is a continuous distribution. The smaller α is, the fewer atoms in G occupy most of its total probability mass. This means that G becomes more sparse.

Then, a concrete sample $\theta \in \Theta$ is drawn from G as follows:

$$\theta \sim G. \tag{37}$$

An alternative constructive definition of the DP is known as the stick-breaking construction (SBC) [41]. As illustrated in Fig. 5, a random distribution G can be written explicitly as a *countably* infinite sum of point masses ("atoms"):

$$\theta_k \sim G_0 \tag{38}$$

$$G(\theta) = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k}(\theta) \tag{39}$$

where $\delta_a(x)$ is the Dirac delta function that diverges to positive infinity at $x = a$, is otherwise equal to 0, and integrates to 1 with respect to x. The point mass π_k of θ_k is given by

$$\pi'_k \sim \mathrm{Beta}(1, \alpha) \tag{40}$$

$$\pi_k = \pi'_k \prod_{i=1}^{k-1} (1 - \pi'_i) \quad .(41)$$

The distribution on the infinite number of mixing weights $\boldsymbol{\pi} = \{\pi_1, \pi_2, \ldots, \pi_\infty\}$ is often written $\boldsymbol{\pi} \sim \mathrm{GEM}(\alpha)$, where the letters stand for Griffiths, Engen, and McCloskey.

An important property of the DP is that G must be a discrete distribution. As shown in Fig. 6, G is an infinite-dimensional discrete distribution when G_0 is a continuous distribution. The DP can therefore be used as a prior distribution to formulate an infinite mixture model. In case of an infinite GMM (iGMM), for example, Θ is a space of Gaussians (i.e., a space of means and variances). G_0 is usually set to a Gaussian–Wishart distribution, which is a conjugate prior distribution over Gaussians. G drawn from the DP is also a distribution over Gaussians. Every time an observation is generated, Gaussian $\theta \in \Theta$ is drawn from G, where θ is *selected* from the infinite number of Gaussians $\{\theta_1, \theta_2, \ldots, \theta_\infty\}$ according to their probabilities $\{\pi_1, \pi_2, \ldots, \pi_\infty\}$. This is a straightforward extension of a conventional finite GMM.

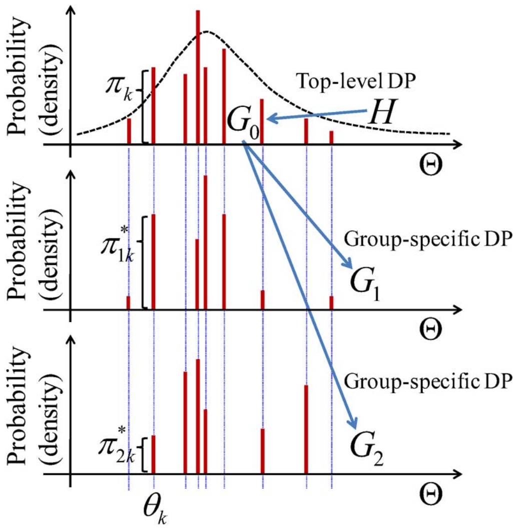

Fig. 7. Overview of hierarchical Dirichlet process. G_0 becomes an infinite-dimensional discrete distribution when H is a continuous distribution. The smaller α is, the fewer atoms in G_n occupy most of its total probability mass. This means that G_n becomes more sparse.

Several extensions increasing a degree of freedom of the standard DP have been proposed. For example, a beta two-parameter process [42] is obtained when

$$\pi'_k \sim \mathrm{Beta}(\alpha, \beta) \tag{42}$$

where positive real numbers α and β are adjustable parameters of the beta distribution.

2) Hierarchical Dirichlet Process: We discuss how to simultaneously train *tied* infinite mixture models when observed data consists of multiple *groups*, e.g., spectral strips (frames). Here a set of component distributions should be shared across mixture models trained for different groups. Such parameter tying enables us to directly compare compositions of different groups in terms of mixing weights of component distributions. This is similar to vector quantization (VQ) [43]. Let N be the number of groups. In this setting, it is natural to use a DP for modeling observed data of each group as follows:

$$G_n \sim \mathrm{DP}(\alpha, G_0) \quad (1 \le n \le N) \tag{43}$$

where G_n is a random distribution on Θ for group n.

A problem is that if G_0 is a continuous distribution, atoms (component distributions) drawn from G_n for generating observations are almost surely disjointed from those drawn from $G_{n'} (n' \neq n)$. This is because N DPs can independently determine the positions of the *countably* infinite number of discrete atoms $\boldsymbol{\theta} = \{\theta_1, \ldots, \theta_\infty\}$ (cardinality $\aleph_0$) from the *uncountably* infinite continuous space Θ (cardinality $\aleph$).

To solve this problem, we use a HDP [44] as a nonparametric prior distribution. As shown in Fig. 7, we consider the base measure G_0 itself to be distributed according to a top-level DP as follows:

$$G_0 \sim \mathrm{DP}(\gamma, H) \tag{44}$$

where γ is a concentration parameter and H a base measure over Θ. In this model, G_0 always becomes a discrete distribution. The SBC of the top-level DP is given by

$$\theta_k \sim H \tag{45}$$

$$G_0(\theta) = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k}(\theta) \tag{46}$$

where $\boldsymbol{\pi} = \{\pi_1, \ldots, \pi_\infty\}$ and $\boldsymbol{\theta} = \{\theta_1, \ldots, \theta_\infty\}$ are the point masses and positions of atoms and we have $\boldsymbol{\pi} \sim \mathrm{GEM}(\gamma)$. Similarly, the SBC of a lower-level DP is given by

$$\theta_{nk} \sim G_0 \tag{47}$$

$$G_n(\theta) = \sum_{k=1}^{\infty} \pi_{nk} \delta_{\theta_{nk}}(\theta) \tag{48}$$

where $\boldsymbol{\pi}_n \sim \mathrm{GEM}(\alpha)$ and each θ_{nk} is selected from $\boldsymbol{\theta}$. Note that θ_{nk} can be equal to $\theta_{nk'}$ if $k \neq k'$ because G_0 is a discrete distribution. Another direct representation based on $\boldsymbol{\theta}$ determined by the top-level DP is as follows:

$$G_n(\theta) = \sum_{k=1}^{\infty} \pi^*_{nk} \delta_{\theta_k}(\theta) \tag{49}$$

where $\boldsymbol{\pi}^*_n \sim \mathrm{DP}(\alpha, \boldsymbol{\pi})$ and the hyperparameter α controls the difference between $\boldsymbol{\pi}$ and $\boldsymbol{\pi}^*$. Therefore, only point masses $\boldsymbol{\pi}^*_n$ (mixture weights) differ between groups while positions $\boldsymbol{\theta}$ (component distributions) are shared across groups.

A remaining problem is how to adjust the influential hyperparameters α and γ. This problem is often solved by putting vague gamma hyperprior distributions on these hyperparameters and inferring the posterior distributions.

B. Model Formulation

We explain how to formulate nested infinite GMMs based on a HDP and generalized DPs by extending the nested finite GMMs described in Section IV.

First we discuss $K \to \infty$. An important requirement is that basis models [harmonic GMMs represented by (1)] should be shared as a global set across all D frames because each basis sound has a duration and may appear in different frames while only its weight varies. The HDP can satisfy this requirement and we can explain the HDP from the generative point of view. After an unbounded number of bases are initially *generated* according to a top-level DP, an unbounded number of bases are *selected* in each frame according to a frame-specific DP. In practice, a limited number of bases are used to represent a spectral strip because the number of observed frequency particles ($n_{d\cdot\cdot}$) is limited. Mathematically speaking, in (6) we consider infinite-dimensional Dirichlet distributions, which are equivalent to the frame-specific DPs, and assume hyperparameter $\boldsymbol{\nu}$ to be distributed according to the top-level DP as follows:

$$\tilde{\nu}_k \sim \mathrm{Beta}(1, \gamma) \tag{50}$$

$$\nu_k = \tilde{\nu}_k \prod_{k'=1}^{k-1} (1 - \tilde{\nu}_{k'}) \tag{51}$$

where γ is a concentration parameter of the top-level DP.

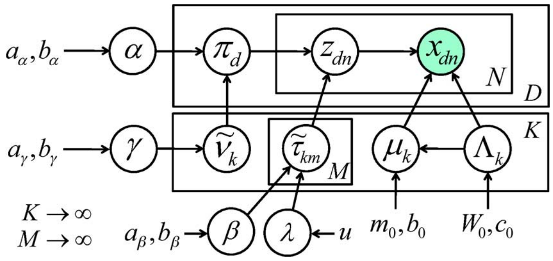

Fig. 8. Graphical representation of nested infinite Gaussian mixture models for iLHA. First the infinite sets of mixing weights $\boldsymbol{\pi}$ and $\boldsymbol{\tau}$ are stochastically generated according to a HDP and beta two-parameter processes (generalized DPs). At the same time, the infinite number of Gaussian distributions are stochastically generated according to a Gaussian–Wishart prior distribution. Then one of the harmonic partials contained in one of the bases is stochastically selected as a latent variable $\boldsymbol{z}_{dn}$ according to multinomial distributions defined by $\boldsymbol{\pi}$ and $\boldsymbol{\tau}$. Finally, frequency $\boldsymbol{x}_{dn}$ is stochastically generated according to a Gaussian distribution specified by $\boldsymbol{z}_{dn}$.

Now we discuss $M \to \infty$. Because each basis is allowed to consist of a unique infinite set of harmonic partials (basis models are independent of each other), instead of (7) we can use beta two-parameter processes as follows:

$$\tilde{\tau}_{km} \sim \mathrm{Beta}(\beta\lambda_1, \beta\lambda_2) \tag{52}$$

$$\tau_{km} = \tilde{\tau}_{km} \prod_{m'=1}^{m-1} (1 - \tilde{\tau}_{km'}) \tag{53}$$

where β is a positive real number and we let λ_1 and λ_2 sum to unity. Note that we used the size-biased permutation property of the SBC to encourage lower harmonic partials to have larger weights because roughly speaking, the weights of harmonic partials of an instrument sound decrease exponentially.

Because hyperparameters α, β, γ, and $\boldsymbol{\lambda}$ are influential, we put hyperprior distributions on them as follows:

$$p(\alpha) = \mathrm{Gam}(\alpha | a_\alpha, b_\alpha) \tag{54}$$

$$p(\gamma) = \mathrm{Gam}(\gamma | a_\gamma, b_\gamma) \tag{55}$$

$$p(\beta) = \mathrm{Gam}(\beta | a_\beta, b_\beta) \tag{56}$$

$$p(\boldsymbol{\lambda}) = \mathrm{Beta}(\lambda_1 | u_1, u_2) \tag{57}$$

where $a_{\{\alpha,\beta,\gamma\}}$ and $b_{\{\alpha,\beta,\gamma\}}$ are shape and rate parameters of the gamma distributions. u_1 and u_2 are parameters of the beta distribution. These distributions are set to be vague ($a_{\{\alpha,\beta,\gamma\}} = 1.0$, $b_{\{\alpha,\beta,\gamma\}} = 0.001$, and $u_1 = u_2 = 1.0$ in our experiments described in Section VI).

Fig. 8 shows a graphical representation of the iLHA model. The full joint distribution is given by

$$p(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\pi}, \tilde{\boldsymbol{\tau}}, \boldsymbol{\mu}, \boldsymbol{\Lambda}, \alpha, \beta, \gamma, \boldsymbol{\lambda}, \tilde{\boldsymbol{\nu}}) = p(\boldsymbol{X}|\boldsymbol{Z}, \boldsymbol{\mu}, \boldsymbol{\Lambda}) p(\boldsymbol{\mu}, \boldsymbol{\Lambda}) \times p(\boldsymbol{Z}|\boldsymbol{\pi}, \tilde{\boldsymbol{\tau}}) p(\boldsymbol{\pi}|\alpha, \tilde{\boldsymbol{\nu}}) p(\tilde{\boldsymbol{\tau}}|\beta, \boldsymbol{\lambda}) p(\alpha) p(\beta) p(\gamma) p(\boldsymbol{\lambda}) p(\tilde{\boldsymbol{\nu}}|\gamma) \tag{58}$$

where $p(\boldsymbol{Z}|\boldsymbol{\pi}, \tilde{\boldsymbol{\tau}})$ is given by plugging (52) into (5) and $p(\boldsymbol{\pi}|\alpha, \tilde{\boldsymbol{\nu}})$ is given by (6). $p(\tilde{\boldsymbol{\nu}}|\gamma)$ and $p(\tilde{\boldsymbol{\tau}}|\beta, \boldsymbol{\lambda})$ are defined according to (50) and (52) as follows:

$$p(\tilde{\boldsymbol{\nu}}|\gamma) = \prod_k \mathrm{Beta}(\tilde{\nu}_k | 1, \gamma) \tag{59}$$

$$p(\tilde{\boldsymbol{\tau}}|\beta, \boldsymbol{\lambda}) = \prod_{km} \mathrm{Beta}(\tilde{\tau}_{km} | \beta\lambda_1, \beta\lambda_2). \tag{60}$$

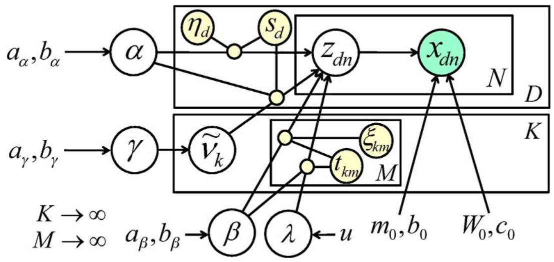

Fig. 9. Graphical representation of collapsed nested infinite mixture models for iLHA. After the original parameters $\boldsymbol{\pi}$, $\tilde{\boldsymbol{\tau}}$, $\boldsymbol{\mu}$, and $\boldsymbol{\Lambda}$ are integrated out, the auxiliary variables $\boldsymbol{\eta}$, $\boldsymbol{\xi}$, $\boldsymbol{s}$, and $\boldsymbol{t}$ are introduced to set up conjugacy between hyperprior distributions and a marginalized likelihood function.

C. Collapsed Variational Bayesian Inference

There are two problems in training the HDP mixture model. The first problem is that VB needs to assume the independence between latent variables and parameters to factorize a posterior distribution as in (9). This assumption is sometimes too strong and leads to incorrect posterior approximation. The second problem is that applying VB to hierarchical Bayesian models that have no conjugacy between priors and hyperpriors is generally difficult.

To solve these problems, we use a sophisticated version of VB called collapsed variational Bayes (CVB) [45]. It instead assumes independence between individual latent variables in a "collapsed" space in which parameters are integrated out (marginalized out). This is reasonable because the dependence between individual latent variables in the collapsed space is generally much weaker than the dependence between a set of parameters and a set of latent variables in the non-collapsed space. In addition, we introduce auxiliary variables to apply CVB to hierarchical Bayesian models.

Fig. 9 shows a graphical representation of a collapsed iLHA model. Integrating out $\boldsymbol{\pi}$, $\tilde{\boldsymbol{\tau}}$, $\boldsymbol{\mu}$, and $\boldsymbol{\Lambda}$, we obtain the marginal distribution given by

$$\begin{aligned} p(\boldsymbol{X}, \boldsymbol{Z}, \alpha, \beta, \gamma, \boldsymbol{\lambda}, \tilde{\boldsymbol{\nu}}) &= p(\boldsymbol{X}|\boldsymbol{Z}) p(\boldsymbol{Z}|\alpha, \beta, \boldsymbol{\lambda}, \tilde{\boldsymbol{\nu}}) \\ &\quad \times p(\alpha) p(\beta) p(\gamma) p(\boldsymbol{\lambda}) p(\tilde{\boldsymbol{\nu}}|\gamma). \end{aligned} \tag{61}$$

The first term of (61) can be easily calculated by leveraging conjugacy between $p(\boldsymbol{X}|\boldsymbol{Z}, \boldsymbol{\mu}, \boldsymbol{\Lambda})$ and $p(\boldsymbol{\mu}, \boldsymbol{\Lambda})$ as follows:

$$p(\boldsymbol{X}|\boldsymbol{Z}) = (2\pi)^{-\frac{n_{\cdot\cdot\cdot}}{2}} \prod_k \left(\frac{b_0}{b_{zk}}\right)^{\frac{1}{2}} \frac{B(\boldsymbol{W}_0, c_0)}{B(\boldsymbol{W}_{zk}, c_{zk})} \tag{62}$$

where $B(\boldsymbol{W}_0, c_0)$ and $B(\boldsymbol{W}_{zk}, c_{zk})$ are normalization factors of prior and posterior Gaussian–Wishart distributions. b_{zk}, c_{zk}, and $\boldsymbol{W}_{zk}$ are obtained by substituting z_{dnkm} for γ_{dnkm} in calculating (26), (27), and (29). Similarly, the second term of (61) can be calculated by leveraging conjugacy between $p(\boldsymbol{Z}|\boldsymbol{\pi}, \tilde{\boldsymbol{\tau}})$ and $p(\boldsymbol{\pi}|\alpha, \tilde{\boldsymbol{\nu}}) p(\tilde{\boldsymbol{\tau}}|\beta, \boldsymbol{\lambda})$ as follows:

$$\begin{aligned} p(\boldsymbol{Z}|\alpha, \beta, \boldsymbol{\lambda}, \tilde{\boldsymbol{\nu}}) &= \prod_d \frac{\Gamma(\alpha)}{\Gamma(\alpha + n_{d\cdot\cdot})} \prod_k \frac{\Gamma(\alpha\nu_k + n_{dk\cdot})}{\Gamma(\alpha\nu_k)} \\ &\quad \times \prod_{km} \frac{\Gamma(\beta)\Gamma(\beta\lambda_1 + n_{\cdot km})\Gamma(\beta\lambda_2 + n_{\cdot k>m})}{\Gamma(\beta\lambda_1)\Gamma(\beta\lambda_2)\Gamma(\beta + n_{\cdot k\geq m})} \end{aligned} \tag{63}$$

where Γ is the gamma function.

We then introduce auxiliary variables by using a technique called data augmentation [45]. Let η_d and ξ_{km} be beta-distributed variables and s_{dk} and $\boldsymbol{t}_{km}$ be positive integers that satisfy $1 \leq s_{dk} \leq n_{dk\cdot}$, $1 \leq t_{km1} \leq n_{\cdot km}$, and $1 \leq t_{km2} \leq n_{\cdot k>m}$. We can augment (63) as follows:

$$\begin{aligned} &p(\boldsymbol{Z}, \boldsymbol{\eta}, \boldsymbol{\xi}, \boldsymbol{s}, \boldsymbol{t}|\alpha, \beta, \boldsymbol{\lambda}, \tilde{\boldsymbol{\nu}}) \\ &\quad = \prod_d \frac{\eta_d^{\alpha-1}(1-\eta_d)^{n_{d\cdot\cdot}-1}}{\Gamma(n_{d\cdot\cdot})} \prod_k \begin{bmatrix} n_{dk\cdot} \\ s_{dk} \end{bmatrix} (\alpha\nu_k)^{s_{dk}} \\ &\qquad \times \prod_{km} \frac{\xi_{km}^{\beta-1}(1-\xi_{km})^{n_{\cdot k\geq m}-1}}{\Gamma(n_{\cdot k\geq m})} \\ &\qquad \times \begin{bmatrix} n_{\cdot km} \\ t_{km1} \end{bmatrix} (\beta\lambda_1)^{t_{km1}} \begin{bmatrix} n_{\cdot k>m} \\ t_{km2} \end{bmatrix} (\beta\lambda_2)^{t_{km2}} \end{aligned} \tag{64}$$

where [] denotes a Stirling number of the first kind. We can confirm that (64) reduces to (63) by marginalizing out auxiliary variables $\boldsymbol{\eta}$, $\boldsymbol{\xi}$, $\boldsymbol{s}$, and $\boldsymbol{t}$. The augmented marginal distribution is given by

$$\begin{aligned} p(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{\eta}, \boldsymbol{\xi}, \boldsymbol{s}, \boldsymbol{t}, \alpha, \beta, \gamma, \boldsymbol{\lambda}, \tilde{\boldsymbol{\nu}}) &= p(\boldsymbol{X}|\boldsymbol{Z}) \\ \times p(\boldsymbol{Z}, \boldsymbol{\eta}, \boldsymbol{\xi}, \boldsymbol{s}, \boldsymbol{t}|\alpha, \beta, \boldsymbol{\lambda}, \tilde{\boldsymbol{\nu}}) &p(\alpha) p(\beta) p(\gamma) p(\boldsymbol{\lambda}) p(\tilde{\boldsymbol{\nu}}|\gamma). \end{aligned} \tag{65}$$

To apply CVB to approximate the true posterior distribution $p(\boldsymbol{Z}, \boldsymbol{\eta}, \boldsymbol{\xi}, \boldsymbol{s}, \boldsymbol{t}, \alpha, \beta, \gamma, \boldsymbol{\lambda}, \tilde{\boldsymbol{\nu}}|\boldsymbol{X})$, we assume that the variational posterior distribution can be factorized as follows:

$$\begin{aligned} q(\boldsymbol{Z}, \boldsymbol{\eta}, \boldsymbol{\xi}, \boldsymbol{s}, \boldsymbol{t}, \alpha, \beta, \gamma, \boldsymbol{\lambda}, \tilde{\boldsymbol{\nu}}) &= q(\alpha, \beta, \gamma, \boldsymbol{\lambda}) \\ &\quad \times q(\tilde{\boldsymbol{\nu}}) q(\boldsymbol{\eta}, \boldsymbol{\xi}, \boldsymbol{s}, \boldsymbol{t}|\boldsymbol{Z}) \prod_{dn} q(\boldsymbol{z}_{dn}) \end{aligned} \tag{66}$$

where we assumed independence between hyperparameters, auxiliary variables, and elements of $\boldsymbol{Z}$. We also use an approximation technique called variational posterior truncation. More specifically, we assume that $q(z_{dnkm}) = 0$ when $k > K^+$ and $m > M^+$. In practice, we set K^+ and M^+ to sufficiently large integers. This does not mean that effective model complexities are fixed in advance. The larger the truncation levels we use, the more the accurate approximations we obtain.

To optimize $q(\boldsymbol{Z}, \boldsymbol{\eta}, \boldsymbol{\xi}, \boldsymbol{s}, \boldsymbol{t}, \alpha, \beta, \gamma, \boldsymbol{\lambda}, \tilde{\boldsymbol{\nu}})$, we use a variational EM algorithm that iterates the following steps:

$$q^*(\boldsymbol{z}_{dn}) \propto \exp\left(\mathbb{E}_{\boldsymbol{Z}^{\neg dn}, \alpha, \beta, \gamma, \boldsymbol{\lambda}, \tilde{\boldsymbol{\nu}}} [\log \text{ Eqn. (61)}]\right) \tag{67}$$

$$q^*(\alpha, \beta, \gamma, \boldsymbol{\lambda}) \propto \exp(\mathbb{E}_{\boldsymbol{Z}, \boldsymbol{\eta}, \boldsymbol{s}, \boldsymbol{\xi}, \boldsymbol{t}, \tilde{\boldsymbol{\nu}}} [\log \text{ Eqn. (65)}]) \tag{68}$$

$$q^*(\tilde{\boldsymbol{\nu}}) \propto \exp(\mathbb{E}_{\boldsymbol{Z}, \boldsymbol{\eta}, \boldsymbol{s}, \boldsymbol{\xi}, \boldsymbol{t}, \alpha, \beta, \gamma, \boldsymbol{\lambda}} [\log \text{ Eqn. (65)}]) \tag{69}$$

$$q^*(\boldsymbol{\eta}, \boldsymbol{\xi}, \boldsymbol{s}, \boldsymbol{t}|\boldsymbol{Z}) \propto \exp(\mathbb{E}_{\boldsymbol{Z}, \alpha, \beta, \gamma, \boldsymbol{\lambda}, \tilde{\boldsymbol{\nu}}} [\log \text{ Eqn. (65)}]) \tag{70}$$

where $\neg dn$ denotes a set of indices without d and n.

D. Variational Posterior Distributions

We derive the formulas for updating variational posterior distributions according to (67)–(70).

1) CVB-E Step: An optimal variational distribution of $\boldsymbol{Z}$ can be obtained as the product of multinomial distributions. The posterior probability that $\boldsymbol{x}_{dn}$ was generated from the mth harmonic partial of basis k is given by

$$\begin{aligned}\log q^*(z_{dnkm}=1) &= \mathbb{E}_{\boldsymbol{z}^{\neg dn}}\left[\log\left(\mathbb{G}[\alpha\nu_k]+n_{dk\cdot}^{\neg dn}\right)\right]\\ &\quad+\mathbb{E}_{\boldsymbol{z}^{\neg dn}}\left[\log\left(\frac{\mathbb{G}[\beta\lambda_1]+n_{\cdot km}^{\neg dn}}{\mathbb{E}[\beta]+n_{\cdot k\geq m}^{\neg dn}}\prod_{m'=1}^{m-1}\frac{\mathbb{G}[\beta\lambda_2]+n_{\cdot k>m'}^{\neg dn}}{\mathbb{E}[\beta]+n_{\cdot k\geq m'}^{\neg dn}}\right)\right]\\ &\quad+\mathbb{E}_{\boldsymbol{z}^{\neg dn}}\left[\log\mathcal{S}(\boldsymbol{x}_{dnm}|\boldsymbol{m}_{zk}^{\neg dn},\boldsymbol{L}_{zk}^{\neg dn},c_{zk}^{\neg dn})\right]+\text{const.}\end{aligned}\tag{71}$$

where $\mathbb{G}[x]$ is the geometric average ($\mathbb{G}[x]=\exp(\mathbb{E}[\log x])$) and $\mathcal{S}$ is the Student-t distribution defined by the three parameters $\boldsymbol{m}_{zk}^{\neg dn}$, $\boldsymbol{L}_{zk}^{\neg dn}$, and $c_{zk}^{\neg dn}$. $\boldsymbol{L}_{zk}^{\neg dn}$ is given by

$$\boldsymbol{L}_{zk}^{\neg dn}=\frac{b_{zk}^{\neg dn}}{1+b_{zk}^{\neg dn}}c_{zk}^{\neg dn}\boldsymbol{W}_{zk}^{\neg dn}\tag{72}$$

where $b_{zk}^{\neg dn}$, $c_{zk}^{\neg dn}$, $\boldsymbol{m}_{zk}^{\neg dn}$, and $\boldsymbol{W}_{zk}^{\neg dn}$ are obtained according to (26)–(29) in which z_{dnkm} is substituted for γ_{dnkm} and the sums are calculated without $\boldsymbol{z}_{dn}$. Each term of (71) can be approximated efficiently by using first-order and second-order Taylor expansions [45]–[47].

Equation (71) calculates the geometric averages of three *predictive distributions* under posterior distributions. These predictive distributions are derived from an infinite-dimensional Dirichlet distribution (a DP for an infinite mixture of iGMMs), stick-breaking construction (a DP for an iGMM), and a Gaussian distribution. Interestingly, this corresponds to (13) based on the geometric averages of three *likelihood functions* under posterior distributions. This implies that CVB is more robust to the local-optima problem than standard VB is.

2) CVB-M Step: We can optimize the variational posterior distributions of the hyperparameters analytically by optimizing those of the auxiliary variables. First, α, β, and γ are gamma distributed as follows:

$$q^*(\alpha)\propto\alpha^{a_\alpha+\mathbb{E}[s_{\cdot\cdot}]-1}e^{-\alpha\left(b_\alpha-\sum_d\mathbb{E}[\log\eta_d]\right)}\tag{73}$$

$$q^*(\beta)\propto\beta^{a_\beta+\mathbb{E}[t_{\cdots}]-1}e^{-\beta\left(b_\beta-\sum_{km}\mathbb{E}[\log\xi_{km}]\right)}\tag{74}$$

$$q^*(\gamma)\propto\gamma^{a_\gamma+K-1}e^{-\gamma\left(b_\gamma-\sum_k\mathbb{E}[\log(1-\tilde{\nu}_k)]\right)}\tag{75}$$

and $\boldsymbol{\lambda}$ and $\tilde{\boldsymbol{\tau}}$ are beta distributed as follows:

$$q^*(\boldsymbol{\lambda})\propto\lambda_1^{u_1+\mathbb{E}[t_{\cdot\cdot1}]-1}\lambda_2^{u_2+\mathbb{E}[t_{\cdot\cdot2}]-1}\tag{76}$$

$$q^*(\tilde{\nu}_k)\propto\tilde{\nu}_k^{1+\mathbb{E}[s_{\cdot k}]-1}(1-\tilde{\nu}_k)^{\mathbb{E}[\gamma]+\mathbb{E}[s_{\cdot>k}]-1}.\tag{77}$$

Then $\boldsymbol{\eta}$ and $\boldsymbol{\xi}$ are beta distributed as follows:

$$q^*(\eta_d)\propto\eta_d^{\mathbb{E}[\alpha]-1}(1-\eta_d)^{n_{d\cdot\cdot}-1}\tag{78}$$

$$q^*(\xi_{km}|\boldsymbol{Z})\propto\xi_{km}^{\mathbb{E}[\beta]-1}(1-\xi_{km})^{n_{\cdot k\geq m}-1}\tag{79}$$

and $\boldsymbol{s}$ and $\boldsymbol{t}$ are multinomial distributed as follows:

$$q^*(s_{dk}=s|\boldsymbol{Z})\propto\begin{bmatrix}n_{dk\cdot}\\ s\end{bmatrix}\mathbb{G}[\alpha\nu_k]^s\tag{80}$$

$$q^*(t_{km1}=t|\boldsymbol{Z})\propto\begin{bmatrix}n_{\cdot km}\\ t\end{bmatrix}\mathbb{G}[\beta\lambda_1]^t\tag{81}$$

$$q^*(t_{km2}=t|\boldsymbol{Z})\propto\begin{bmatrix}n_{\cdot k>m}\\ t\end{bmatrix}\mathbb{G}[\beta\lambda_2]^t.\tag{82}$$

To optimize the variational posterior distributions, we need to calculate the expectations of these variables. If a random variable x follows $\mathrm{Gam}(x|a,b)$ with shape parameter a and rate parameter b, its expectations are given by $\mathbb{E}[x]=a/b$ and $\mathbb{E}[\log x]=\psi(a)-\log(b)$. If x follows $\mathrm{Beta}(x|c,d)$ with parameters c and d, its expectations are given by $\mathbb{E}[x]=c/(c+d)$ and $\mathbb{E}[\log x]=\psi(c)-\psi(c+d)$. Note that the distributions given by Equations (79)–(82) are conditioned by $\boldsymbol{Z}$. The expectations must therefore be averaged over $\boldsymbol{Z}$. For example, we now have the following conditional expectation:

$$\mathbb{E}[\log\xi_{km}|\boldsymbol{Z}]=\psi\left(\mathbb{E}[\beta]\right)-\psi\left(\mathbb{E}[\beta]+n_{\cdot k\geq m}\right).\tag{83}$$

We use Taylor expansion to average $\mathbb{E}[\log\xi_{km}|\boldsymbol{Z}]$ over $\boldsymbol{Z}$, but the digamma function ψ diverges to negative infinity much faster than the logarithmic function does in the vicinity of the origin. To solve this problem, we use a method that treats the case $n_{\cdot k\geq m}=0$ exactly and applies second-order approximation when $n_{\cdot k\geq m}>0$ [45]. We can similarly average the following conditional expectations:

$$\mathbb{E}[s_{dk}|\boldsymbol{Z}]=\mathbb{G}[\alpha\nu_k]\left(\psi\left(\mathbb{G}[\alpha\nu_k]+n_{dk\cdot}\right)-\psi\left(\mathbb{G}[\alpha\nu_k]\right)\right)\tag{84}$$

$$\mathbb{E}[t_{km1}|\boldsymbol{Z}]=\mathbb{G}[\beta\lambda_1]\left(\psi\left(\mathbb{G}[\beta\lambda_1]+n_{\cdot km}\right)-\psi\left(\mathbb{G}[\beta\lambda_1]\right)\right)\tag{85}$$

$$\mathbb{E}[t_{km2}|\boldsymbol{Z}]=\mathbb{G}[\beta\lambda_2]\left(\psi\left(\mathbb{G}[\beta\lambda_2]+n_{\cdot k>m}\right)-\psi\left(\mathbb{G}[\beta\lambda_2]\right)\right).\tag{86}$$

To estimate F0s in the end, we explicitly compute the variational posterior distributions of the integrated-out parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Lambda}$. To do this, we need to execute the standard VB-M step once using $q(\boldsymbol{Z})$ obtained in the CVB-E step.

E. Variational Lower Bound

As in LHA, we monitor the increase of the variational lower bound of evidence $p(\boldsymbol{X})$, which is given by

$$\begin{aligned}\mathcal{L}&=\mathbb{E}\left[\log p(\boldsymbol{X},\boldsymbol{Z},\alpha,\beta,\gamma,\boldsymbol{\lambda},\tilde{\boldsymbol{\nu}})\right]-\mathbb{E}\left[\log q(\boldsymbol{Z},\alpha,\beta,\gamma,\boldsymbol{\lambda},\tilde{\boldsymbol{\nu}})\right]\\ &=\mathbb{E}\left[\log p(\boldsymbol{X}|\boldsymbol{Z})\right]+\mathbb{E}\left[\log p(\boldsymbol{Z}|\alpha,\beta,\boldsymbol{\lambda},\tilde{\boldsymbol{\nu}})\right]-\mathbb{E}\left[\log q(\boldsymbol{Z})\right]\\ &\quad+\mathbb{E}\left[\log p(\alpha)\right]+\mathbb{E}\left[\log p(\beta)\right]+\mathbb{E}\left[\log p(\gamma)\right]\\ &\quad-\mathbb{E}\left[\log q(\alpha)\right]-\mathbb{E}\left[\log q(\beta)\right]-\mathbb{E}\left[\log q(\gamma)\right]\\ &\quad+\mathbb{E}\left[\log p(\boldsymbol{\lambda})\right]+\mathbb{E}\left[\log p(\tilde{\boldsymbol{\nu}}|\gamma)\right]-\mathbb{E}\left[\log q(\boldsymbol{\lambda})\right]\\ &\quad-\mathbb{E}\left[\log q(\tilde{\boldsymbol{\nu}})\right].\end{aligned}\tag{87}$$

The calculation of these terms is described in Appendix II.

VI. Evaluation

This section reports the results of two comparative evaluation experiments. We compared LHA and iLHA with PreFEst and HTC because these four methods are based on the same idea for modeling harmonic structures. Using a different data set, we then compared iLHA with NMF-based methods and other methods. In the latter experiment, we investigated how significantly the value of the scaling factor ω (i.e., how many frequency particles are assumed to be observed in total) affects the accuracy of multipitch analysis.

TABLE II
FRAME-LEVEL F-MEASURES OF F0 DETECTION

Piece number	Optimized		Automated	
RWC-MDB-	PreFEst [2]	HTC [4]	LHA	iLHA
J-2001 No.1	75.8	79.0	70.7	**82.2**
J-2001 No.2	**78.5**	78.0	69.1	77.9
J-2001 No.6	70.4	**78.3**	49.8	71.2
J-2001 No.7	83.0	**86.0**	70.2	85.5
J-2001 No.8	**85.7**	84.4	55.9	84.6
J-2001 No.9	85.9	**89.5**	68.9	84.7
C-2001 No.30	76.0	**83.6**	81.4	81.6
C-2001 No.35	72.8	76.0	58.9	**79.6**
Total	78.5	**81.8**	68.6	80.9

TABLE III
FRAME-LEVEL ACCURACY OF F0 DETECTION

Method	$\mathcal{P}$	$\mathcal{R}$	$\mathcal{F}$
Unconstrained NMF	58.9	60.0	57.8
NMF under harmonicity constraints	63.2	60.9	60.5
NMF under harmonicity and source-filter constraints [14]	60.1	59.1	57.5
NMF under harmonicity and spectral smoothness constraints [16]	**71.6**	**65.5**	**67.0**
Harmonic sum [24]	62.1	21.6	31.5
Correlogram [25]	43.1	23.9	30.3
Spectral peak clustering [26]	65.7	57.4	60.2
iLHA	65.8	59.4	61.2

A. Comparison with Conventional Parametric Methods

1) Experimental Conditions: We evaluated LHA and iLHA on a test that was used in [4] and consisted of eight pieces of piano and guitar solo performances excerpted from the RWC music database [48]. The first 23 s of each piece were used for evaluation. Spectral analysis with a 16-ms time resolution was conducted using a wavelet transform with Gabor wavelets. The correct values and temporal positions of actual F0s were prepared by hand as ground truth. Denoting by g_d, e_d, and c_d the respective numbers of ground-truth, estimated, and correct F0s on frame d, we calculated the following frame-level recall and precision rates and F-measure for each piece:

$$\mathcal{R} = 100 \cdot \frac{\sum_d c_d}{\sum_d g_d} \quad \mathcal{P} = 100 \cdot \frac{\sum_d c_d}{\sum_d e_d} \quad \mathcal{F} = \frac{2\mathcal{R}\mathcal{P}}{\mathcal{R}+\mathcal{P}} \tag{88}$$

and we averaged each of these measures over all pieces.

The prior and hyperprior distributions of LHA and iLHA were set to noninformative distributions. In LHA, K and M were set to 60 and 15. In iLHA, K^+ and M^+ were also set to 60 and 15. iLHA is not sensitive to these values, and no other tuning was needed for either method. To output F0s at each frame, we extracted bases whose expected weights $\boldsymbol{\pi}$ were over a threshold that was optimized as in [4].

For comparison, we referred to the PreFEst and HTC experimental results reported in [4]. Although the ground-truth data in that study was slightly different from ours, it was close enough for roughly evaluating performance comparatively. The number of bases, priors, and weighting factors of the PreFEst and HTC were carefully tuned to optimize the results. Although this is not realistic, the *upper bounds* of potential performance were investigated in the literature.

2) Experimental Results: The results listed in Table II show that the performance of iLHA closely approached and sometimes surpassed that of HTC. This is consistent with the empirical findings of many studies on Bayesian nonparametrics that nonparametric models were competitive with optimally tuned parametric models. HTC outperformed PreFEst because HTC can appropriately deal with temporal continuity of spectral bases. This implies that incorporating temporal modeling would improve the performance of iLHA.

The results of LHA were worse than those of iLHA because LHA is not formulated in a hierarchical Bayesian manner and requires precise priors. In fact, we confirmed that the results of PreFEst and HTC based on MAP estimation were drastically degraded when using noninformative priors. Automated iLHA, in contrast, stably showed the good performance.

We found that model flexibility can be greatly enhanced by making time-consuming fine tuning unnecessary. Conventional studies assumed that appropriate prior knowledge is required to constrain flexibility (called *regularization*). By using a truly flexible hierarchical model based on Bayesian nonparametrics, however, we can let the data speak for itself. This naturally results in optimal performance.

B. Comparison With NMF-Based Methods and Other Methods

1) Experimental Conditions: We then evaluated iLHA on a test set that was used in [16] and consisted of 50 pieces of piano solo performances excerpted from the MAPS piano database [8]. The first 30 s of each piece were used for evaluation. Spectral analysis with a 10-ms time resolution was conducted using a Gabor wavelet transform. The value of K^+ was increased to 88, (the number of notes in a standard piano) because the piano pieces were much sophisticated than those used in first experiment. The time resolution and the value of K^+ were equal to those used in [16], and performance was evaluated in terms of F-measures.

For comparison, we referred to the experimental results of seven methods reported in [16]. We compared iLHA with four NMF-based methods: one using no constraints, one using harmonicity constraints (a subset of [13]), one using harmonicity and source-filter constraints [14], and one using harmonicity and spectral smoothness constraints [16]. Note that only the last one was manually tuned to yield the best results (the effect of hyperparameter tuning was investigated in [16]). We also compared it with a method based on harmonic sums [24], a method based on correlograms [25], and a method based on spectral peak clustering [26].

2) Experimental Results: The results listed in Table III show that iLHA was the second best among the seven methods. Although the best variant of NMF gained the better F-measure (67.0%) than iLHA did (61.2%), we can say that well-automated iLHA is still competitive because it is reported that non-optimal settings deteriorated the performance of NMF moderately [16]. The F-measure of iLHA (61.2%) was close to that of NMF using only harmonicity constraints (60.5%). As discussed in Section III-C2, pLSA and PLCA are proven to have a close connection to NMF. Therefore, the similarity between iLHA based on harmonic GMMs and NMF based on harmonicity constraints was experimentally and theoretically supported. In addition, the difference between NMF using only harmonicity constraints and NMF adding spectral smoothness constraints implies that the performance of iLHA would be improved by incorporating spectral smoothness modeling.

It is interesting that in almost all methods the $\mathcal{P}$ was higher than the $\mathcal{R}$. This means that there were many F0s that were hard to detect because of the complex overlapping of multiple F0s. To solve this problem, more accurate spectral modeling would be required by removing the assumption of amplitude additivity that forms a basis of iLHA and NMF.

3) Impact of Scaling Factor: We investigated the impact of the scaling factor ω described in Section III-C. We tested three different values: $\omega = 0.1, 1, 10$. The similarity of respective F-measures—61.2%, 60.6%, and 60.1%—indicates that the results are not sensitive to the value of the scaling factor ω. The automatic optimization of ω would be an interesting research topic that tackles the limitation of many methods based on the assumption of amplitude quantization.

VII. Conclusion

This paper presented a novel statistical method for detecting multiple F0s in polyphonic music audio signals. In this method, which is called iLHA and is the first to apply Bayesian nonparametrics to multipitch analysis, we formulated nested infinite GMMs that represent polyphonic spectral strips in a hierarchical nonparametric Bayesian manner. More specifically, each spectral strip is allowed to contain an unbounded number of spectral bases, each of which can contain an unbounded number of harmonic partials. The method was fully automated by putting noninformative hyperprior distributions on influential hyperparameters except for the final thresholding process. The joint posterior distribution of all unknown variables can be inferred efficiently according to the VB framework. In our experiments comparing iLHA with the state-of-the-art methods manually optimized by trial and error, we found that iLHA is competitive enough and there is room for improvement based on modeling of temporal continuity and spectral smoothness. One interesting future direction is to use MCMC methods such as Gibbs sampling and more efficient variants for training the iLHA model.

Bayesian nonparametrics is a powerful framework avoiding the model selection problem faced in various areas of music information retrieval (MIR). For example, how many sections are required for structuring a musical piece? How many groups are required for clustering listeners according to their tastes or musical pieces according to their contents? We can avoid these problems by assuming that in theory there is an infinite number of objects (sections or groups) behind available observed data. Unnecessary objects are automatically removed from consideration through statistical inference. Hoffman *et al.* recently successfully applied this framework to the calculation of musical similarity [49] and the detection of repeated patterns [32], and we also plan to use this powerful framework in a wide range of applications.

Appendix I

The nine terms of the variational lower bound of LHA in (34) can be calculated as follows:

$$\begin{aligned}
&\mathbb{E}\left[\log p(\boldsymbol{X}|\boldsymbol{Z},\boldsymbol{\mu},\boldsymbol{\Lambda})\right]\\
&\quad= \sum_{dnkm} \gamma_{dnkm} \mathbb{E}_{\boldsymbol{\mu},\boldsymbol{\Lambda}}\left[\log \mathcal{N}\left(\boldsymbol{x}_{dnm}|\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k^{-1}\right)\right]\\
&\mathbb{E}\left[\log p(\boldsymbol{Z}|\boldsymbol{\pi},\boldsymbol{\tau})\right]\\
&\quad= \sum_{dnkm} \gamma_{dnkm} \left(\mathbb{E}_{\boldsymbol{\pi}_d}[\log \pi_{dk}] + \mathbb{E}_{\boldsymbol{\tau}_k}[\log \tau_{km}]\right)\\
&\mathbb{E}\left[\log p(\boldsymbol{\pi})\right]\\
&\quad= D \log C(\alpha\boldsymbol{\nu}) + \sum_{dk}(\alpha\nu_k - 1)\mathbb{E}_{\boldsymbol{\pi}_d}[\log \pi_{dk}]\\
&\mathbb{E}\left[\log p(\boldsymbol{\tau})\right]\\
&\quad= K \log C(\beta\boldsymbol{\upsilon}) + \sum_{km}(\beta\upsilon_m - 1)\mathbb{E}_{\boldsymbol{\tau}_k}[\log \tau_{km}]\\
&\mathbb{E}\left[\log p(\boldsymbol{\mu},\boldsymbol{\Lambda})\right]\\
&\quad= \sum_k \mathbb{E}_{\boldsymbol{\mu}_k,\boldsymbol{\Lambda}_k}\left[\log \mathcal{N}\left(\boldsymbol{\mu}_k|\boldsymbol{m}_0, (b_0\boldsymbol{\Lambda}_k)^{-1}\right) \mathcal{W}(\boldsymbol{\Lambda}_k|\boldsymbol{W}_0, c_0)\right]\\
&\mathbb{E}\left[\log q(\boldsymbol{Z})\right]\\
&\quad= \sum_{dnkm} \gamma_{dnkm} \log \gamma_{dnkm}\\
&\mathbb{E}\left[\log q(\boldsymbol{\pi})\right]\\
&\quad= \sum_d \log C(\boldsymbol{\alpha}_d) + \sum_{dk}(\alpha_{dk} - 1)\mathbb{E}_{\boldsymbol{\pi}_d}[\log \pi_{dk}]\\
&\mathbb{E}\left[\log q(\boldsymbol{\tau})\right]\\
&\quad= \sum_k \log C(\boldsymbol{\beta}_k) + \sum_{km}(\beta_{km} - 1)\mathbb{E}_{\boldsymbol{\tau}_u}[\log \tau_{km}]\\
&\mathbb{E}\left[\log q\left(\boldsymbol{\mu},\boldsymbol{\Lambda}\right))\right]\\
&\quad= \sum_k \mathbb{E}_{\boldsymbol{\mu}_k,\boldsymbol{\Lambda}_k}\left[\log \mathcal{N}\left(\boldsymbol{\mu}_k|\boldsymbol{m}_k, (b_k\boldsymbol{\Lambda}_k)^{-1}\right) \mathcal{W}(\boldsymbol{\Lambda}_k|\boldsymbol{W}_k, c_k)\right]
\end{aligned}$$

where the fifth and last terms can be obtained as follows:

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{\mu}_k,\boldsymbol{\Lambda}_k}\left[\log \mathcal{N}\left(\boldsymbol{\mu}_k|\boldsymbol{m}_0, (b_0\boldsymbol{\Lambda}_k)^{-1}\right) \mathcal{W}(\boldsymbol{\Lambda}_k|\boldsymbol{W}_0, c_0)\right]\\
&\quad= \frac{1}{2}\log\left(\frac{b_0}{2\pi}\right) + \frac{1}{2}\mathbb{E}_{\boldsymbol{\Lambda}_k}\left[\log|\boldsymbol{\Lambda}_k|\right] + \log B(\boldsymbol{W}_0, c_0)\\
&\qquad - \frac{b_0}{2}\left(c_k(\boldsymbol{m}_k - \boldsymbol{m}_0)^T \boldsymbol{W}_k(\boldsymbol{m}_k - \boldsymbol{m}_0) + \frac{1}{b_k}\right)\\
&\qquad + \frac{c_0 - 2}{2}\mathbb{E}_{\boldsymbol{\Lambda}_k}\left[\log|\boldsymbol{\Lambda}_k|\right] - \frac{c_k}{2}\mathrm{Tr}\left(\boldsymbol{W}_0^{-1}\boldsymbol{W}_k\right)\\
&\mathbb{E}_{\boldsymbol{\mu}_k,\boldsymbol{\Lambda}_k}\left[\log \mathcal{N}\left(\boldsymbol{\mu}_k|\boldsymbol{m}_k, (b_k\boldsymbol{\Lambda}_k)^{-1}\right) \mathcal{W}(\boldsymbol{\Lambda}_k|\boldsymbol{W}_k, c_k)\right]\\
&\quad= -\mathbb{E}_{\boldsymbol{\Lambda}_k}\left[H\left[q(\boldsymbol{\mu}_k|\boldsymbol{\Lambda}_k)\right]\right] - H\left[q(\boldsymbol{\Lambda}_k)\right]\\
&\quad= \frac{1}{2}\mathbb{E}_{\boldsymbol{\Lambda}_k}\left[\log|\boldsymbol{\Lambda}_k|\right] + \frac{1}{2}\log\left(\frac{b_k}{2\pi}\right) - \frac{1}{2} + \log B(\boldsymbol{W}_k, c_k)\\
&\qquad + \frac{c_k - 2}{2}\mathbb{E}_{\boldsymbol{\Lambda}_k}\left[\log|\boldsymbol{\Lambda}_k|\right] - \frac{c_k}{2}
\end{aligned}$$

Appendix II

The 13 terms of the variational lower bound of iLHA in (87) can be calculated as follows:

$$\begin{aligned}
\mathbb{E}[\log p(\boldsymbol{X}|\boldsymbol{Z})] = &-\frac{n_{\cdots}}{2}\log(2\pi) + \frac{1}{2}\sum_k \log b_0\\
&- \frac{1}{2}\sum_k \mathbb{F}_2[\log b_{zk}]\\
&+ \sum_k \log B(\boldsymbol{W}_0, c_0)\\
&- \sum_k \mathbb{F}_1\left[\log B(\boldsymbol{W}_{zk}, c_{zk})\right]
\end{aligned}$$

$$
\begin{aligned}
\mathbb{E}[\log p(\boldsymbol{Z}|\alpha,\beta,\boldsymbol{\lambda},\tilde{\boldsymbol{\nu}})] =& \sum_{d} \log\left(\frac{\Gamma(\mathbb{E}[\alpha])}{\Gamma(\mathbb{E}[\alpha]+n_{d\cdot\cdot})}\right) \\
&+ \sum_{dk} \mathbb{F}_2\left[\log\left(\frac{\Gamma(\mathbb{G}[\alpha\nu_k]+n_{dk\cdot})}{\Gamma(\mathbb{G}[\alpha\nu_k])}\right)\right] \\
&+ \sum_{km} \mathbb{F}_2\left[\log\left(\frac{\Gamma(\mathbb{E}[\beta])}{\Gamma(\mathbb{E}[\beta]+n_{\cdot k\geq m})}\right)\right] \\
&+ \sum_{km} \mathbb{F}_2\left[\log\left(\frac{\Gamma(\mathbb{G}[\beta\lambda_1]+n_{\cdot km})}{\Gamma(\mathbb{G}[\beta\lambda_1])}\right)\right] \\
&+ \sum_{km} \mathbb{F}_2\left[\log\left(\frac{\Gamma(\mathbb{G}[\beta\lambda_2]+n_{\cdot k>m})}{\Gamma(\mathbb{G}[\beta\lambda_2])}\right)\right] \\
\mathbb{E}[\log p(\alpha)] =& -\log\Gamma(a_\alpha)+a_\alpha\log b_\alpha \\
&+ (a_\alpha-1)\mathbb{E}_\alpha[\log\alpha] - b_\alpha\mathbb{E}_\alpha[\alpha] \\
\mathbb{E}[\log p(\beta)] =& -\log\Gamma(a_\beta)+a_\beta\log b_\beta \\
&+ (a_\beta-1)\mathbb{E}_\beta[\log\beta] - b_\beta\mathbb{E}_\beta[\beta] \\
\mathbb{E}[\log p(\gamma)] =& -\log\Gamma(a_\gamma)+a_\gamma\log b_\gamma \\
&+ (a_\gamma-1)\mathbb{E}_\gamma[\log\gamma] - b_\gamma\mathbb{E}_\gamma[\gamma] \\
\mathbb{E}[\log p(\boldsymbol{\lambda})] =& \log\frac{\Gamma(u_1+u_2)}{\Gamma(u_1)\Gamma(u_2)} \\
&+ (u_1-1)\mathbb{E}[\log\lambda_1] \\
&+ (u_2-1)\mathbb{E}[\log\lambda_2] \\
\mathbb{E}[\log p(\tilde{\boldsymbol{\nu}}|\gamma)] =& K\mathbb{E}[\log\gamma] \\
&+ \sum_{k}\left(\mathbb{E}[\gamma]-1\right)\mathbb{E}[\log(1-\tilde{\nu}_k)] \\
\mathbb{E}[q(\boldsymbol{Z})] =& \sum_{dnkm}\gamma_{dnkm}\log\gamma_{dnkm} \\
\mathbb{E}[\log q(\alpha)] =& -H\left[\text{PosteriorGamma}(\alpha)\right] \\
\mathbb{E}[\log q(\beta)] =& -H\left[\text{PosteriorGamma}(\beta)\right] \\
\mathbb{E}[\log q(\gamma)] =& -H\left[\text{PosteriorGamma}(\gamma)\right] \\
\mathbb{E}[\log q(\boldsymbol{\lambda})] =& -H\left[\text{PosteriorBeta}(\lambda_1)\right] \\
\mathbb{E}[\log q(\tilde{\boldsymbol{\nu}})] =& \sum_{k} -H\left[\text{PosteriorBeta}(\tilde{\nu}_k)\right]
\end{aligned}
$$

where $\mathbb{F}_1$ and $\mathbb{F}_2$ mean the first-order and second-order approximations based on Taylor expansion (see [45]–[47]).

ACKNOWLEDGMENT

The authors would like to thank Dr. H. Kameoka (The University of Tokyo/NTT, Japan) for providing the ground-truth transcriptions of eight pieces included in the RWC music databases [48]. They would also like to thank Dr. V. Emiya (INRIA, France) for allowing them to use the valuable MAPS piano database [8].

REFERENCES

[1] P. Orbanz and Y. W. Teh, "Bayesian nonparametric models," in *Encyclopedia of Machine Learning*. New York: Springer, 2010.

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

[3] H. Kameoka, T. Nishimoto, and S. Sagayama, "Separation of harmonic structures based on tied Gaussian mixture model and information criterion for concurrent sounds," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2004, vol. 4, pp. 297–300.

[4] H. Kameoka, T. Nishimoto, and S. Sagayama, "A multipitch analyzer based on harmonic temporal structured clustering," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 3, pp. 982–994, Mar. 2007.

[5] C. Raphael, "Automatic transcription of piano music," in *Proc. 3rd Int. Conf. Music Inf. Retrieval (ISMIR)*, 2002, pp. 161–166.

[6] A. T. Cemgil, H. J. Kappen, and D. Barber, "A generative model for music transcription," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 2, pp. 679–694, Mar. 2006.

[7] S. A. Raczyński, E. Vincent, F. Bimbot, and S. Sagayama, "Multiple pitch transcription using DBN-based musicological models," in *Proc. 11th Int. Conf. Music Inf. Retrieval (ISMIR)*, 2010, pp. 363–368.

[8] V. Emiya, R. Badeau, and B. David, "Multipitch estimation of piano sounds using a new probabilistic spectral smoothness principle," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 6, pp. 1643–1654, Aug. 2010.

[9] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2000, pp. 556–562.

[10] P. Smaragdis and J. C. Brown, "Non-negative matrix factorization for polyphonic music transcription," in *Proc. IEEE Workshop Applicat. Signal Process. Audio Acoust. (WASPAA)*, 2003, pp. 177–180.

[11] T. O. Virtanen, A. T. Cemgil, and S. J. Godsill, "Bayesian extensions to nonnegative matrix factorisation for audio signal modelling," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2008, pp. 45–48.

[12] P. H. Peeling, A. T. Cemgil, and S. J. Godsill, "Generative spectrogram factorization models for polyphonic piano transcription," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 3, pp. 519–527, Mar. 2010.

[13] S. A. Raczyński, N. Ono, and S. Sagayama, "Multipitch analysis with harmonic non-negative matrix approximation," in *Proc. 6th Int. Conf. Music Inf. Retrieval (ISMIR)*, 2007, pp. 381–386.

[14] D. FitzGerald, M. Cranitch, and E. Coyle, "Extended nonnegative tensor factorisation models for musical sound source separation," *Comput. Intell. Neurosci.*, vol. 2008, 2008.

[15] N. Bertin, R. Badeau, and E. Vincent, "Enforcing harmonicity and smoothness in Bayesian non-negative matrix factorization applied to polyphonic music transcription," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 3, pp. 538–549, Mar. 2010.

[16] E. Vincent, N. Bertin, and R. Badeau, "Adaptive harmonic spectral decomposition for multiple pitch estimation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 3, pp. 528–537, Mar. 2010.

[17] A. Cont, "Realtime multiple pitch observation using sparse non-negative constraints," in *Proc. 7th Int. Conf. Music Inf. Retrieval (ISMIR)*, 2006, pp. 206–211.

[18] H. Kameoka, T. Nishimoto, and S. Sagayama, "Complex NMF: A new sparse representation for acoustic signals," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2009, pp. 45–48.

[19] C. Févotte, N. Bertin, and J.-L. Durrieu, "Nonnegative matrix factorization with the Itakura-Saito divergence: With application to music analysis," *Neural Comput.*, vol. 21, no. 3, pp. 793–830, 2009.

[20] M. Hoffman, D. Blei, and P. Cook, "Bayesian nonparametric matrix factorization for recorded music," in *Proc. 27th Int. Conf. Mach. Learn. (ICML)*, 2010, pp. 439–446.

[21] *Signal Processing Methods for Music Transcription*, A. Klapuri and M. Davy, Eds. New York: Springer, 2010.

[22] M. Marolt, "A connectionist approach to transcription of polyphonic piano music," *IEEE Trans. Multimedia*, vol. 6, no. 3, pp. 439–449, 2004.

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

[24] A. Klapuri, "Multiple fundamental frequency estimation by summing harmonic amplitudes," in *Proc. 7th Int. Conf. Music Inf. Retrieval (ISMIR)*, 2006, pp. 216–221.

[25] T. Tolonen and M. Karjalainen, "A computationally efficient multipitch analysis model," *IEEE Trans. Speech Audio Process.*, vol. 8, no. 6, pp. 708–716, Nov. 2000.

[26] A. Pertusa and J. M. Iñesta, "Multiple fundamental frequency estimation using Gaussian smoothness," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2008, pp. 105–108.

[27] J. P. Bello, L. Daudet, and M. B. Sandler, "Automatic piano transcription using frequency and time-domain information," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 6, pp. 2242–2251, Nov. 2006.

[28] K. Dressler, "Extraction of the melody pitch contour from polyphonic audio," in *Proc. 2nd Music Inf. Retrieval Eval. eXchange (MIREX)*, 2005 [Online]. Available: http://www.musicir.org/evaluation/mirex-results/articles/melody/dressler.pdf

[29] M. P. Ryynänen and A. P. Klapuri, "Transcription of the singing melody in polyphonic music," in *7th Int. Conf. Music Inf. Retrieval (ISMIR)*, 2006, pp. 206–211.

[30] J.-L. Durrieu, G. Richard, B. David, and C. Févotte, "Source/filter model for unsupervised main melody extraction from polyphonic audio signals," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 3, pp. 564–575, Mar. 2010.

[31] G. E. Poliner, D. P. Ellis, A. F. Ehmann, E. Gómez, S. Streich, and B. Ong, "Melody transcription from music audio: Approaches and evaluation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 4, pp. 1247–1256, May 2007.

[32] M. Hoffman, D. Blei, and P. Cook, "Finding latent sources in recorded music with a shift-invariant HDP," in *Proc. 12th Int. Conf. Digital Audio Effects (DAFX)*, 2009.

[33] D. Blei, A. Ng, and M. Jordan, "Latent Dirichlet allocation," *Mach. Learn. Res.*, vol. 3, pp. 993–1022, 2003.

[34] T. Hofmann and J. Puzicha, "Probabilistic latent semantic indexing," in *Proc. 22nd Int. Conf. Res. Develop. Inf. Retrieval (SIGIR)*, 1999, pp. 50–57.

[35] M. Shashanka, B. Raj, and P. Smaragdis, "Probabilistic latent variable models as non-negative factorizations," *Comput. Intell. Neuosci.*, vol. 2008, 2008.

[36] E. Gaussier and C. Goutte, "Relation between pLSA and NMF and implications," in *Proc. 28th Int. Conf. Res. Develop. Inf. Retrieval (SIGIR)*, 2005, pp. 601–602.

[37] H. Attias, "A variational Bayesian framework for graphical models," in *Adv. Neural Inf. Process. Syst. (NIPS)*, 2000, pp. 209–215.

[38] R. M. Neal, "Probabilistic inference using Markov chain Monte Carlo methods," Dept. of Comput. Sci., Univ. of Toronto, Toronto, ON, Canada, Tech. Rep. CRG-TR-93-1, 1993.

[39] Y. W. Teh, "Dirichlet processes," in *Encyclopedia of Machine Learning*. New York: Springer, 2010.

[40] T. Ferguson, "Bayesian analysis of some nonparametric problems," *Ann. Statist.*, vol. 1, no. 2, pp. 209–230, 1973.

[41] J. Sethuraman, "A constructive definition of Dirichlet priors," *Statist. Sinica*, vol. 4, pp. 639–650, 1994.

[42] H. Ishwaran and L. F. James, "Gibbs sampling methods for stick-breaking priors," *J. Amer. Statist. Assoc.*, vol. 96, no. 453, pp. 161–173, 2001.

[43] R. Gray, "Vector quantization," *IEEE ASSP Mag.*, vol. 1, no. 2, pp. 4–29, Apr. 1984.

[44] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei, "Hierarchical Dirichlet processes," *J. Amer. Statist. Assoc.*, vol. 101, no. 476, pp. 1566–1581, 2006.

[45] Y. W. Teh, K. Kurihara, and M. Welling, "Collapsed variational inference for HDP," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, 2007.

[46] J. Sung, Z. Ghahramani, and S.-Y. Bang, "Latent-space variational Bayes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 12, pp. 2236–2242, Dec. 2008.

[47] J. Sung, Z. Ghahramani, and S.-Y. Bang, "Second-order latent-space variational Bayes for approximate Bayesian inference," *IEEE Signal Process. Lett.*, vol. 15, pp. 918–921, 2008.

[48] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka, "RWC music database: Popular, classical, and jazz music database," in *Proc. 3th Int. Conf. Music Inf. Retrieval (ISMIR)*, 2002, pp. 287–288.

[49] M. Hoffman, D. Blei, and P. Cook, "Content-based musical similarity computation using the hierarchical Dirichlet process," in *Proc. 9th Int. Conf. Music Inf. Retrieval (ISMIR)*, 2008, pp. 349–354.

Kazuyoshi Yoshii (M'08) received the Ph.D. degree in informatics from Kyoto University, Kyoto, Japan, in 2008.

He is currently a Research Scientist at the National Institute of Advanced Industrial Science and Technology (AIST), Tsukuba, Japan. His research interests include probabilistic music analysis, blind source separation, and Bayesian nonparametrics.

Dr. Yoshii has received several awards including the IPSJ Yamashita SIG Research Award and the Best-in-Class Award of MIREX 2005. He is a member of the Information Processing Society of Japan (IPSJ) and Institute of Electronics, Information, and Communication Engineers (IEICE).

Masataka Goto received the D.Eng. degree from Waseda University, Tokyo, Japan, in 1998.

He is currently the leader of the Media Interaction Group, National Institute of Advanced Industrial Science and Technology (AIST), Tsukuba, Japan. He serves concurrently as a Visiting Professor at the Institute of Statistical Mathematics, an Associate Professor (Cooperative Graduate School Program) in the Graduate School of Systems and Information Engineering, University of Tsukuba, and a Project Manager of the MITOH Program (the Exploratory IT Human Resources Project) Youth division by the Information Technology Promotion Agency (IPA).

Dr. Goto received 25 awards over the past 19 years, 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," the IPSJ Nagao Special Researcher Award, and the IPSJ Best Paper Award.