BAYESIAN NONNEGATIVE HARMONIC-TEMPORAL FACTORIZATION AND ITS APPLICATION TO MULTIPITCH ANALYSIS

Daichi Sakaue Takuma Otsuka Katsutoshi Itoyama Hiroshi G. Okuno
Graduate School of Informatics, Kyoto University
{dsakaue,ohtsuka,itoyama,okuno}@kuis.kyoto-u.ac.jp

ABSTRACT

Since important musical features are mutually dependent, their relations should be analyzed simultaneously. Their Bayesian analysis is particularly important to reveal their statistical relation. As the first step for a unified music content analyzer, we focus on the harmonic and temporal structures of the wavelet spectrogram obtained from harmonic sounds. In this paper, we present a new Bayesian multipitch analyzer, called Bayesian nonnegative harmonic-temporal factorization (BNHTF). BNHTF models the harmonic and temporal structures separately based on Gaussian mixture model. The input signal is assumed to contain a finite number of harmonic sounds. Each harmonic sound is assumed to emit a large number of sound quanta over the time-log-frequency domain. The observation probability is expressed as the product of two Gaussian mixtures. The number of quanta is calculated in the ϵ-neighborhood of each grid point on the spectrogram. BNHTF integrates latent harmonic allocation (LHA) and nonnegative matrix factorization (NMF) to estimate both the observation probability and the number of quanta. The model is optimized by newly designed deterministic procedures with several approximations for the variational Bayesian inference. Results of experiments on multipitch estimation with 40 musical pieces showed that BNHTF outperforms the conventional method by 0.018 in terms of F-measure on average.

1. INTRODUCTION

Multipitch estimation [5, 7, 10, 19] is one of the most fundamental techniques of music information retrieval (MIR) because the temporal pattern of pitch strongly expresses the content of musical pieces, especially in Western music. It is useful for a wide range of applications, including content-based music search [2], musical instrument identification [9], and chord recognition [15].

One promising technique in multipitch analysis is to assume a probabilistic generative model of musical signals and then perform pattern matching between the model and recorded signals using EM algorithm. Variational Bayesian methods are particularly valuable because they can flexibly model the probabilistic relation between the occurrence pattern of pitch and many important musical aspects, including harmonic structure, musical instrument, musical structure [11], chord, onset, and emotion [8]. Our goal is to formulate a music analyzer that estimates the relation between all such latent variables. At present, latent harmonic allocation (LHA) is the most suitable candidate for further extensions.

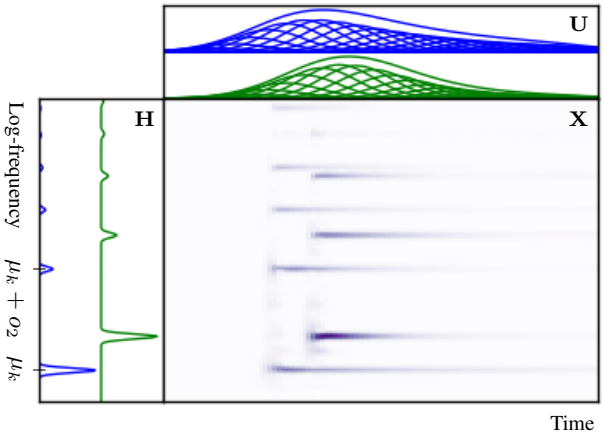

Figure 1. Illustration of Bayesian nonnegative harmonic-temporal factorization. Basis and activation matrices are explicitly modeled using Gaussian mixtures.

The most important features in an observed wavelet spectrogram are the harmonic and temporal structures. These structures are mutually dependent and should therefore be analyzed simultaneously. Conventionally, LHA defines the volume of a sound at a time frame as the relative coefficient to the total volume of that time frame. As a result, there is no explicit variable that describes the actual volume of each sound. This makes it difficult to enhance the model so that it considers the temporal envelope of the sounds.

In this paper, we present a new method that explicitly models the volume of each harmonic sound using a time series of Poisson distributions and its harmonic and temporal structure using mixtures of Gaussians. This is illustrated in Figure 1. Here, the observed spectrogram is interpreted as a histogram of a large number of statistically independent particles. This interpretation corresponds to

the integration of LHA and Bayesian NMF [3], because both of them assumes the spectrogram as a histogram. In our model, the number of observations at each time-log-frequency point is calculated in an ϵ-neighborhood of the point. To do this, we integrate the probabilistic density functions (pdfs) of harmonic structures around the f-th frequency bin, with a quite small amount of a band of width ϵ_{F}. For the temporal structure, we introduce a similar assumption using ϵ_{T}. The objective is to fit the generative model of BNHTF to the standard formulation of NMF. The variational posterior distributions are approximately conjugate, when we take the limit $\epsilon_{\mathrm{F}}, \epsilon_{\mathrm{T}} \to 0$. The formulation strategy is similar to those formerly obtained by Ochiai [12, 13]. Ours seem to deliver a more concrete interpretation of the generative process, rather than their construction which depends on the direct use of Dirac delta function.

Our method can also be viewed as a Bayesian NMF-based method with adaptive harmonic basis. This method should be quite valuable for many researchers, because such a model has been searched for long years [1, 14, 18]. As the proposed method is a variation of NMF, our method can easily be extended further by using recent improvements of NMF.

2. CONVENTIONAL METHODS

2.1 Harmonic Clustering

The spectral envelope of harmonic sound has several peaks at its fundamental frequency and the overtone frequencies. This structure, known as harmonic structure, can be approximated by using a mixture of Gaussians. In this paper, we call this approach *harmonic clustering*. Harmonic clustering methods represent the wavelet spectrum of each harmonic sound as a probabilistic density function described by a mixture of Gaussians. PreFEst [5], harmonic temporal clustering (HTC) [7], and LHA [19] are notable examples of harmonic clustering.

The mathematical representation is as follows. Let x be the log-frequency, μ_k be the logarithm of the fundamental frequency of the k-th harmonic sound, and λ_k be the precision of the Gaussian components. Moreover, let M be the number of harmonic partials considered in the model. The relative weight of each harmonic partial is indicated using $\eta_k = [\eta_{k1}, \cdots, \eta_{kM}]$. Following this, the k-th harmonic sound is represented as:

$$p_k(x|\eta_k, \mu_k, \lambda_k) = \sum_{m=1}^{M} \eta_{km} \mathcal{N}(x|\mu_k + o_m, \lambda_k^{-1}), \quad (1)$$

where $\mathcal{N}$ denotes normal distribution and o_m denotes the relative position of the m-th overtone component on the log-frequency axis. To retain the versatility of the model, in most cases, o_m is set so that the model represents completely harmonic sounds. The relationship between the log and linear frequency scales is defined as:

$$f_{\log} = 1200(\log_2 f - \log_2 440 + 4.75). \quad (2)$$

2.2 Latent Harmonic Allocation

Latent harmonic allocation (LHA) is a variational Bayesian method of harmonic clustering that represents each time frame spectrum of the observed spectrogram using a mixture of harmonic sound models. The observed spectrum is interpreted as a histogram of numerous sound quanta that are generated by the observation model. The t-th time frame spectrum is represented as a linear combination of K harmonic sounds with mixing coefficients $\pi_t = [\pi_{t1} \cdots \pi_{tK}]$. The emission probability of a sound quantum is described as:

$$p_t(x|\pi, \eta, \mu, \lambda) = \sum_{k=1}^{K} \pi_{tk} p_k(x|\eta_k, \mu_k, \lambda_k). \quad (3)$$

Let x_{tf} be the value of the wavelet spectrogram at the t-th time frame and the f-th frequency bin. The overall observation probability of the t-th time frame $X_t = [x_{t1}, \cdots, x_{tF}]$ is described as:

$$p_t(X_t|\pi, \eta, \mu, \lambda) = \prod_{f=1}^{F} \left(\sum_{km} \pi_{tk} \eta_{km} \mathcal{N}(x_f|\mu_k + o_m, \lambda_k^{-1}) \right)^{x_{tf}}, \quad (4)$$

where x_f denotes the log-frequency of the f-th frequency bin.

The volume of the k-th sound at the t-th time frame is implicitly determined by π_{tk}, which describes the relative weight of the k-th sound in the time frame. Because the total volume of each time frame differs between the frames, it can be difficult to discuss the temporal envelope of π_{tk}. To solve this problem, we explicitly model the volume using a time series of Poisson distributions, which is similar to the construction of Bayesian NMF.

2.3 Bayesian Nonnegative Matrix Factorization

Nonnegative matrix factorization (NMF) is a promising information retrieval method that factorizes the observed matrix of size $N \times M$ into two matrices of size $N \times K$ and $K \times M$. Naturally, we set $K \ll N$ and M. In music analysis, NMF is often applied to an STFT or a wavelet spectrogram. The two matrices are learned by minimizing a cost function, $D(\mathbf{X}||\mathbf{UH})$, where $\mathbf{X} = \{x_{tf}\}$ denotes the observed spectrogram and $\mathbf{U} = \{u_t^k\}$ and $\mathbf{H} = \{h_f^k\}$ denotes the factorized matrices. The cost function is usually customized for specific objectives [1, 4, 14, 17]. Following this, x_{tf} is approximated as:

$$x_{tf} \approx \sum_{k=1}^{K} h_f^k u_t^k. \quad (5)$$

As in LHA, the index k stands for the specific spectral pattern. This pattern is described by the vector $[h_1^k, \cdots, h_F^k]$. The volume of the k-th basis in the t-th time frame is denoted by u_t^k. $\mathbf{H}$ is often called the basis matrix, and $\mathbf{U}$ is called the activation matrix.

Cemgil [3] proposed a full Bayesian inference of NMF, in which the joint posterior probability of latent variables

$p(S, \mathbf{H}, \mathbf{U}|\mathbf{X})$ is estimated. Here, S denotes the set of K separated spectrograms. The likelihoods and prior distributions are written as:

$$p(x_{tf}|s_{tf}^{[k]}) = \delta(x_{tf} - \sum_{k=1}^{K} s_{tf}^k), \tag{6}$$

$$p(s_{tf}^k|h_f^k, u_t^k) = \mathcal{P}(s_{tf}^k|h_f^k u_t^k), \tag{7}$$

$$p(h_f^k) = \mathrm{Gam}(h_f^k|a_0, b_0), \tag{8}$$

$$p(u_t^k) = \mathrm{Gam}(u_t^k|a_0, b_0), \tag{9}$$

where s_{tf}^k denotes the observation of k-th sound. Hereafter, a square bracket indicates a set over the index variable, δ denotes delta function, $\mathcal{P}$ denotes Poisson distribution, and Gam denotes Gamma distribution. The prior distributions of h_f^k and u_t^k are the conjugate priors, where a_0, b_0 are the hyperparameters of the distributions. A variational EM algorithm is obtained based on a mean-field approximation, $p(S, \mathbf{H}, \mathbf{U}|\mathbf{X}) \approx q(S)q(\mathbf{H})q(\mathbf{U})$.

The main drawback of NMF is its inability to model spectral and temporal continuity using Gaussian mixtures. Though one can say that template-based approach can model these continuities, in that case, we cannot update the envelope of basis and activation matrices adaptively. Many methods have been proposed to solve this problem [1, 14, 16, 18], but these methods are difficult to extend further because they do not estimate Gaussian mixture densities based on variational Bayes.

3. SIGNAL MODEL

In this section, we describe how to integrate the two promising methods: LHA and NMF. At first, we describe our idea by introducing the spectral continuity. In these methods, the wavelet spectrogram is interpreted as a histogram of sound quanta. Here, the volume of each sound is interpreted as the number of quanta. The number is determined from a Poisson distribution for each time frame. Next, the distribution of sound quanta is determined from a mixture of Gaussians. In the next subsection, we describe a straightforward observation model. This model is not suitable for the estimation, so we introduce several approximations afterward to formulate a VB-EM algorithm.

3.1 Generative Model

The volume of the k-th sound at the t-th time frame is drawn from a Poisson distribution $\mathcal{P}(S_t^k|u_t^k)$, similar to NMF. Next, we draw the number of observations of the m-th harmonic partial of the k-th sound following a multinomial distribution. The relative weight of each component is determined by η_{km}.

$$p(S_t^k) = \mathcal{P}(S_t^k|u_t^k) \tag{10}$$

$$p(S_t^{k[m]}) = \mathcal{M}(S_t^{k[m]}|S_t^k, \eta_{k[m]}) \tag{11}$$

Here, $\mathcal{M}$ denotes multinomial distribution.

Next, we draw a set of observed particles: $X_t^{km} = [x_{t1}^{km}, \cdots, x_{tS_t^{km}}^{km}]$. The value of each particle follows the Gaussian distribution of the corresponding the m-th harmonic partial. The likelihood of the observation is written as:

$$p(x_{tn}^{km}|\mu_k, \lambda_k) = \mathcal{N}(x_{tn}^{km}|\mu_k + o_m, \lambda_k^{-1}), \tag{12}$$

$$p(X_t^{km}|S_t^{km}, \mu_k, \lambda_k) = \prod_{n=1}^{S_t^{km}} \mathcal{N}(x_{tn}^{km}|\mu_k + o_m, \lambda_k^{-1}). \tag{13}$$

To generate the spectrogram of each harmonic partial, we assume the number of particles observed in spectrograms as a histogram of particles that have a value error in the range of $\epsilon_{\mathrm{F}}/2$. This corresponds with our transformation of the continuous probabilistic density functions into discrete probabilistic mass functions.

$$x_{tf}^{km} = \#\{n|x_f - \epsilon_{\mathrm{F}}/2 \le x_{tn}^{km} \le x_f + \epsilon_{\mathrm{F}}/2\} \tag{14}$$

$$x_{t\neg}^{km} = S_t^{km} - \sum_{f=1}^{F} x_{tf}^{km} \tag{15}$$

$$\hat{r}_f^{km} = \int_{x_f - \epsilon_{\mathrm{F}}/2}^{x_f + \epsilon_{\mathrm{F}}/2} \mathcal{N}(x|\mu_k + o_m, \lambda_k^{-1})dx \tag{16}$$

$$p(x_{t[f]}^{km}, x_{t\neg}^{km}) = \mathcal{M}(x_{t[f]}^{km}, x_{t\neg}^{km}|S_t^{km}, \hat{r}_{[f]}^{km}, \hat{r}_{\neg}^{km}) \tag{17}$$

Here, $\#$ denotes the number of elements in the set, x_{tf}^{km} denotes the number of particles allocated the f-th frequency bin, $x_{t\neg}^{km}$ denotes the number of particles which are not allocated any frequency bin, and $\hat{r}_f^{km}$ denotes the relative weight of each frequency bin. Further, $\hat{r}_f^{km}$ is approximated as:

$$\hat{r}_f^{km} \approx \epsilon_{\mathrm{F}} \mathcal{N}(x_f|\mu_k + o_m, \lambda_k^{-1}). \tag{18}$$

We denote the right hand of the equation r_f^{km}. Finally, the observed spectrogram is obtained as a summation of the all harmonic components.

$$x_{tf} = \sum_{km} x_{tf}^{km} \tag{19}$$

3.2 Approximations

For the above formulations are not appropriate for Bayesian estimation, we introduce the following approximation. The main objective is to marginalize the volume variables S_t^k and S_t^{km}. To do this, we inspect the following characteristics of Poisson distribution.

Poisson distribution gives the probability of n event observations in a unit time when the average occurrence interval is λ^{-1}. Next, we consider to distribute the observations into K classes, following the distribution: $\mathcal{M}(n_{[k]}|n, p_{[k]})\mathcal{P}(n|\lambda)$. The marginal probability of each class of the multinomial distribution follows a binomial distribution, so $p(n_k|n, p_k) = \mathrm{Bin}(n_k|n, p_k)$. Further, we assume that $p(n_k|p_k, \lambda) = \mathcal{P}(n_k|p_k\lambda)$ because the event of the k-th class occurs in an average time interval of $(p_k\lambda)^{-1}$. In the following section, we formulate a VB-EM algorithm based on this observation model.

4. BAYESIAN NONNEGATIVE HARMONIC FACTORIZATION

In this section, we describe the formulation of our model and the update procedures of Bayesian nonnegative harmonic factorization (BNHF). The probabilistic mass function of the intermediate spectrogram of the m-th harmonic partial and the k-th harmonic sound is first described. The spectrogram is generated following the activation u_t^k, which is the relative weight of each harmonic partial η_{km}. The prior distributions are selected to imitate LHA and NMF. This is formulated as follows.

$$p(s_{tf}^{km}|u_t^k, \eta_k, \mu_k, \lambda_k) \approx \mathcal{P}(s_{tf}^{km}|\epsilon_{\mathrm{F}} u_t^k \eta_{km} \mathcal{N}(x_f|\mu_k + o_m, \lambda_k^{-1})) \quad (20)$$

$$p(u_t^k) = \mathrm{Gam}(u_t^k|a_0, b_0) \quad (21)$$

$$p(\eta_k) = \mathrm{Dir}(\eta_k|\alpha_m^0) \propto \prod_{m=1}^{M} \eta_{km}^{\alpha_m^0 - 1} \quad (22)$$

$$p(\mu_k, \lambda_k) = \mathcal{N}(\mu_k|m_0, (\beta_0\lambda_k)^{-1})\mathcal{W}(\lambda_k|w_0, \nu_0) \quad (23)$$

Here, $\mathcal{W}$ denotes Wishart distribution and a_0, b_0, α_m^0, m_0, β_0, w_0, and ν_0 are the hyperparameters. The prior distributions are not conjugate, and thus the analytic variational Bayesian inference of the posterior distributions is intractable. Instead, we will follow a limit that $\epsilon_{\mathrm{F}} \to 0$. Under this condition, the posterior distributions are written in an approximately conjugate form. First, we assume the following factorization.

$$q(S, u, \eta, \mu, \lambda) = q(S) \prod_{tk} q(u_t^k) \prod_{k=1}^{K} \{q(\eta_k) q(\mu_k, \lambda_k)\} \quad (24)$$

This is known as mean-field approximation.

4.1 VB-E Step

During the VB-E step, we calculate the temporal estimation of separated source spectrogram s_{tf}^{km}.

$$\begin{aligned} \ln q^*(s_{tf}^{[km]}) &= \ln(X|S) + \mathbb{E}[p(S|u, \eta, \mu, \lambda)] \\ &= \ln \delta(x_{tf} - \sum_{km} s_{tf}^{km}) + \mathbb{E}[\ln u_t^k + \ln \eta_{km} \\ &\quad + \ln \mathcal{N}(x_f|\mu_k + o_m, \lambda_k^{-1})] \end{aligned} \quad (25)$$

Hereafter, all constant variables that do not affect the inference are omitted. The optimal posterior distribution is a multinomial distribution.

$$q(s_{tf}^{[km]}) = \mathcal{M}(s_{tf}^{[km]}|x_{tf}, \gamma_{tf}^{[km]}) \quad (26)$$

$$\ln \tilde{\gamma}_{tf}^{km} = \mathbb{E}[\ln u_t^k + \ln \eta_{km} + \ln \mathcal{N}(x_f|\mu_k + o_m, \lambda_k^{-1})] \quad (27)$$

$$\gamma_{tf}^{km} = \frac{\tilde{\gamma}_{tf}^{km}}{\sum_{k'm'} \tilde{\gamma}_{tf}^{k'm'}} \quad (28)$$

Here, $\mathcal{M}$ denotes multinomial distribution.

4.2 VB-M Step

During the VB-M step, we update the posterior distributions of u_t^k, η_{km}, μ_k, and λ_k. For example, we describe the Bayesian estimation of u_t^k in detail. The logarithm of the optimal posterior distribution is written as:

$$\begin{aligned} \ln q^*(u_t^k) &= \mathbb{E}_S[\ln p(S|u, \eta, \mu, \lambda)] + \ln p(u_t^k) \\ &= \sum_{fm} \mathbb{E}[s_{tf}^{km}] \ln u_t^k + (a_0 - 1) \ln u_t^k \\ &\quad - \sum_{fm} \epsilon_{\mathrm{F}} u_t^k \eta_{km} \mathcal{N}(x_f|\mu_k + o_m, \lambda_k^{-1}). \end{aligned} \quad (29)$$

Taking the limit $\epsilon_{\mathrm{F}} \to 0$, we obtain the following update:

$$q^*(u_t^k) \approx \mathrm{Gam}(u_t^k|a_t^k, b_0), \text{where} \quad (30)$$

$$a_t^k = a_0 + \sum_{fm} \mathbb{E}[s_{tf}^{km}]. \quad (31)$$

The same is true for η_{km}, μ_k, and λ_k: the optimal posterior distributions have the conjugate form when we take the limit $\epsilon_{\mathrm{F}} \to 0$. The approximated posterior distributions are written as:

$$q^*(\eta_k) \approx \mathrm{Dir}(\eta_k|\alpha_k), \quad (32)$$

$$q^*(\mu_k, \lambda_k) \approx \mathcal{N}(\mu_k|m_k, (\beta_k\lambda_k)^{-1})\mathcal{W}(\lambda_k|w_k, \nu_k), \quad (33)$$

where the posterior hyperparameters are written as:

$$\alpha_{km} = \alpha_m^0 + \sum_{tf} \mathbb{E}[s_{tf}^{km}], \quad (34)$$

$$m_k = \frac{m_0\beta_0 + \sum_{tfm} \mathbb{E}[s_{tf}^{km}](x_f - o_m)}{\beta_0 + \sum_{tfm} \mathbb{E}[s_{tf}^{km}]}, \quad (35)$$

$$\beta_k = \beta_0 + \sum_{tfm} \mathbb{E}[s_{tf}^{km}], \quad (36)$$

$$w_k^{-1} = w_0^{-1} + \beta_0 m_0^2 + \sum_{tfm} \mathbb{E}[s_{tf}^{km}](x_f - o_m)^2 - \beta_k m_k^2, \quad (37)$$

$$\nu_k = \nu_0 + \sum_{tf} \mathbb{E}[s_{tf}^{km}]. \quad (38)$$

The update equation of BNHF is quite similar to that of LHA, and these two methods had exactly the same result in our experiment. The difference is that our model explicitly models the volume of each sound, which makes it easier to consider the temporal continuity. This is described in more detail in the next section. We can also formulate the Gibbs sampler by using a similar approximation in which the space complexity is of the order $O(TFKM)$, instead of the $O(TNKM)$ for LHA. The derivations are omitted due to space restrictions.

5. BAYESIAN NONNEGATIVE HARMONIC-TEMPORAL FACTORIZATION

Here, we describe how to introduce temporal continuity to BNHF. The temporal structure is modeled using a mixture of Gaussians arranged at regular intervals. The intensity of

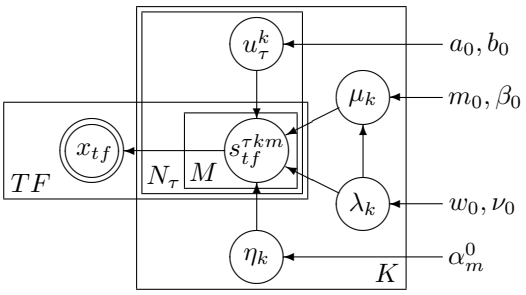

Figure 2. Graphical model of proposed method. Single solid lines indicate latent variables and double solid lines indicate observed variables.

each Gaussian component is decided using a Gamma distribution. Let T be the interval, τ be the index of Gaussian components, λ_{T} be the precision of each Gaussian component, and u_τ^k be the intensity of τ-th Gaussian component of k-th sound. The joint distribution can be written as $p(X, S, u, \eta, \mu, \lambda)$. A graphical model of the method is shown in Figure 2.

The conditional probability of $s_{tf}^{\tau km}$ is:

$$p(s_{tf}^{\tau km}|u,\eta,\mu,\tau) = \mathcal{P}(s_{tf}^{\tau km}|\epsilon_{\mathrm{T}} u_\tau^k \mathcal{N}(y_t|\tau T, \lambda_{\mathrm{T}}^{-1}) \times \epsilon_{\mathrm{F}}\eta_{km}\mathcal{N}(x_f|\mu_k + o_m, \lambda_k^{-1})), \quad (39)$$

where y_t is the temporal position of the t-th time frame. The corresponding optimal posterior distribution is:

$$q(s_{tf}^{[\tau km]}) = \mathcal{M}(s_{tf}^{[\tau km]}|x_{tf}, \gamma_{tf}^{[\tau km]}), \quad (40)$$

$$\gamma_{tf}^{\tau km} \propto \exp(\mathbb{E}[\ln u_\tau^k + \ln \eta_{km} + \ln \mathcal{N}(x_f|\mu_k + o_m, \lambda_k^{-1})] + \ln \mathcal{N}(t|\tau T, \lambda_{\mathrm{T}}^{-1})) \quad (41)$$

Further, the optimal posterior distribution of u_τ^k is approximated as:

$$q(u_\tau^k) \approx \mathrm{Gam}(u_\tau^k|a_\tau^k, b_0), \text{where} \quad (42)$$

$$a_\tau^k = a_0 + \sum_{tfm} \mathbb{E}[s_{tf}^{\tau km}]. \quad (43)$$

6. EVALUATION

In this section, we compare the performance of three multi-pitch estimation methods: LHA, NHF, and NHTF. We then discuss their performance in detail.

6.1 Estimation Target

For the experiment, we used 40 musical pieces from the RWC Music Database [6]. These included five piano solo pieces (RM-J001 to RM-J005), five guitar solo pieces (RM-J006 to RM-J010), ten jazz duo pieces (RM-J011 to RM-J020), ten jazz pieces played with three or more players (RM-J021 to RM-J030), and ten classical chamber pieces (RM-C012 to RM-C021). All the pieces were recorded from MIDI files using a MIDI synthesizer (Yamaha MOTIF-XS.) The drum tracks were muted, and the number of players was counted without including the drum player. The recorded signals were truncated to the first 32 seconds to reduce the large computational time needed for the experiment. They were transformed into wavelet spectrograms using Gabor wavelets with a time resolution of 16 [msec], frequency bins from 30 to 3000 [Hz], and a frequency resolution of 12 [cents]. The ground truths were constructed using the reference MIDI files.

Table 1. Calculated F-measures.

	Non-infomative			Infomative		
Music Type	LHA	BNHF	BNHTF	LHA	BNHF	BNHTF
Piano Solo	0.558	0.558	0.590	0.584	0.584	**0.590**
Guitar Solo	0.684	0.684	0.726	0.728	0.728	**0.740**
Jazz (Duo)	0.524	0.524	0.545	0.552	0.552	**0.556**
Jazz (Trio∼)	0.523	0.523	**0.548**	0.536	0.536	0.541
Chamber	0.481	0.481	0.508	0.503	0.503	**0.512**

6.2 Experimental Settings

Here, we describe the experimental settings. For the estimation, two prior distribution settings were evaluated. In the first one, all priors were set to be non-informative. That is, $a_0, b_0, \alpha_m^0, \beta_0, w_0$, and ν_0 were set to unity and m_0 was set to zero. In the other one, the prior distribution of the harmonic structure was set appropriately. That is, $\alpha_m^0 = 0.6547Nm^{-2}$, where N is the number of total observations. The other hyperparameters were set to be non-informative. The setting of α_m^0 was the same setting as HTC.

As an initialization, the relative weight of the source model in the t-th frame was set to the sum of the amplitudes of the nearest frequency bins of its overtones. The initial overtone weights were set to decay exponentially.

Model orders K and M were set to 73 and 6, respectively, where the number of overtones M is equal to HTC [7]. The EM algorithm was truncated at 200 iterations. The number of iterations was determined experimentally on the basis of estimation accuracy saturation.

After the iterations, the estimated pitches were extracted from the posterior hyperparameters. Let r be the threshold. The effective observation count of the k-th basis in the t-th time frame, N_{tk}, satisfies $N_{tk} \geq r \max_{tk} N_{tk}$, is considered to be sounding. The threshold was optimized experimentally for each piece and the method used to evaluate the potential performance of each model. The model frequencies were allocated the nearest note number. The estimated result and the ground truth were transformed into $T \times 128$ binary maps, and the F-measure was then calculated. Let N be the number of entries in the estimated map, C be the number of entries in the truth map, and R be the number of correct entries in the estimated map. The resultant F-measure is calculated as $F = 2R/(N + C)$. The larger the value of the F-measure, the better the performance.

6.3 Results

The results are shown in Table 1. Our method outperforms the conventional method for all the music type and both prior settings. The highest performance was attained with

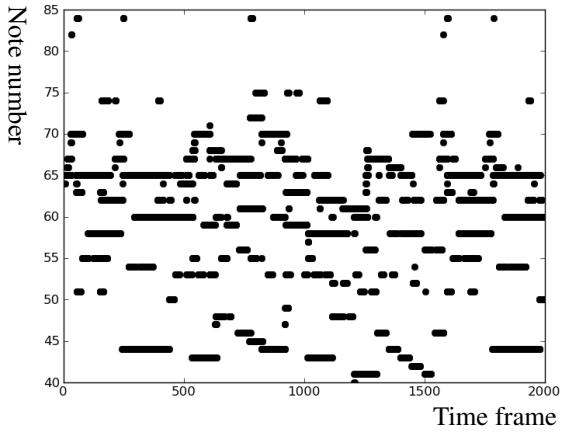

Figure 3. Example of estimated result of the proposed method with musical piece *RM-J007*.

appropriate prior for four of the five groups of musical pieces and with non-informative for the remaining group. This indicates the importance of the joint estimation of the harmonic and temporal structures. Figure 3 illustrates an example of the estimated result of the proposed model.

7. RELATED WORKS

Our method have a strong relation to the model proposed by Ochiai *et., al* [12,13]. Ours and theirs are notably different in three points. Firstly, they do not explicitly model the spectral continuity. Secondly, they use Dirac delta function instead of the ϵ-neighborhood to achieve the estimation of Gaussian mixture distribution. By following this strategy, the estimation is in conjugate form, but the interpretation of the model becomes difficult. Thirdly, they model the temporal structure based on the joint estimation of probabilistic context-free grammar and Gaussian mixture distribution.

8. CONCLUSION

We presented a new multipitch analyzer based on variational Bayes that explicitly models the harmonic and temporal structures separately based on Gaussian mixture model. Several priors were set to be not conjugate, in order to integrate NMF and LHA. The variational posterior distributions become conjugate under several approximations. Our method can be viewed as a Bayesian NMF method with adaptive harmonic basis. Evaluation results showed that the proposed method outperform the conventional multipitch analyzer, latent harmonic allocation (LHA). In the future, we will propose a more precise modeling using recent improvements to NMF, including the source-filter model [17] and nonparametric models [11].
This research is partially supported by KAKENHI (S) No. 24220006.

9. REFERENCES

[1] A. Bertin et al. Enforcing harmonicity and smoothness in bayesian non-negative matrix factorization applied to polyphonic music transcripntion. *IEEE Trans. on ASLP*, 18(3):538–549, 2010.

[2] B. A. Casey et al. Content-based music information retrieval: Current directions and future challenges. *Proc. of the IEEE*, 96(4):668–696, 2008.

[3] A. T. Cemgil. Bayesian inference for nonnegative matrix factorization models. *Technical Report CUED/F-INFENG/TR.609*, 2008.

[4] D. FitzGerald et al. On the use of the beta divergence for musical source separation. In *Proc. ISSC*, pages 1–6, 2010.

[5] M. Goto. A real-time music-scene-analysis system: Predominant-F0 estimation for detecting melody and bass lines in real-world audio signals. *Speech Communication*, 43(4):311–329, September 2004.

[6] M. Goto et al. RWC music database: Popular, classical, and jazz music databases. In *Proc. ISMIR*, pages 287–288, 2002.

[7] H. Kameoka et al. A multipitch analyzer based on harmonic temporal structured clustering. *IEEE Trans. on ASLP*, 15(3):982–994, 2007.

[8] Y. E. Kim et al. Music emotion recognition: A state of the art review. In *Proc. ISMIR*, pages 255–266, 2010.

[9] T. Kitahara et al. Instrument identification in polyphonic music: feature weighting to minimize influence of sound overlaps. *EURASIP J. Appl. Signal Process.*, 2007:1–15, 2007.

[10] A. Klapuri. Multipitch analysis of polyphonic music and speech signals using an auditory model. *IEEE Trans. on ASLP*, 16(2):255–266, 2008.

[11] M. Nakano et al. Nonparametric Bayesian music parser. In *Proc. ICASSP*, pages 461–464, 2012.

[12] K. Ochiai et al. Explicit beat structure modeling for nonnegative matrix factorization-based multipitch analysis. In *Proc. ICASSP*, pages 133–136, 2012.

[13] K. Ochiai et al. Hierarchical Bayesian modeling of the generating process of music signals for automatic transcription. In *ASJ Spring Meeting*, 2012 (in Japanese).

[14] S. A. Raczyński et al. Multipitch analysis with harmonic nonnegative matrix approximation. In *Proc. ISMIR*, pages 381–386, 2007.

[15] Y. Ueda et al. HMM-based approach for automatic chord detection using refined acoustic features. In *Proc. ICASSP*, pages 5518–5521, 2010.

[16] E. Vincent et al. Adaptive harmonic spectral decomposition for multiple pitch estimation. *IEEE Trans. on ASLP*, 18(3):528–537, 2010.

[17] T. Virtanen et al. Analysis of polyphonic audio using source-filter model and nonnegative matrix factorization. In *Advances in Models for Acoustic Processing*, 2006.

[18] N. Yasuraoka et al. I-Divergence-based dereverberation method with auxiliary function approach. In *Proc. ICASSP*, pages 369–372, 2011.

[19] K. Yoshii and M. Goto. A nonparametric Bayesian multipitch analyzer based on infinite latent harmonic allocation. *IEEE Trans. on ASLP*, 20(3):717–730, 2012.