INITIALIZATION-ROBUST BAYESIAN MULTIPITCH ANALYZER BASED ON PSYCHOACOUSTICAL AND MUSICAL CRITERIA

Daichi Sakaue, Takuma Otsuka, Katsutoshi Itoyama, Hiroshi G. Okuno

Graduate School of Informatics, Kyoto University, Japan

ABSTRACT

We present a new Bayesian multipitch analyzer that dispenses with a precise optimization of parameter initialization or hyperparameters. Our method uses a new family of prior distribution, *characteristic prior*; it efficiently restricts the existence region of the latent variables, that is, the product of a conjugate prior and a characteristic function. The update formulas become a simple form that is actually suitable for Gibbs sampling. We construct characteristic priors of harmonic structures based on psychoacoustical and musical knowledge and apply them to nonnegative harmonic factorization. Experimental results improve 5.2 points in F-measure under a tough condition, random initialization with no hyperparameter optimization.

Index Terms— multipitch estimation, nonnegative matrix factorization, harmonic clustering, overtone corpus, Bayesian analysis

1. INTRODUCTION

Multipitch analysis [1–13] is one of the most appropriate music signal processing techniques today because it enables precise, higher-level analysis and manipulations, including music structure visualization [14], musical instrument identification [15], instrument-based equalization [16], and musical signal manipulation [17]. Recent Bayesian methods [3, 4] are particulary important because they can explicitly model the relationship between multiple pitch activities and other musical features, such as musical instrument, chord, and onset [18, 19]. Although further improvement of multipitch estimation is highly demanded for actual applications, it remains still a challenging problem due to its optimization difficulty.

The main objective of Bayesian multipitch analyzers is to estimate the most likely combination of latent variables, including pitch, volume, and timber. This can be difficult because there are a number of inappropriate combinations that superficially describe the observed spectrogram well. For example, a harmonic sound with only its fifth overtone is quite rarely included in a musical piece, but it is very likely to be estimated by untrained multipitch analyzers. To avoid such errors, many authors have proposed precise heuristic initialization [2, 5] and parametric [2, 4] and nonparametric [3] optimizations of the prior distributions. These techniques have worked well to a certain extent, but further refinement is difficult because their optimization requires expensive procedures such as cross-validation. One candidate to overcome the optimization problem is an initialization-robust method with no hyperparameter optimization, and so our research has been focused in this direction [5].

In this paper, we present a new method of optimizing the prior distribution based on psychoacoustical and musical knowledge. Our method uses a new family of prior distribution, *characteristic prior*, that forces each latent variable to exist in a desirable range. The distributions are the product of a non-informative conjugate prior and a characteristic function. All we need to do is determine the desirable range of each latent variable and simply represent it as the characteristic function. This reduces the range of latent variables in searching for an optimal solution. Our update formula takes a simple form similar to conjugate variational Bayesian methods. We applied these priors to nonnegative harmonic factorization (NHF) [20] and confirmed an average F-measure improvement of 5.2 points against random initialization.

2. CONVENTIONAL METHODS

In this section, we briefly discuss two conventional methods: Bayesian nonnegative matrix factorization (NMF) [21] and Bayesian nonnegative harmonic factorization (NHF) [20]. These two methods constitute the basis of the proposed method. In the following, $\mathcal{N}, \mathcal{W}, \mathcal{P}, \mathcal{M}, \mathcal{G}$, and $\mathcal{D}$ denotes normal, Wishart, Poisson, multinomial, gamma, and Dirichlet distributions, respectively. Further, t, f, k, and m denotes index of time frame, log-frequency bin, basis, and overtone, respectively, and are numbered T, F, K, and M. A bracket set [ ] denotes a set or vector over the index contained within.

2.1. Bayesian Nonnegative Matrix Factorization

An audio signal is composed of a small number of frequent sounds, e.g., the C4 of a violin, the E5 of a piano, and the A5 of a flute. The aim of NMF is to extract these template sounds and their temporal activities. In their time-frequency representation, the timber of a sound is visualized as a constant spectral pattern called 'basis'. NMF assumes that each time frame spectrum of the observed spectrogram $Y_{t[f]}$ is the linear combination of K bases. This is represented as $Y_{tf} \approx \sum_k u_t^k h_f^k$, where u_t^k represents the volume of the k-th basis at the t-th frame and $h_{[f]}^k$ represents the spectrum of the basis. u_t^k and h_f^k are learned by minimizing the cost function $\sum_{tf} D(Y_{tf} || \sum_k u_t^k h_f^k)$. Here, D is a measure of difference, and usually Kullback-Leibrer (KL) or Itakura-Saito (IS) divergences are selected. A Bayesian counterpart of KL-NMF has been proposed by Cemgil [21]. Its likelihood function is

$$p(Y_{tf}|X_{tf}^{[k]}) = \delta\left(Y_{tf} - \sum_{k=1}^{K} X_{tf}^k\right), \tag{1}$$

$$p(X_{tf}^k) = \mathcal{P}(X_{tf}^k|u_t^k h_f^k), \tag{2}$$

where X_{tf}^k represents the hidden spectrogram generated by the k-th basis. To perform variational Bayesian estimation, a conjugate prior distribution of the model parameters is assumed:

$$p(u_t^k) = \mathcal{G}(u_t^k|a_0, b_0), \qquad p(h_f^k) = \mathcal{G}(h_f^k|a_0, b_0). \tag{3}$$

Since a standard NMF separately formulates multiple bins of a basis, its spectral shape is not limited. This is sometimes inconvenient because one basis does not always correspond to one harmonic

sound nor have the explicit parameter of a fundamental frequency. Overall, it makes multipitch analysis especially difficult. As an alternative, we represent the basis using a Gaussian mixture model (GMM) that represents a harmonic structure, as

$$h_f^k \propto \sum_{m=1}^{M} \tau_{km} \mathcal{N}(x_f | \mu_k + o_m, \lambda_k^{-1}), \tag{4}$$

where x_f denotes the log-frequency of the f-th bin, μ_k denotes the k-th fundamental frequency, λ_k denotes the precision of the harmonic components, o_m denotes the offset of the m-th component, and τ_{km} denotes the relative weight of the m-th component. The adaptive estimation of these parameters enables an accurate estimation. This attempt has previously been successful using Bayesian nonnegative harmonic factorization (NHF) [20].

2.2. Bayesian Nonnegative Harmonic Factorization

Bayesian NHF extends Bayesian NMF to represent its basis with a Gaussian mixture. Since Bayesian NMF uses Poisson distributions to represent a quantized observed spectrogram, each basis at each time frame is assumed to generate a discrete number of observation energy, $\sum_f X_{tf}^k$. Our method denotes the energy with S_t^k and represents it using a Poisson distribution. Further, the quantized energy is assumed to distribute on the log-frequency axis based on a corresponding GMM distribution. The observation likelihood of an energy quantum is

$$p(x | \tau_k, \mu_k, \lambda_k) = \sum_{m=1}^{M} \tau_{km} \mathcal{N}(x | \mu_k + o_m, \lambda_k^{-1}). \tag{5}$$

Next, we assume the observed spectrogram is a histogram of the particles. The f-th frequency bin counts the number of the quanta with a frequency of $x_f - \epsilon/2 \leq x \leq x_f + \epsilon/2$. The likelihood function of the k-th basis and the m-th component at the t-th frame and the f-th bin are derived as

$$\begin{aligned} p(s_{tf}^{km} | u_t^k, \tau_k, \mu_k, \lambda_k) &= \sum_{S_t^k=0}^{\infty} p(s_{tf}^{km} | S_t^k, \tau_k, \mu_k, \lambda_k) p(S_t^k | u_t^k) \\ &\approx \mathcal{P}(s_{tf}^{km} | \epsilon u_t^k \tau_{km} \mathcal{N}(x_f | \mu_k, \lambda_k^{-1})). \end{aligned} \tag{6}$$

This is known as the *thinning* of a Poisson distribution [22]. We perform a full Bayesian estimation by constructing the joint model with the following prior distributions:

$$p(Y_{tf} | s_{tf}^{[km]}) = \delta \left(Y_{tf} - \sum_{km} s_{tf}^{km} \right), \tag{7}$$

$$p(u_t^k) = \mathcal{G}(u_t^k | a_0, b_0), \quad p(\tau_k) = \mathcal{D}(\tau_k | \alpha_0), \tag{8}$$

$$p(\mu_k, \lambda_k) = \mathcal{N}(\mu_k | m_k, (\beta_0 \lambda_k)^{-1}) \mathcal{W}(\lambda_k | w_0, \nu_0). \tag{9}$$

2.3. Derivation of Update Equations

Although Dirichlet and normal-Wishart distributions are not conjugate of a Poisson distribution, their posterior distribution become a conjugate form when we take a limit $\epsilon \to 0$. This makes it easy to estimate the latent variables using variational Bayes or Gibbs sampling. If we use variational Bayes, the optimal variational posterior distribution of u_t^k becomes

$$q^*(u_t^k) = \mathcal{G}(u_t^k | a_t^k, b_0), \qquad a_t^k = a_0 + \sum_{fm} \mathbb{E}[s_{tf}^{km}]. \tag{10}$$

The posterior distributions of s_{tf}^{km}, τ_k, μ_k, and λ_k are obtained in a similar manner.

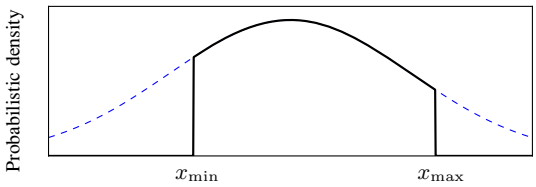

Fig. 1. Example of characteristic prior. Blue dashed line indicates normal distribution and black solid line indicates characteristic normal distribution. Corresponding latent variable is effectively captured as $x_{\min} \leq x \leq x_{\max}$.

Since the form of update equations is the same as those used for LHA, the performances of the methods using variational Bayes are strictly equal [20]. However, our method is notable for the following two reasons.

1. Our method can easily represent the inter-frame dynamics of volume. This is because it explicitly models the absolute volume of a basis using u_t^k, while LHA models the relative volume of a basis using $\pi_{t[k]} \sim \mathcal{D}(\alpha_{t[k]})$ to the total volume at each frame. We have previously introduced a GMM structure for the temporal dynamics in respect to HTC [2] and verified the estimation accuracy improvement [20].
2. The space complexity of the latent variables in LHA is $O(NKM)$, where N is the number of energy quanta. This prevents us from using Gibbs sampling and the rich variety of non-conjugate models. Since the complexity remains $O(TFKM)$ in our model, it allows us to try a new family of prior distributions, *characteristic priors*. This is the focus of the next section.

3. CHARACTERISTIC PRIOR

3.1. Definition

The aim of Bayesian analysis is the joint estimation of latent variables. For GMM-based multipitch analyzers, including PreFEst [1], HTC [2], LHA [3], and NHF [20], the main objective is that of the volume u_t^k, timber τ_k, pitch μ_k, and component precision λ_k. Sometimes the estimated results of the parameters are obviously wrong because they do not forbid parameters from having inappropriate values. For example, since the aim of harmonic clustering is to capture the sharp peak of the harmonic components, the precision parameter λ_k should be larger than a certain threshold.

We propose a new family of prior distribution to force each latent variable to exist in a desirable range. This is called *characteristic prior*, which is the product of a conjugate prior and a characteristic function. A characteristic function χ assigns 0 or 1 for each entry of the domain, A, and then specifies a subset B:

$$\chi(x) = \begin{cases} 1 & (x \in B) \\ 0 & (\text{otherwise}) \end{cases}. \tag{11}$$

Hereafter, we represent the subset B using the same symbol of the function χ. A characteristic Dirichlet distribution $\mathcal{D}^*$ is defined as

$$\mathcal{D}^*(x | \alpha, \chi) = \frac{\mathcal{D}(x | \alpha)}{\int_\chi \mathcal{D}(x' | \alpha) dx'} \quad (x \in \chi), \tag{12}$$

where $\int_\chi \mathcal{D}(x' | \alpha) dx'$ is a normalization constant. Characteristic normal distribution $\mathcal{N}^*$ and characteristic Wishart distribution $\mathcal{W}^*$ are defined in a similar manner. An example of the characteristic distribution is shown in Fig. 1.

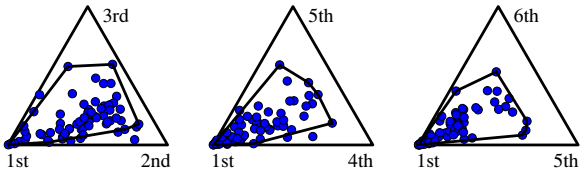

Fig. 2. Relative overtone weights of MIDI instruments at A4 (440 Hz). Solid lines indicate convex hull of characteristic function.

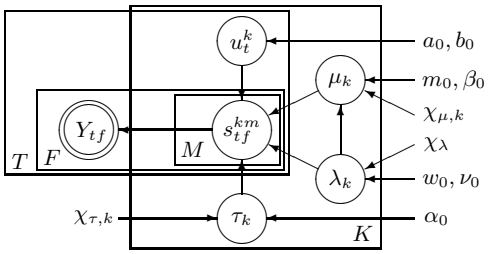

Fig. 3. Graphical model of proposed method. Single solid lines indicate latent variables and double solid line indicates observed one.

3.2. Prior Distributions of Pitch, Shape, and Timber

Here, we determine a desirable range for the NHF parameters. First, we examine the pitch μ_k. Since we hope one basis will correspond to one note number, we set the characteristic function that forces each basis to appear in the semitone range:

$$\chi_{\mu,k}(x) = \begin{cases} 1 & (\hat{\mu}_k - 50 \leq x \leq \hat{\mu}_k + 50) \\ 0 & (\text{otherwise}) \end{cases}. \quad (13)$$

We set the basis number as $K = 73$ to cover six octaves. The constant $\hat{\mu}_k$ corresponds to the log-frequency of the k-th note number. The characteristic function of λ_k is also defined so that the standard deviation of any harmonic component does not exceed 100 [cents]:

$$\chi_\lambda(x) = \begin{cases} 1 & (x \geq 1/10000[\text{cents}^{-2}]) \\ 0 & (\text{otherwise}) \end{cases}. \quad (14)$$

The characteristic function of harmonic structure is determined by psychoacoustic basis, the basic idea of which has been described in our previous paper [5]. Essentially, the excessive weight of the upper overtones is psychoacoustically inappropriate and causes a virtual pitch [23] of the overtone frequency. To avoid this type of incorrect estimation, we set the upper bound of overtone weights by examining individual note sounds generated using a MIDI synthesizer (Fig. 2). The complete procedure will be described later, in the evaluation section. Hereafter, we denote the convex hull for the k-th basis using $\chi_{\tau,k}$.

4. CHARACTERISTIC NONNEGATIVE HARMONIC FACTORIZATION

As a composition of NHF and characteristic priors, we formulated a new Bayesian multipitch analyzer, characteristic NHF (CNHF). The prior distribution of NHF is modified as

$$p(u_t^k) = \mathcal{G}(u_t^k | a_0, b_0), \quad (15)$$

$$p(\tau_k) = \mathcal{D}^*(\tau_{k[m]} | \alpha_0, \chi_{\tau,k}), \quad (16)$$

$$p(\mu_k | \lambda_k) = \mathcal{N}^*(\mu_k | m_0, (\beta_0 \lambda_k)^{-1}, \chi_{\mu,k}), \quad (17)$$

$$p(\lambda_k) = \mathcal{W}^*(\lambda_k | w_0, \nu_0, \chi_\lambda). \quad (18)$$

For the volume parameter u_t^k, we retain the original prior distribution because it is difficult to specify a desirable range for it.

4.1. Inference with Gibbs Sampling

Since the proposed method assumes non-conjugate priors, it is difficult to use deterministic procedures such as variational Bayes. Instead, we use Gibbs sampling to update the latent variables.

The latent variables are iteratively drawn from their conditional posterior distributions. For example, the posterior probability of s_{tf}^{km} is described as

$$p(s_{tf}^{[km]} | Y, S_{\neg tf[km]}, u, \pi, \mu, \lambda) = \mathcal{M}(s_{tf}^{[km]} | Y_{tf}, \rho_{tf}^{[km]}), \quad (19)$$

$$\tilde{\rho}_{tf}^{km} = u_t^k \tau_{km} \mathcal{N}(x_f | \mu_k + o_m, \lambda_k^{-1}), \quad (20)$$

$$\rho_{tf}^{km} = \frac{\tilde{\rho}_{tf}^{km}}{\sum_{k'm'} \tilde{\rho}_{tf}^{k'm'}}, \quad (21)$$

where ρ_{tf}^{km} is the parameter for updating s_{tf}^{km}. The posterior distribution of u_t^k, τ_k, μ_k, and λ_k are obtained by taking a limit $\epsilon \to 0$:

$$p(u_t^k | Y, S, u_{\neg kt}, \tau, \mu, \lambda) \approx \mathcal{G}(u_t^k | a_t^k, b_0), \quad (22)$$

$$p(\tau_k | Y, S, u, \tau_{\neg k}, \mu, \lambda) \approx \mathcal{D}^*(\tau_k | \alpha_{k[m]}, \chi_{\tau,k}), \quad (23)$$

$$p(\mu_k | Y, S, u, \tau, \mu_{\neg k}, \lambda) \approx \mathcal{N}^*(\mu_k | m_k, (\beta_k \lambda_k)^{-1}, \chi_{\mu,k}), \quad (24)$$

$$p(\lambda_k | Y, S, u, \tau, \mu, \lambda_{\neg k}) \approx \mathcal{W}^*(\lambda_k | w_k, \nu_k, \chi_\lambda). \quad (25)$$

The posterior hyperparameters $a_t^k, \alpha_{km}, m_k, \beta_k, w_k$, and ν_k are represented as

$$a_t^k = a_0 + \sum_{fm} s_{tf}^{km}, \qquad \alpha_{km} = \alpha_0 + \sum_{tf} s_{tf}^{km}, \quad (26)$$

$$\beta_k = \beta_0 + \sum_{tfm} s_{tf}^{km}, \qquad \nu_k = \nu_0 + \sum_{tfm} s_{tf}^{km}, \quad (27)$$

$$m_k = \frac{m_0 \beta_0 + \sum_{tfm} s_{tf}^{km} (x_f - o_m)}{\beta_0 + \sum_{tfm} s_{tf}^{km}}, \quad (28)$$

$$w_k^{-1} = w_0^{-1} + \beta_0 m_0^2 + \sum_{tfm} s_{tf}^{km} (x_f - o_m)^2 - \beta_k m_k^2. \quad (29)$$

The graphical model of the proposed method is shown in Fig. 3.

4.2. Implementation Issue

During the estimation, it is required to draw samples from characteristic Dirichlet, normal, and Wishart distributions. One possible and strict algorithm is to draw samples from corresponding standard distributions and reject ones that are not included in the characteristic region. The problem with this algorithm is the unbounded number of rejections. For example, the expected number of drawing is $(\int_\chi \mathcal{D}(x'|\alpha)dx')^{-1}$ for a characteristic Dirichlet distribution, which may become an intractably large number. Instead, we use the following approximate sampling algorithm.

Characteristic normal and Wishart distributions We first try the strict algorithm described above with 100 samples and use the first appropriate one if any sample within the range has been sampled. Otherwise, the value of the last sample is rounded to the appropriate range. This optimization is similar to the iterative conditional mode (ICM) [24].

Characteristic Dirichlet distribution In this case, sampling is performed within a high-dimensional space. This makes the sampling more difficult. The strict algorithm is performed with 100 samples and afterward an approximated algorithm is evaluated. One possible approximation is a Metropolis-Hasting algorithm using some

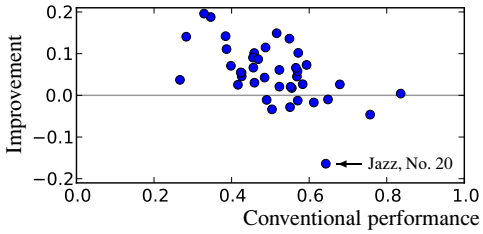

Fig. 4. Performance comparison of conventional and proposed methods. Horizontal axis shows conventional performance ($\mathrm{F}_{\mathrm{OCLHA}}$), vertical axis shows improvement ($\mathrm{F}_{\mathrm{CNHF}} - \mathrm{F}_{\mathrm{OCLHA}}$), and each dot indicates one musical piece.

proposal distributions, such as normal or Cauchy, but it does not always work properly because an appropriate proposal distribution is difficult to determine due to the complex shape of the convex hull. Instead, we try the Metropolis algorithm with 200 iterations with the proposal of uniform distribution over the convex hull and use the last sample. This is performed using Delaunay triangulation and the linear projection of the samples from a Dirichlet distribution $\mathcal{D}(x|1)$. We used the Qhull library for this procedure [25].

5. EVALUATION

To evaluate the performance of the proposed method with no initialization or parameter optimization, we conducted a multipitch estimation experiment with three multipitch analyzers: LHA [3], which is a conventional Bayesian method, OC-LHA [5], which is our previous method that is robust against initialization and parameter optimization, and the proposed method.

5.1. Harmonic Region

As the first step of the experiment, we determined the characteristic function of the harmonic structure. For the 73 note numbers, we recorded the musical instrument sounds of program 1 to 80 for one second using a MIDI synthesizer (Roland SD-80). The signals were then transformed into wavelet spectrograms. Since some recorded sounds are inappropriate for retrieving the overtone weight, we removed them using two criteria. First, sounds that had more than 50% energy on their inharmonic part were removed. To do this, we integrated the amplitude spectrogram over its overtone frequencies (6 overtones and $\pm$ 100 [cents]) and compared it with the total amount. Second, we calculated their virtual pitch using subharmonic summation (SHS) [23] and retained ones whose virtual pitch corresponded to the correct note number. Further, the spectrograms are integrated over the overtone frequency bands to obtain an overtone vector. The redundant vertices were efficiently reduced by using an approximation algorithm [5].

5.2. Experimental Data

For the experiment, we used 40 musical pieces from the RWC Music Database [26]. Five piano solo pieces (Jazz, No. 1–5), five guitar solo pieces (Jazz, No. 6–10), ten jazz duo pieces (Jazz, No. 11–20), and ten jazz pieces with more than three players (Jazz, No. 21–30) were excerpted. The database includes both MIDI and audio versions of the pieces, but we used only the MIDI version because there is no temporal alignment between the two versions. The corresponding audio signals were recorded using a MIDI synthesizer (YAMAHA MOTIF-XS). The drum tracks were all muted and removed from the overall experiment, and the number of players excludes the drum player. For the experiment, we used only the first 32 seconds of each piece due to the heavy computational time.

Table 1. F-measure performance of three multipitch analyzers. Bold values indicate maximum performance.

Music type	LHA	OC-LHA	CNHF
Piano solo	0.339	0.563	**0.610**
Guitar solo	0.137	0.659	**0.678**
Jazz (Duo)	0.228	0.484	**0.547**
Jazz (Trio$\sim$)	0.258	0.474	**0.520**
Chamber	0.247	0.464	**0.529**

5.3. Evaluation Conditions

To evaluate the robustness against initialization, we set the latent variables so that they reflected none of their prior information. That is, we initialized the responsibility parameters of the EM algorithms with uniformly random distribution. All hyperparameters were set as non-informative to evaluate the robustness against parameter optimization. The number of iterations was set to 1000, 1000, and 100, respectively, for the three methods. This was determined experimentally due to estimation accuracy saturation.

After the iterations, we calculated the sound activity in a binary form. This was done by setting a threshold on a posterior hyperparameter, N_{tk}. Further, the indices of basis were exchanged to fit 128 MIDI note numbers. Finally, the results were represented as $T \times 128$ binary matrices. Estimation accuracy was calculated by comparison with ground truths generated from the MIDI files. The thresholds were separately optimized for each method and piece to compare the potential performance.

5.4. Results

The proposed method outperformed the conventional methods for all five data sets (Table 1). The improvement of F-measure was 5.2 points on average. Fig. 4 shows the ratio of improvement against the previous method. The fact that the improvement is big when the performance of the previous method is weak, indicates that the proposed method works fine for a wider variety of musical pieces. For one musical piece, the performance was substantially worse than the previous method (Jazz, No. 20). In this case, the sustain pedal of a piano is pressed down throughout the musical piece, and the evaluation did not work properly. If we exclude this piece, the improvement was 5.8 points on average.

6. CONCLUSION

In this paper, we presented a new Bayesian multipitch analyzer that does not require a precise optimization of parameter initialization or hyperparameters. A new family of prior distribution, *characteristic prior*, was introduced to restrict the existence region of the latent variables. The update formulas become a simple form and suitable for Gibbs sampling. We applied characteristic priors to NHF and obtained an initialization-robust multipitch analyzer. Experimental results showed a 5.2 points F-measure improvement against random initialization with no hyperparameter optimization. In future, we intend to integrate higher-level structures of musical pieces such as musical instruments and verses to improve the estimation accuracy further. This research was partially supported by KAKENHI (S) No. 24220006 and (B) No. 24700168.

7. REFERENCES

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

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

[3] K. Yoshii and M. Goto, "A nonparametric Bayesian multipitch analyzer based on infinite latent harmonic allocation," *IEEE Trans. on ASLP*, vol. 20, no. 3, pp. 717–730, 2012.

[4] H. Kameoka, K. Ochiai, M. Nakano, M. Tsuchiya, and S. Sagayama, "Context-free 2D tree structure model of musical notes for Bayesian modeling of polyphonic spectrograms," in *Proc. ISMIR*, 2012, pp. 307–312.

[5] D. Sakaue, K. Itoyama, T. Ogata, and H. G. Okuno, "Initialization-robust multipitch estimation based on latent harmonic allocation using overtone corpus," in *Proc. ICASSP*, 2012, pp. 425–428.

[6] A. Klapuri, "Multipitch analysis of polyphonic music and speech signals using an auditory model," *IEEE Trans. on ASLP*, vol. 16, no. 2, pp. 255–266, 2008.

[7] E. Vincent, N. Bertin, and R. Badeau, "Adaptive harmonic spectral decomposition for multiple pitch estimation," *IEEE Trans. on ASLP*, vol. 18, no. 3, pp. 528–537, 2010.

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

[9] S. A. Raczyński, N. Ono, and S. Sagayama, "Multipitch analysis with harmonic nonnegative matrix approximation," in *Proc. ISMIR*, 2007, pp. 381–386.

[10] S. A. Raczyński, E. Vincent, F. Bimbot, and S. Sagayama, "Multiple pitch transcription using DBN-based musicological models," in *Proc. ISMIR*, 2010, pp. 363–368.

[11] T. Tolonen and M. Karjalainen, "A computationally efficient multipitch analysis model," *IEEE Trans. on SAP*, vol. 8, no. 6, pp. 708–716, 2000.

[12] A. Koretz and J. Tabrikian, "Maximum a posteriori probability multi-pitch tracking using the harmonic model," *IEEE Trans. on ASLP*, vol. 19, no. 7, pp. 2210–2221, 2011.

[13] J. K. Nielsen, M. G. Christensen, and S. H Jensen, "Default Bayesian estimation of the fundamental frequency," *IEEE Trans. on ASLP*, vol. 21, no. 3, pp. 598–610, 2013.

[14] M. Goto, "A chorus-section detection method for musical audio signals and its application to a music listening station," *IEEE Trans. on ASLP*, vol. 14, no. 5, pp. 1783–1794, 2006.

[15] T. Kitahara, M. Goto, K. Komatani, T. Ogata, and H. G. Okuno, "Instrument identification in polyphonic music: feature weighting to minimize influence of sound overlaps," *EURASIP J. Appl. Signal Process.*, vol. 2007, pp. 1–15, 2007.

[16] K. Itoyama, M. Goto, K. Komatani, T. Ogata, and H. G. Okuno, "Query-by-example music information retrieval by score-informed source separation and remixing technologies," *EURASIP J. Adv. in Signal Process.*, vol. 2010, pp. 1–14, 2010.

[17] N. Yasuraoka, T. Abe, K. Itoyama, T. Takahashi, T. Ogata, and H. G. Okuno, "Changing timbre and phrase in existing musical performances as you like: manipulations of single part using harmonic and inharmonic models," in *Proc. ACM Multimedia*, 2009, pp. 203–212.

[18] A. Maezawa, H. G. Okuno, T. Ogata, and M. Goto, "Polyphonic audio-to-score alignment based on Bayesian latent harmonic allocation hidden Markov model," in *Proc. ICASSP*, 2011, pp. 185–188.

[19] K. Miyamoto, H. Kameoka, T. Nishimoto, N. Ono, and S. Sagayama, "Harmonic-temporal-timbral clustering (HTTC) for the analysis of multi-instrument polyphonic music signals," in *Proc. ICASSP*, 2008, pp. 113–116.

[20] D. Sakaue, T. Otsuka, K. Itoyama, and H. G. Okuno, "Bayesian nonnegative harmonic-temporal factorization and its application to multipitch analysis," in *Proc. ISMIR*, 2012, pp. 91–96.

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

[22] R. Durrett, *Essentials of Stochastic Processes*, Springer-Verlag, 2nd edition, 2012.

[23] D. J. Hermes, "Measurement of pitch by subharmonic summation," *J. Acoust. Soc. Am.*, vol. 83, no. 1, pp. 257–264, 1988.

[24] J. Kittler and J. Föglein, "Contextual classification of multispectral pixel data," *Image and Vision Computing*, vol. 2, no. 1, pp. 13–29, 1984.

[25] C. B. Barber, D. P. Dobkin, and H. Huhdanpaa, "The quickhull algorithm for convex hulls," *ACM Trans. on Mathematical Software*, vol. 22, no. 4, pp. 469–483, 1996.

[26] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka, "RWC music database: Popular, classical, and jazz music databases," in *Proc. ISMIR*, 2002, pp. 287–288.