INITIALIZATION-ROBUST MULTIPITCH ESTIMATION BASED ON LATENT HARMONIC ALLOCATION USING OVERTONE CORPUS

Daichi Sakaue, Katsutoshi Itoyama, Tetsuya Ogata, Hiroshi G. Okuno

Graduate School of Informatics, Kyoto University, Japan

ABSTRACT

We present a new method for modeling the overtone structures of musical instruments that uses an overtone corpus generated using a MIDI synthesizer. Since multipitch estimation requires a joint estimation of F0's and their overtone structures, one of the most important problems is the overtone structure modeling. Latent harmonic allocation (LHA), a promising multipitch estimation method, is difficult to use for various applications because it requires appropriate prior distributions of the overtone structures, which cannot be determined from statistical evidence. Our method uses an overtone corpus to avoid the problem of setting prior distributions and instead restricts the lower and upper bounds of each overtone weight. The bounds are determined from reference signals generated by a MIDI synthesizer. Experimental results demonstrated that the overtone structures were stably and accurately estimated for a wide variety of initial settings.

Index Terms— Multipitch estimation, harmonic clustering, overtone estimation, musical instrument sounds

1. INTRODUCTION

Multipitch estimation [1, 2, 3, 4] is an important research area because it has a wide range of applications, such as blind source separation [5], musical signal manipulation [6, 7], musical instrument identification [8], and musical chord recognition [9], which require musical scores as prior knowledge. Bayesian probabilistic models [1, 2, 3] are particularly valuable because they can model musical features, such as musical instrument, chord and onset as probabilistic latent variables, and thus enable joint estimation of the source model and these features [10, 11]. Latent harmonic allocation (LHA) [3] is a promising Bayesian method based on variational Bayes, which can estimate the posterior probabilities of each latent variable.

Currently, the estimation of LHA often fails because the model contains a lot of inappropriate optima, and thus sensitive to the initializations (such as F0s and relative weights of each sound.) A common way to avoid such optima is to set precise prior distributions of the overtone structure, but this is not a universal solution because there has been no method to estimate them under statistical evidence. For this reason, applications of LHA are limited because advanced Bayesian models based on LHA inherit such local optima in addition to the strength of it.

To overcome the weakness, we have developed a new method for modeling the overtone structure. Relative weights of instrument sounds, which represent the relative amplitudes of harmonic components observed in wavelet spectrograms, can be represented as a point on a simplex. We assume the simplex can be divided into two regions, one corresponding to appropriate overtone structures and the other corresponding to inappropriate overtone structures. We approximate the former region as a convex hull, and formulate a method based on variational Bayes that forces every overtone structure in the proposed model to have overtone weights contained in the convex hull.

2. OVERTONE STRUCTURE MODELING

Here we define harmonic clustering as Bayesian models representing the wavelet spectrum of each instrument sound as a probabilistic density function of a finite or infinite mixture of Gaussians. PreFEst [1], harmonic temporal clustering (HTC) [2], and LHA are examples of harmonic clustering. Let x be the log-frequency, μ_k be the fundamental frequency of the k-th instrument sound, λ_k be the precision of the mixture components, M be the number of overtones considered in the model, and $\tau_k = [\tau_{k1}, \cdots, \tau_{kM}]$ be the relative weight of each overtone. The k-th instrument sound is thus represented as:

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

$$o_m = 1200 \log_2 m, \qquad (2)$$

where $\mathcal{N}$ denotes the normal distribution, and o_m denotes the relative position of the m-th overtone component on the log-frequency axis. The relationship between the log and linear frequency scales is defined as:

$$f_{\log} = 1200(\log_2 f - \log_2 440 + 4.75). \qquad (3)$$

Because each τ_{km} is nonnegative and their summation is unity, the weight vector, $[\tau_{k1}, \cdots, \tau_{kM}]$, can be represented as a point on an $(M-1)$-simplex. Furthermore, the d-th time frame spectrum of the observed spectrogram is represented as a linear combination of K instrument sounds with mixing coefficients $\pi_d = [\pi_{d1} \cdots \pi_{dK}]$:

$$p(x) = \sum_{k=1}^{K} \pi_{dk} p_k(x|\tau_k, \mu_k, \lambda_k). \qquad (4)$$

2.1. Prior-based Approach

The performance of harmonic clustering depends on the estimation accuracy of the overtone structures, in other words, how the model can avoid the wrong estimation of them. For example, if a 440-Hz sound has an overtone weight $\tau_k = [0, 0, 1]$, the sound should be heard as 1320 Hz. More generally, we have prior knowledge that some overtone structures are obviously inappropriate. This is shown in Fig. 1. The aim of the overtone structure modeling is to avoid such wrong estimations, and the prior knowledge has been represented as a prior distribution of the overtone weight in former methods [1, 2].

However, this approach has a weakness that we cannot estimate the appropriate priors automatically on the basis of statistical evidence. This is because such priors should reflect the distribution of the musical instruments used to play the selected piece, and the distribution of the notes played. These distributions vary significantly for each piece, thus the prior learning of the prior distribution is almost impossible. Although infinite LHA [3] is a hierarchical Bayesian model that estimates the priors using a statistical model, it is not based on statistical evidence of the overtone structure because it only estimates the priors that maximize the model likelihood. For

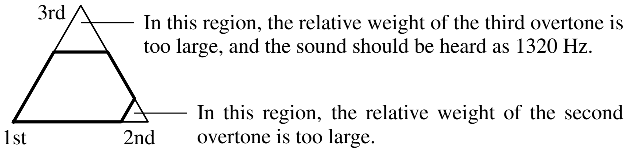
3rd
In this region, the relative weight of the third overtone is too large, and the sound should be heard as 1320 Hz.
In this region, the relative weight of the second overtone is too large.
1st
2nd

Fig. 1. Relative overtone weights of first three harmonic components of a 440-Hz sound, which can be represented as a point on a 2-simplex.

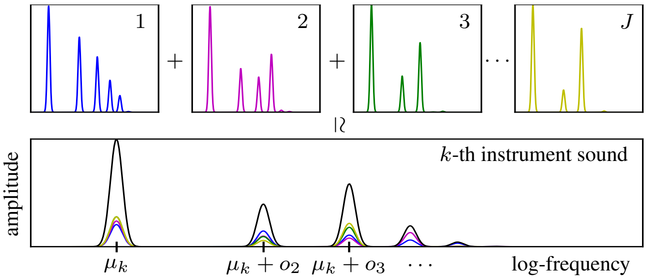

Fig. 2. Overtone corpus harmonic clustering. Each observed sound is approximated as nonnegative linear summation of J reference overtone structures.

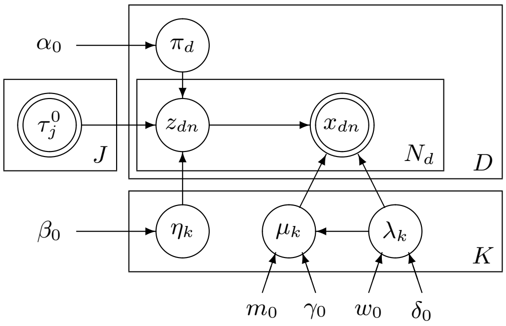

Fig. 3. Graphical model of proposed method. Single solid lines indicate latent variables, and double solid lines indicate observed variables.

these reasons, a de facto standard method to estimate the appropriate priors is not yet evident.

2.2. Overtone Corpus

In this paper, we introduce a new method to model the overtone structure, which restricts the relative overtone weights to existing in a convex hull. The vertices of the convex hull represent the overtone weights of reference signals. To do this, each instrument sound is represented as a nonnegative linear combination of J harmonic cluster templates. This is illustrated in Fig. 2. Let τ^0_{jm} be the weight of the m-th overtone component of the j-th template and $\eta_k = [\eta_{k1}, \cdots, \eta_{kJ}]$ be the mixing coefficients for each template. In the model, the k-th instrument sound is represented as:

$$p_k(x|\eta_k, \mu_k, \lambda_k) = \sum_{j=1}^{J} \eta_{kj} \sum_{m=1}^{M} \tau^0_{jm} \mathcal{N}(x|\mu_k + o_m, \lambda_k^{-1}), \quad (5)$$

where the τ^0_{jm} are precalculated using reference sounds and fixed in the estimation. The total weight of each overtone is represented as:

$$\tau_{km} = \sum_{j=1}^{J} \eta_{kj} \tau^0_{jm}. \quad (6)$$

With this model, it is easy to prove that there are upper and lower bounds on each overtone weight. Let $\tau^{(\min)}_m$ be the smallest m-th weight of the J templates and $\tau^{(\max)}_m$ be the largest one:

$$\tau^{(\min)}_m = \min_j \tau^0_{jm}, \qquad \tau^{(\max)}_m = \max_j \tau^0_{jm}. \quad (7)$$

As a result, τ_{km} is restricted to $\tau^{(\min)}_m \leq \tau_{km} \leq \tau^{(\max)}_m$ because mixing weights η_{kj} are nonnegative. More precisely, J weight vectors $[\tau^0_{j1}, \cdots, \tau^0_{jM}]$ consist a convex hull in the simplex, and τ_{km} is restricted to existing in it. This two-level construction is similar to latent variable decomposition [12].

Since the computational time of the model is proportional to the number of templates J, we introduce a method to reduce the number J efficiently. Because the interior points of the convex hull can be represented as a nonnegative linear combination of the vertices, we can remove such points from the model. Moreover, we introduce a selection criterion, which approximates the exact convex hull with a limited number of vertices. This is represented as:

$$\tilde{J} = \cup_m \left\{ \mathrm{argmin}_j \tau^0_{jm}, \mathrm{argmax}_j \tau^0_{jm} \right\}. \quad (8)$$

In addition to this, we add the second smallest and the second largest elements to the corpus to improve the estimation accuracy. By following this criterion, $\tilde{J}$ is smaller than $4M$ and is thus independent of the corpus size.

3. LATENT HARMONIC ALLOCATION COMBINED WITH OVERTONE CORPUS

Here we construct a Bayesian network of the proposed model on basis of the original LHA. Let D be the number of time frames and F be the number of frequency bins. In LHA, observed spectrogram X_{df} is interpreted as a histogram of a large number of independent particles. In other words, the particles in the d-th frame and the f-th frequency bin are observed X_{df} times. In the following description, $X = [X_1, \cdots, X_D]$ denotes the frequencies of the observed particles and $X_d = [x_{d1}, \cdots, x_{dN_d}]$ denotes the set of particles in the d-th frame, where N_d is the number of particles observed in the d-th frame. To perform variational Bayesian inference, we introduce a corresponding latent variable, Z, as the class allocation; z_{dn} indicates that observation x_{dn} is produced by the m-th overtone of the j-th template of the k-th instrument sound. α_0, β_0, γ_0, δ_0, m_0 and w_0 denote the hyperparameters of the model. The likelihoods and prior probabilities of the proposed model are stated as:

$$p(X|Z, \mu, \lambda) = \prod_{dnkjm} \mathcal{N}(x_{dn}|\mu_k + o_m, \lambda_k^{-1})^{z_{dnkjm}}, \quad (9)$$

$$p(Z|\pi, \eta) = \prod_{dnkjm} \left(\pi_{dk} \eta_{kj} \tau^0_{jm}\right)^{z_{dnkjm}}, \quad (10)$$

$$p(\pi) = \prod_{d=1}^{D} \mathrm{Dir}(\pi_d|\alpha_0) = \prod_{d=1}^{D} \prod_{k=1}^{K} \pi_{dk}^{\alpha^0_k - 1}, \quad (11)$$

$$p(\eta) = \prod_{k=1}^{K} \mathrm{Dir}(\eta_k|\beta_0) = \prod_{k=1}^{K} \prod_{j=1}^{J} \eta_{kj}^{\beta^0_j - 1}, \quad (12)$$

$$p(\mu, \lambda) = \prod_{k=1}^{K} \mathcal{N}(\mu_k|m_0, (\gamma_0 \lambda_k)^{-1}) \mathcal{W}(\lambda_k|w_0, \delta_0), \quad (13)$$

where Dir denotes the Dirichlet distribution and $\mathcal{W}$ denotes the Wishart distribution. A graphical model of the proposed method is shown in Fig. 3.

This model contains the latent variables Z, π, η, μ, and λ. The aim of Bayesian inference is to estimate the posterior distribution,

$p(Z, \pi, \eta, \mu, \lambda|X)$. In variational Bayesian inference, this is approximated as $p(Z, \pi, \eta, \mu, \lambda|X) \simeq q(Z)q(\pi, \eta, \mu, \lambda)$, and the factorized distributions are updated iteratively.

3.1. VB-E Step

In the VB-E step, the expected values ρ_{dnkjm} of class allocation z_{dnkjm} are calculated using the temporal estimation of π, η, μ, and λ:

$$\begin{aligned}\log q^*(Z) &= \mathbb{E}_{\pi,\eta,\mu,\lambda}\left[\log p(X, Z, \pi, \eta, \mu, \lambda)\right] + \text{const.} \\ &= \sum_{dnkjm} z_{dnkjm} \log \rho_{dnkjm} + \text{const.}, \end{aligned} \tag{14}$$

$$\begin{aligned}\log \tilde{\rho}_{dnkjm} &= \mathbb{E}\left[\log \pi_{dk}\right] + \mathbb{E}\left[\log \eta_{kj}\right] + \log \tau^0_{jm} \\ &+ \mathbb{E}\left[\log \mathcal{N}\left(x_{dn}|\mu_k + o_m, \lambda_k^{-1}\right)\right], \end{aligned} \tag{15}$$

$$\rho_{dnkjm} = \frac{\tilde{\rho}_{dnkjm}}{\sum_{kjm} \tilde{\rho}_{dnkjm}}. \tag{16}$$

3.2. VB-M Step

In the VB-M step, the sufficient statistics, N_k, N_{dk}, N_{kj}, and N_{fkm} are collected, and the variational posterior distribution of π, η, μ, and λ are calculated. Since all the prior distributions are conjugate, their variational posterior probability can be decomposed:

$$q(\pi, \eta, \mu, \lambda) = \prod_{d=1}^{D} q(\pi_d) \prod_{k=1}^{K} \left\{q(\eta_k) q(\mu_k, \lambda_k)\right\}, \tag{17}$$

$$q(\pi_d) = \mathrm{Dir}(\pi_d|\alpha_d), \qquad q(\eta_k) = \mathrm{Dir}(\eta_k|\beta_k), \tag{18}$$

$$q(\mu_k, \lambda_k) = \mathcal{N}(\mu_k|m_k, (\gamma_k \lambda_k)^{-1}) \mathcal{W}(\lambda_k|w_k, \delta_k). \tag{19}$$

Further, variational posterior hyperparameters $\alpha_{dk}, \beta_{kj}, \gamma_k, \delta_k, m_k$, and w_k can be calculated:

$$\log q^*(\pi, \eta, \mu, \lambda) = \mathbb{E}_Z\left[\log p(X, Z, \pi, \eta, \mu, \lambda)\right] + \text{const.}, \tag{20}$$

$$N_k = \sum_{dnjm} \rho_{dnkjm}, \qquad N_{dk} = \sum_{njm} \rho_{dnkjm}, \tag{21}$$

$$N_{kj} = \sum_{dnm} \rho_{dnkjm}, \qquad N_{fkm} = \sum_{dj} \sum_{x_{dn} = x_f} \rho_{dnkjm}, \tag{22}$$

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

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

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

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

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

4. EVALUATION

To evaluate the robustness of the proposed model, we conducted multipitch estimation experiments using 20 musical pieces and 3 initialization conditions.

4.1. Corpus Construction

We used a commercial musical instrument digital interface (MIDI) synthesizer (Roland SD-80) to produce reference signals and recorded the sounds of 70 General MIDI instruments (1 to 80, excluding 15, 17, 18, 19, 20, 32, 46, 48, 56, and 62.) The ones omitted have an illegal overtone structure. Instruments 81 to 128 are mostly artificial, so it is difficult to select ones with appropriate overtone structures. The sounds were recorded at 440 Hz for one second and transformed into wavelet spectrograms using Gabor wavelets. The spectrograms were integrated over time and each overtone frequency band, $f_m \le x_f < f_{m+1}$, to create the overtone weights:

$$f_m = \left(m - \frac{1}{2}\right) \times f_0, \quad \tau^0_{jm} \propto \sum_{d=1}^{D} \sum_{f_m^{(\log)} \le x_f < f_{m+1}^{(\log)}} Y_{df}^{(j)}, \tag{27}$$

where f_0 is the fundamental frequency, $f_m^{(\log)}$ is the corresponding log-frequency of f_m, and $Y_{df}^{(j)}$ is the wavelet spectrogram of the j-th instrument sound. The obtained parameters, τ^0_{jm}, were filtered using the criterion explained in 2.2. Finally, the size of the corpus, J, was set to 16.

4.2. Estimation Target

From the RWC Music Database [13], five piano solo pieces (RM-J001 to RM-J005), five guitar solo pieces (RM-J006 to RM-J010), and ten classical chamber pieces (RM-C012 to RM-C021) were selected and used to compare the performance of LHA with that of the proposed method. All the pieces were recorded from MIDI files using another MIDI synthesizer (Yamaha MOTIF-XS.) The recorded signals were truncated to the first 30 seconds and transformed into wavelet spectrograms using Gabor wavelets with a its time resolution of 16 [msec], frequency bins from 30 to 3000 [Hz], and frequency resolution of 12 [cents].

4.3. Experimental Settings

We used three types of initialization: random, linear, and exponential. The first starts the estimation from the VB-E step, and the other two start from the VB-M step. For the random initialization, we sampled the responsibilities ρ_{dnkm} or ρ_{dnkjm} from a uniform distribution. This is considerably the worst case and thus tests model stability for initialization. For the latter two, initial fundamental frequencies m_k were set from 33 Hz (C1) to 2093 Hz (C7), and the standard deviation $\sigma_k = (w_k \delta_k)^{-1/2}$, was set to 50 [cents]. The initial overtone weights were set to be uniform, or decaying exponentially. The relative weight of the source model in the d-th frame, π_{dk}, was set to the sum of the amplitudes of the nearest frequency bins of its overtones. Initialization of LHA for the exponential initialization was done using

$$\alpha_{dk} \propto \sum_{m=1}^{M} 2^{-m} X_{df_{km}}, \qquad \beta_{km} \propto 2^{-m}, \tag{28}$$

$$\gamma_k = \sum_{d=1}^{D} \alpha_{dk}, \quad \delta_k = \sum_{d=1}^{D} \alpha_{dk}, \quad w_k^{-1} = \delta_k \left(50\,[\text{cents}]\right)^2 \tag{29}$$

to imitate the update equations of the variational Bayes. For the proposed model, we cannot initialize overtone weights τ_{km} directly because they are represented as the summation of J harmonic templates. Instead, we optimized β_{kj} by using EUC-NMF [14] for 100 iterations, so that the total overtone weights, $\sum_j \beta_{kj} \tau_{jm}$, approximated the initial overtone weights.

During the overall estimation, all priors were set to be non-informative. That is, $\alpha_0, \beta_0, \delta_0$, and w_0 were set to unity, γ_0 was set to 10^{-3}, and m_0 was set to zero. Model orders K, J, and M were set to 73, 16, and 6, respectively, where the number of overtone M is equal to HTC [2]. For the random initialization, estimations were truncated at 1000 iterations and for the linear and exponential ones, they were truncated at 100 iterations. The numbers of iterations were 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 $N_d \pi_{dk}$ satisfies $N_d \pi_{dk} \ge r \max_{dk} N_d \pi_{dk}$,

Table 1. Calculated F-measures: *rand* stands for random initialization, *linear* stands for linear initialization, and *exp* stands for exponential initialization.

	LHA			OC-LHA (proposed)		
Music Type	rand	linear	exp	rand	linear	exp
Piano Solo	31.1	51.3	58.5	55.0	**62.3**	58.1
Guitar Solo	12.8	48.8	**76.7**	64.3	73.9	72.6
Chamber	23.8	36.1	49.1	46.8	**53.6**	50.9

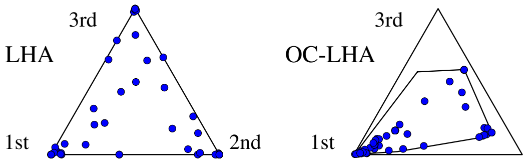

Fig. 4. Estimated relative weights of first three harmonic components of 73 sound source models obtained for random initialization with musical piece *RM-C012*. Convex hull projected to 2-simplex is displayed with solid lines. Corresponding F-measures are 19.4 for LHA and 63.4 for OC-LHA.

is considered to be sounding. The threshold was optimized experimentally for each piece, initialization, and method 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 $D \times 128$ binary maps, and the F-measure was 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 $F = 200R/(N + C)$. The larger the value of the F-measure, the better the performance.

4.4. Results

The results are summarized in Table 1. LHA with exponential initialization and the proposed model with linear initialization had similar performance, which means LHA and the proposed model work properly once the appropriate initial parameters are given. Comparing the linear and exponential columns, we see that LHA was more sensitive to the initialization of overtone weights. The estimation accuracy of LHA was substantially worse with random initialization because LHA did not handle the overtone weights appropriately. In contrast, the accuracy of the proposed model was consistent among the three initializations.

The estimated overtone weights are plotted in Fig. 4. LHA favored source models with only a second overtone or with only a third overtone while the proposed model did not estimate such wrong structures.

5. DISCUSSIONS

5.1. Properties of the Overtone Corpus

Here we explain the validity of the assumption that the region of appropriate overtone weights is a convex hull. Let us have J template sounds of the same pitch, and plot their relative overtone weights in a simplex. The plotted points make a convex hull. If we pick up any point inside it, the corresponding sound can be obtained as a linear combination of the template sounds which share the same phase. In this case, the pitch of the sound is equal to that of the template sounds. Therefore, our assumption can be justified whenever the perceptual pitch and the fundamental frequency are equal.

Another important property we should discuss is, the assumption that the simplex can be divided into two regions, in one region the corresponding overtone structures are appropriate and in the other the corresponding overtone structures are inappropriate. For the perceived pitch depends on the listener's perception, a more better model can be obtained if we take into account another region where the corresponding pitch is unknown. This should be considered in the future work.

5.2. Relationship with the Conventional Method

Our model includes the original LHA as a special case whenever the harmonic templates is set to $\tau_{jm} = \omega_{jm}$, where ω_{jm} is the Dirac delta function. The proof is omitted for reasons of space.

6. CONCLUSIONS

Our proposed method for overtone structure modeling represents the overtone weights as a nonnegative linear combination of harmonic templates. The templates are collected from audio signals generated using a MIDI synthesizer to avoid overtone structures that are not observed in audio signals. We assumed that the appropriate overtone region is a convex hull and introduced an initialization-robust multipitch estimation. Experimental results demonstrated that the overtone structures were stably and accurately estimated for a wide variety of initial settings. We will extend this model and construct a joint estimation framework of multipitch estimation and instrument identification in the future work. This research was partially supported by Grant-in-Aid for Scientific Research (S) and the Global COE Program.

7. REFERENCES

[1] M. Goto, "PreFEst: A predominant-F0 estimation method for polyphonic musical audio signals," in *MIREX*, 2005.

[2] H. Kameoka et al., "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, "Infinite latent harmonic allocation: A nonparametric Bayesian approach to multipitch analysis," in *Proc. ISMIR*, 2010, pp. 309–314.

[4] 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.

[5] K. Itoyama et al., "Simultaneous processing of sound source separation and musical instrument identification using Bayesian spectral modeling," in *Proc. ICASSP*, 2011, pp. 3816–3819.

[6] N. Yasuraoka et al., "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.

[7] N. Yasuraoka et al., "I-Divergence-based dereverberation method with auxiliary function approach," in *Proc. ICASSP*, 2011, pp. 369–372.

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

[9] Y. Ueda et al., "HMM-based approach for automatic chord detection using refined acoustic features," in *Proc. ICASSP*, 2010, pp. 5518–5521.

[10] A. Maezawa et al., "Polyphonic audio-to-score alignment based on Bayesian latent harmonic allocation hidden Markov model," in *Proc. ICASSP*, 2011, pp. 185–188.

[11] K. Miyamoto et al., "Harmonic-temporal-timbral clustering (HTTC) for the analysis of multi-instrument polyphonic music signals," in *Proc. ICASSP*, 2008, pp. 113–116.

[12] B. Raj et al., "Latent variable decomposition of spectrograms for single channel speaker separation," in *Proc. WASPAA*, 2005, pp. 17–20.

[13] M. Goto et al., "RWC music database: Popular, classical, and jazz music databases," in *Proc. ISMIR*, 2002, pp. 287–288.

[14] M. Nakano et al., "Convergence-guaranteed multiplicative algorithms for nonnegative matrix factorization with beta-divergence," in *Proc. MLSP*, 2010, pp. 283–288.