INTRA-NOTE SEGMENTATION VIA STICKY HMM WITH DP EMISSION

Yuma Koizumi, Katunobu Itou

Graduate School of Computer and Information Sciences, Hosei University
3-7-2 Kajinocho, Koganei, Tokyo, 184-8584, Japan
12t0005@cis.k.hosei.ac.jp

ABSTRACT

This paper presents an intra-note segmentation method for monophonic recordings based on acoustic feature variation; each musical note is separated into onset, steady and offset states. The task of intra-note segmentation from audio signals is detecting change points of acoustic feature. In proposed method, the Markov process is assumed on state transition, and time-varying acoustic feature is represented by three Dirichlet processes (DP) that are emitted by the each state. In order to express the generative process, the sticky hidden Markov model (HMM) with DP emission is employed. This modeling allows us to automatically estimate the state transition while avoiding the model selection problem by assuming countably infinite of possible acoustic feature in musical notes. Experimental result shows that the detection accuracy of onset–to–steady and steady–to–offset were improved 2.3 points and 20.7 points from previous method, respectively.

Index Terms— intra-note segmentation, music information retrieval, hidden Markov model, Dirichlet process

1. INTRODUCTION

Musicians do not play exactly what is written in the score because they interpret the music in their own way. Deviances, such as vibrato or rubato, are included in the tempo, amplitude, timbre and pitch in their performance. These deviances are among the factors that make a listener judge a performance as "expressive" and/or "individual". Hence, more or less musical applications [1, 2, 3] require understanding players' performance expression/intention. For this reason, many attempts of analyzing and modeling its have been made up to this day [4, 5, 6, 7, 8, 9, 10, 11, 12].

A musical tone generally has three possible states: *onset*, *steady*, and *offset*. In particular, "*local deviances*" in each note (e.g. vibrato or articulation) have different performance effects depending on the states. For example, if a player uses fast vibrato at around onset timing (i.e. onset state), the vibrato effects "*accent*" called as "vibrato accent". Thus, as a pre-processing for musical performance analysis, we need to deal with intra-note segmentation; a musical tone is needed to be separated into the three states. As an application example of intra-note segmentation, a timbre model of musical instruments is proposed [13].

As literatures on this subject, methods based on collinear approximation of amplitude variation with the decided number of straight lines were proposed [14, 15]. In these methods, the states are estimated via gradient of these straight lines. However, because observed amplitudes have various shapes depending on musical expression, complex amplitude variations of excitation-continuous musical instruments (e.g. wind instruments or bowed strings) with including vibrato or tremolo could not be approximated by the decided number of straight lines.

In this paper, we propose more flexible intra-note segmentation for excitation-continuous musical instruments based on sticky *hidden Markov model* (sticky HMM) with Dirichlet process (DP) emission [16]. In the proposed method, countable infinite acoustic variations are considered in the three states. The state transition is detected by clustering of observed acoustic features.

In section 2 we begin by describing acoustic characteristics of each state. In section 3, generative model of intra-note segments is described and section 4 describes inference of these states. Finally, the experimental result is presented in section 5.

2. ACOUSTIC CHARACTERISTICS OF EACH STATE

Onset, steady and offset states are sectionalized depending on difference in vibration of excitation source. The task of intra-note segmentation from audio signals is detecting change points of acoustic feature due to the excitation differences. Figure 1 shows an example of the differences in acoustic features on violin recordings.

The onset state is the interval between onset timing[1] and stabilizing timing of excitation source. As acoustic features, the amplitude is increased on almost instruments and playing styles [17, 18]. Moreover, in a part of playing style, the timbre becomes like a "noise" due to instable vibration of excitation source (Fig. 1 (b)).

Unified definition of steady state is quite difficult because not all instruments contain the same temporal events. In this paper, the steady state is defined as almost constant interval of acoustic features. When the note is played with vibrato or some playing style, acoustic feature changes at around the constant value.

The offset state is the interval between exit timing of the excitation control and offset timing[2]. As acoustic features, the amplitude decrease rapidly and high-level harmonics decrease gently (Fig. 1 (a)).

3. GENERATIVE MODEL OF INTRA-NOTE SEGMENTS

In this section, we introduce a generative model of intra-note states and acoustic features. In the following, t is index of time frame and x_t is amplitude at time t dealt in log-domain (dB). Further, $\mathcal{N}, \mathcal{W}, \mathcal{M}, \mathcal{D}$, Ber and Bin denote Gaussian, Wishart, Multinomial, Dirichlet, Bernoulli and Binomial distribution, respectively.

[1] Start timing of the note.

[2] Timing that the note becomes imperceptible.

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

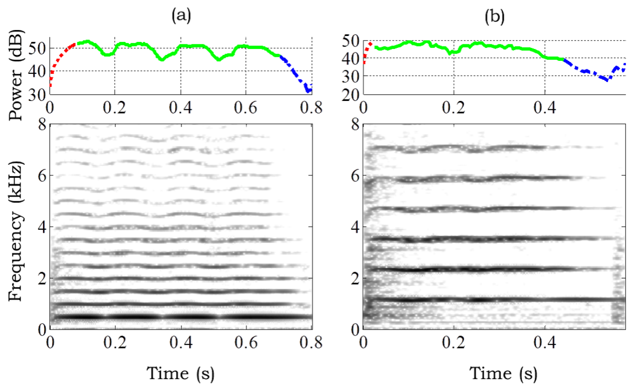

Fig. 1. Examples of acoustic characteristics of each state (top: amplitude, bottom: spectrogram). A normal tone (a) and a strong tone (b) played with the violin. In each top figure, dotted line denotes the onset state, solid line denotes the steady state and dashed line denotes the release state.

3.1. Acoustic feature for intra-note segmentation

As above mentioned, acoustic features that varies depending on the states are mainly amplitude and timbre. Hence, in this study, acoustic feature related with amplitude and timbre are used for the segmentation.

The amplitude is characterized with time-variation such as increase or decrease. Therefore, we use first-order differentiation of the amplitude, $\Delta x_t = (x_t - x_{t-1})/\Delta t$, as amplitude characteristics. This is consistent intuitively to ADSR (Attack Decay Steady/Sustain Release) that is a generative model of amplitude which expressed explicitly intra-note segment. ADSR expresses amplitude modulation with some (decided number of) straight lines or curves.

The timbre is characterized with aperiodicity and harmonic ratio. Hence, the spectral entropy [19] and low-dimensional features of spectrum envelope are used for aperiodicity and harmonic ratio, respectively. In order to express spectral envelope in low-dimensional, spectral envelope is deemed as probability density function, and 1$^{\text{st}}$ to 4$^{\text{th}}$ order moments are calculated [20]. Then, the principal component analysis (PCA) is executed to calculated spectral entropy and the moments, and then the top 3-dimensions (c^1, c^2, c^3) are selected due to contribution ratio.

From the above, $\boldsymbol{y}_t = (\Delta x_t, c_t^1, c_t^2, c_t^3)^\dagger$ is employed as acoustic feature. Here, † denotes transpose of vector or matrix.

3.2. Generative model of acoustic features

The actual amplitude and timbre in a musical note are time-varying due to various factors such as vibrato or playing style. The time-varying is closely related to performance expression. Therefore, in order to express the every variation of performance, it is not validity to fix the complexity of the model, like the ADSR and previous methods [14]. The complexity of the model should be decided according to the complexity of the observed acoustic feature.

Meanwhile, the number of intra-note state is generally three. In some playing style such as *legato*, there are cases that some state vanish. Though, in any playing style, there are no cases that the number of state is increase from three. Therefore, the complexity of the observed acoustic feature should be considered under the state transition.

For these reasons, we employ hierarchical generative process of the states and acoustic features. Namely, first, players generate transition of the $K = 3$ states. Next, the players select acoustic feature patterns from infinite number of own acoustic candidate ($J_k \to \infty$) on each state, and then the player generate a musical tone by combining the selected acoustic feature patterns.

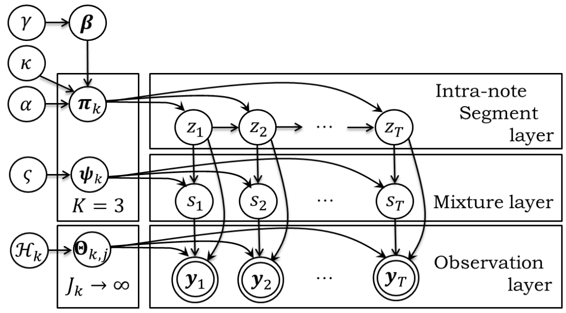

Fig. 2. A graphical representation of a sticky HDP-HMM with nested DP emission for intra-note segmentation.

In order to represent the process as statistical model, we employ sticky HMM with nested DP emission [16] (Fig. 2). In this model, the acoustic feature at t, $\boldsymbol{y}_t$, is generated by infinite Gaussian Mixture Model (infinite-GMM) $\sum_{j=1}^{J_{z_t}} \psi_{z_t,j} \mathcal{N}(\boldsymbol{\mu}_{z_t,j}, \boldsymbol{\Lambda}_{z_t,j}^{-1})$ corresponding to the state z_t. This model is similar to the Infinite-State Spectrum Model presented by Nakano et al. [11] in terms of attempting to express time-varying acoustic feature by infinite number of patterns. Whereas Nakano expressed the time-varying by HMM directly, we attempt to model the state explicitly and express transition of infinite mixture distribution.

Here, we describe generative process of acoustic features $\boldsymbol{y}_{1,\ldots,T}$. First, intra-note state at t, z_t, is generated by Multinomial $\mathcal{M}(\boldsymbol{\pi}_{z_{t-1}})$. The parameter of the Multinomial $\boldsymbol{\pi}_k$ denotes state transition probability of state k to the next state. And its prior distribution is its conjugate distribution, Dirichlet distribution, as follow:

$$\boldsymbol{\pi}_k \sim \mathcal{D}\left(\alpha\beta(Z_1), \ldots, \alpha\beta(k)+\kappa, \ldots, \alpha\beta(Z_K)\right), \tag{1}$$

$$\boldsymbol{\beta} \sim \mathcal{D}\left(\gamma/K, \ldots, \gamma/K\right). \tag{2}$$

Here, $\kappa > 0$ is a parameter for self–transition bias, and $\alpha, \gamma > 0$ are hyper parameters.

Next, indicator of Gaussian at t, s_t, is determined by Multinomial $\mathcal{M}(\boldsymbol{\psi}_{z_t})$. The parameter of the Multinomial $\boldsymbol{\psi}_k$ is mixture weight of k^{th} states' infinite-GMM, and the weight is generated by Stick-breaking process [21] with a parameter $\varsigma > 0$.

Finally, acoustic feature at t, $\boldsymbol{y}_t$, is generated by s_t^{th} Gaussian on state z_t, $\mathcal{N}(\boldsymbol{\mu}_{z_t,s_t}, \boldsymbol{\Lambda}_{z_t,s_t}^{-1})$ with parameters $\Theta_{k,j} = \{\boldsymbol{\mu}_{k,j}, \boldsymbol{\Lambda}_{k,j}\}$. In this study, we employ nested DP [22], and prior distribution of each Gaussians' parameters is Gaussian-Wishart distribution with parameters $\mathcal{H}_k = \{\boldsymbol{\lambda}_k, R_k, \boldsymbol{W}_k, \nu_k\}$.

4. STATES INFERENCE

In this section, we describe the inference for intra-note states $z_{1,\ldots,T}$. Latent variables of DP can be inferred by Variational Bayesian methods (VB) or Markov chain Monte Carlo methods (MCMC). Because the proposed model is quite complex, it is difficult to use deterministic procedures such as VB. Instead, we use Gibbs sampler to update latent variables. The basic algorithm is same as the literature [16], thus we abbreviate its derivation and describe its algorithm and update formulas.

4.1. Parameters inference with Gibbs Sampling

The latent variables are iteratively drawn from their conditional posterior distributions. The sampling order is $z_t, s_t, \boldsymbol{\beta}, \alpha, \kappa, \varsigma$ and $\mathcal{H}_k$.

Step 1: z_t and s_t are drawn from following conditional posterior:

$$z_t \sim \sum_{k=1}^{K} f_k(y_t)\delta(z_t, k), \tag{3}$$

$$s_t \sim \sum_{j=1}^{J} f'_{z_t,j}(y_t)\delta(s_t, j) + f'_{z_t,J_{z_t}+1}(y_t)\delta(s_t, J_{z_t}+1), \tag{4}$$

where

$$f_k(\boldsymbol{y}_t) = \left(\alpha\beta_k + n^-_{z_{t-1},k}\right) \times \left(\frac{\alpha\beta_{z+1} + n^-_{k,z_{t+1}} + \kappa\delta(k, z_{t+1})}{\alpha + n^-_{k,\cdot} + \kappa}\right) \sum_{j=1}^{J_k} \mathcal{N}(\boldsymbol{y}_t|\hat{\boldsymbol{\mu}}_{k,j}, \hat{\boldsymbol{\Lambda}}^{-1}_{k,j}), \tag{5}$$

$$f'_{z_t,j}(\boldsymbol{y}_t) = \left(\frac{m^-_{z_t,j}}{\varsigma + m^-_{z_t,\cdot}} \mathcal{N}(\boldsymbol{y}_t|\hat{\boldsymbol{\mu}}_{z_t,j}, \hat{\boldsymbol{\Lambda}}^{-1}_{z_t,j})\right), \tag{6}$$

$$f'_{z_t,J_{z_t}+1}(\boldsymbol{y}_t) = \left(\frac{\varsigma}{\varsigma + m^-_{z_t,\cdot}} \mathcal{N}(\boldsymbol{y}_t|\hat{\boldsymbol{\mu}}_{z_t,J_{z_t}+1}, \hat{\boldsymbol{\Lambda}}^{-1}_{z_t,J_{z_t}+1})\right). \tag{7}$$

Here, $n_{k,k'}$ represents the number of Markov chain transition from state k to k', $m_{k,j}$ represents the number of active count of j^{th} Gaussian on state k, superscript "$-$" denotes removing information of $\boldsymbol{y}_t$, "$\cdot$" denotes summation of its variable and $\delta(i, j)$ is Kronecker delta. Here, $\hat{\boldsymbol{\mu}}_{z_t,j}$ and $\hat{\boldsymbol{\Lambda}}_{z_t,j}$ are drawn from following equations:

$$\hat{\boldsymbol{\mu}}_{z_t,j} \sim \mathcal{N}\left(\frac{\bar{\boldsymbol{y}}^-_{z_t,j}\hat{\boldsymbol{\Lambda}}_{z_t,j} + \boldsymbol{\lambda}_{z_t}R_{z_t}}{m^-_{z_t,j}\hat{\boldsymbol{\Lambda}}_{z_t,j} + R_{z_t}}, \left(m^-_{z_t,j}\hat{\boldsymbol{\Lambda}}_{z_t,j} + R_{z_t}\right)^{-1}\right), \tag{8}$$

$$\hat{\boldsymbol{\Lambda}}_{z_t,j} \sim \mathcal{W}\left(\left(\nu_{z_t}\boldsymbol{W}_{z_t} + \Phi^-_{z_t,j}\right)^{-1}, \nu_{z_t} + m^-_{z_t,j}\right), \tag{9}$$

$$\bar{\boldsymbol{y}}_{k,j} = \sum_{t' \in (z_t = k, s_t = j)} \boldsymbol{y}_{t'}, \tag{10}$$

$$\Phi_{k,j} = \sum_{t' \in (z_t = k, s_t = j)} (\boldsymbol{y}_{t'} - \hat{\boldsymbol{\mu}}_{z_t,j})(\boldsymbol{y}_{t'} - \hat{\boldsymbol{\mu}}_{z_t,j})^{\dagger}. \tag{11}$$

Please note that in mean variable of Gaussian distribution of equation (8), inverse matrix is written by division due to limitations of space. After sampling for all $t \in 1, ..., T$, if there exist a j such that $m_{z_t,j} = 0$, remove j and decrease J_{z_t}.

Step 2: Sampling $\boldsymbol{\beta}$. State transition inference of a sticky HMM is not Chinese Restaurant Franchise (CRF), but that is CRF with Loyal Customers [23]. Thus, $\boldsymbol{\beta}$ is drawn by using auxiliary random variables $\boldsymbol{q}, \boldsymbol{r}, \bar{\boldsymbol{q}}$ as following:

$$q_{k,k'} = \sum_{i=1}^{n_{k,k'}} u_i, \quad u_i \sim \text{Ber}\left(\frac{\alpha\beta_{k'} + \kappa\delta(k, k')}{i + \alpha\beta_{k'} + \kappa\delta(k, k')}\right) \tag{12}$$

$$r_k \sim \text{Bin}\left(q_{k,k}, \frac{\rho}{\rho + \beta_k(1-\rho)}\right), \tag{13}$$

$$\bar{q}_{k,k'} = \begin{cases} q_{k,k'} & (k \neq k'), \\ q_{k,k'} - r_k & (k = k'), \end{cases} \tag{14}$$

$$\boldsymbol{\beta} \sim \mathcal{D}\left(\bar{q}_{\cdot,1}, \bar{q}_{\cdot,1}, ..., \bar{q}_{\cdot,K}\right), \tag{15}$$

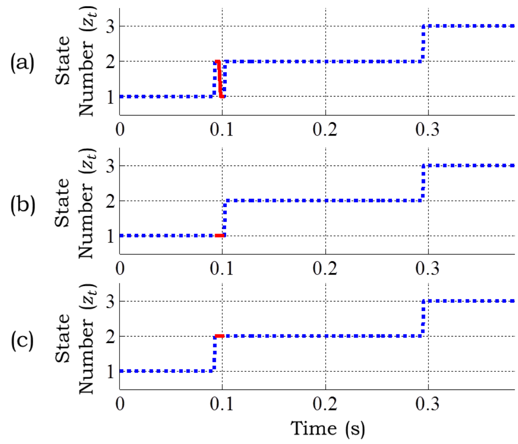

Fig. 3. An example of state adjustment. Estimated state z_t (a), adjustment pattern 1 (b) and adjustment pattern 2 (c).

where $\rho = \kappa/(\alpha + \kappa)$.

Step 3: Smpling hyper-parameters $\alpha, \kappa, \varsigma$ and $\mathcal{H}_k$. Sampling equations of α, κ and ς are omitted since become redundant, but the algorithm is same as the [16]. $\mathcal{H}_k$ is sampled via infinite-GMMs' method [24] by using $\boldsymbol{y}_t \in z_t = k$.

If the iteration count reaches the appointed number, the iteration is exited. Otherwise, the algorithm returns to step 1.

4.2. Post-processing for z_t

State transition in a musical note is a Left-to-Right automaton including state skips. However, the sticky HMM is ergodic HMM, thus there are some cases of state "backset", such as "onset $\rightarrow$ steady $\rightarrow$ onset" (Fig. 3 (a)). In these cases, z_t is adjusted by post-processing.

Let us consider P patterns of adjustable state transition $\hat{z}^p_\tau$ in time interval $\tau \in \{t_1, ..., t_2\}$ (e.g. In Fig. 3, $\hat{z}^1_\tau$ =(b), $\hat{z}^2_\tau$ =(c) and $P = 2$). When HMM parameters $\Upsilon = \{\boldsymbol{\pi}_k, \boldsymbol{\phi}_k, \Theta_k\}$ are given, the likelihood of each pattern can be written as follow:

$$p(\hat{z}^p_\tau, \boldsymbol{y}_\tau|\Upsilon) = \prod_{\tau=t_1}^{t_2+1} \pi_{z^p_{\tau-1}, z^p_\tau} \sum_{j=1}^{J_{z^p_\tau}} \psi_{z^p_\tau,j}\mathcal{N}(\boldsymbol{y}_\tau|\boldsymbol{\mu}_{z^p_\tau,j}, \boldsymbol{\Lambda}^{-1}_{z^p_\tau,j}). \tag{16}$$

In this paper, z_t is adjusted via $\hat{z}^p_\tau$ that maximize equation (16).

Figure 4 shows an result example of intra-note segmentation whose musical note is played by the violin (468Hz). Although estimated state transition has a little difference with the true state transition, the estimated error is less than about 20 ms. Moreover, time-variation of acoustic feature is represented by ($\sum_K J_k =$)11 Gaussians.

5. EXPERIMENT

This section presents the experimental result of proposed method on actual musical recordings. For the experiment data, three phrases (saxophone, clarinet and trumpet) from Music Information Retrieval Evaluation eXchange (MIREX) onset detection dataset [25], two phrases (flute and trumpet) from RWC Music Database (jazz music) [26] and five phrases from a database of solo violin recordings [12] were used. The reason of this selection is these data includes a variety of playing style on classic and jazz. All musical note were

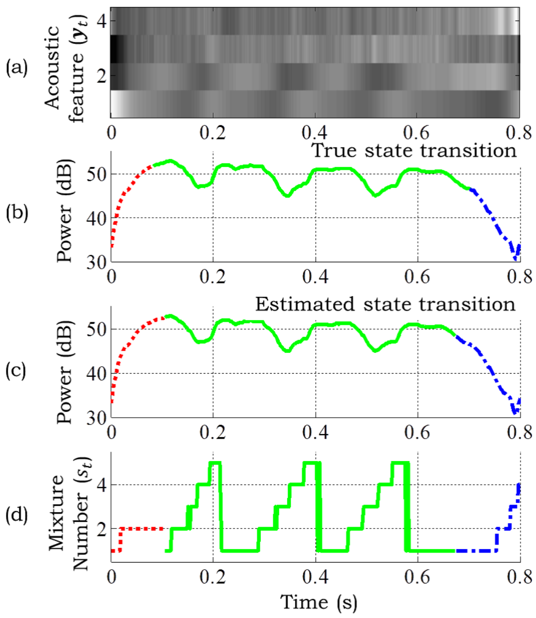

Fig. 4. An result of intra-note segmentation. Acoustic feature (a), true state transition (b), estimated state transition (c) and mixture number (d). In (b), (c) and (d), dotted line denotes the onset state, solid line denotes the steady state and dashed line denotes the release state.

separated into each note by hand-labeling of onset timings and offset timings. There were 349 musical notes in total. All signals were processed as monaural signals sampled at 48 kHz and 24 bit. The correct labels are generated by mean of three musicians' hand labeling result that are based on audio signal, fundamental frequency, spectrogram and amplitude.

For acoustic feature calculation, temporal shift and window length of Short-Time Fourier Transform (STFT) are 1 ms and 20 ms, respectively. The hyperparameters of α, κ and ς in [16] were set to $a, b, c, d = 1$. Gaussian indicator s_t was initialized by random value with $J_k = 30$. To ensure the numerical stability of the algorithm, we placed the initial value of z_t as $z_{1,\dots,T/4} = 1$, $z_{T/4+1,\dots,3T/4} = 2$ and $z_{3T/4+1,\dots,T} = 3$. Appointed number of max iteration was 1000.

5.1. Experiment for intra-note segmentation

The accuracy of proposed intra-note segmentation was compared with a previous method [14] via precision. In intra-note segmentation, a musical note is separated into three states, thus the accuracy was evaluated on detected state transition time of onset–to–steady (A–to–S) and steady–to–offset (S–to–R). In S-to-R, there was significant difference by the 2-sample test for equality of proportions (significance levels were 1 %). Correct matches imply that the target and detected onsets were within a 50-ms window [17]. This window is to allow for the inaccuracy of the hand labeling process.

Figure 5 shows the result of segmentation accuracy. The accuracy of proposed method was 2.3 point and 20.7 point higher than the previous method, A–to–S and S–to–R respectively. The previous method is employed at performer identification [1, 2] and timbre modeling [27], and proposed method can segmentation sophisticatedly than the previous one on excitation-continous musical instruments note. Thus, it can be concluded that the proposed method is efficient for pre-processing of musical performance analysis.

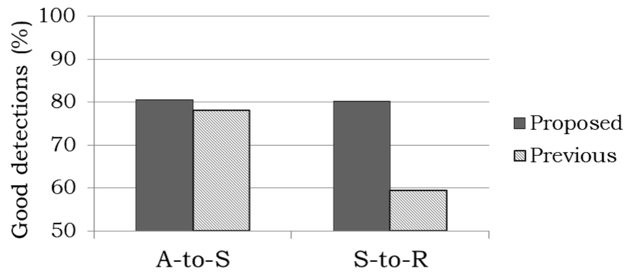

Fig. 5. Evaluation result. "A–to–S" and "S–to–R" denote change point of "onset state to steady state" and "steady state to offset state", respectively.

6. CONCLUSIONS

In this paper, we proposed a flexible intra-note segmentation for excitation-continuous musical instruments based on sticky HMM with DP emission. In the method, we assumed that players perform a musical note by selecting and combining acoustic feature from countable infinite variations in each state. The state transition was detected by clustering of observed acoustic features. Experimental result shows that the detection accuracy of onset–to–steady and steady–to–offset were improved 2.3 point and 20.7 point from previous method, respectively. The previous method is employed at performer identification and timbre modeling, and proposed method can segmentation sophisticatedly than the previous one on excitation-continous musical instruments note. Thus, it can be concluded that the proposed method is efficient for pre-processing of musical performance analysis.

In this study, the issue of state "backset" due to ergodic property of HMM was resolved by post-processing. Meanwhile, by constricting the transition probability matrix $\boldsymbol{\pi}_k$ as upper triangular matrix, the post-processing can be omitted. Moreover, it can be consider that, this constraint can improve inference accuracy of emission distribution of acoustic feature on each state. In the future, we are going to derive the constraint version of update equations.

In fact, there are two causes of amplitude time-varying: articulation and dynamics (e.g. *crecendo*). This study only considered the cause of articulation. Thus, in future, we need to consider preliminarily removing or statistical modeling of effect of the dynamics

As future prospects, it can be considered that the inferred HMM parameters $\Upsilon = \{\boldsymbol{\pi}_k, \boldsymbol{\phi}_k, \Theta_k\}$ and indicator $s_{1,\dots,T}$ can be regarded as analyzing result of performance style characteristics. Thus, we will attempt to apply it for performance modeling or performer identification.

7. REFERENCES

[1] R. Ramirez, E. Maestre, and X. Serra, "Automatic performer identification in commercial monophonic jazz performances," *Pattern Recognition of Non-Speech Audio*, vol. 31, no. 12, pp. 1514–1523, 2010.

[2] R. Ramirez, E. Maestre, A. Perez, and X. Serra, "Automatic performer identification in celtic violin audio recordings," *Music and Machine Learning*, vol. 40, no. 2, pp. 165–174, 2011.

[3] T. Nakano, M. Goto, and Y. Hiraga, "An automatic singing skill evaluation method for unknown melodies using pitch interval

accuracy and vibrato features," in *the International Conference on Spoken Language Processing (INTERSPEECH)*, 2006, pp. 1706–1709.

[4] H. Kawahara and M. Morise, "Analysis and synthesis of strong vocal expressions: Extension and application of audio texture features to singing voice," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2012.

[5] L. Regnier and G. Peeters, "Singer verification: Singer model .vs. song model," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2012.

[6] E. Maestre, M. Blaauw, J. B., E. Guaus, and A. Perez, "Statistical modeling of bowing control applied to violin sound synthesis," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 4, pp. 855–871, 2010.

[7] P. Papiotis, M. Marchini, and E. Maestre, "Computational analysis of solo versus ensemble performance in string quartets: Dynamics and intonation," in *Proceedings of 12th International Conference on Music Perception and Cognition (ICMPC)*, 2012.

[8] M. Marchini, P. Papiotis, and E. Maestre, "Timing synchronization in string quartet performance: a preliminary study," in *Proceedings of International Workshop on Computer Music Modeling and Retrieval (CMMR12)*, 2012, pp. 117–185.

[9] M. Marchini, R. Ramirez, and E. Maestre P. Papiotis, "Inducing rules of ensemble music performance : A machine learning approach," in *the 3rd International Conference on Music & Emotion (ICME3)*, 2013.

[10] Y. Ohishi, H. Kameoka, D. Mochihashi, H. Nagano, and K. Kashino, "Statistical modeling of f0 dynamics in singing voices based on gaussian processes with multiple oscillation bases," in *Proceedings of International Conference on Spoken Language Processing (INTERSPEECH)*, 2012.

[11] M. Nakano, J.L. Roux, H. Kameoka, T. Nakamura, N. Ono, and S. Sagayama, "Bayesian nonparametric spectrogram modeling based on infinite factorial infinite hidden markov model," in *Proceedings of Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2011, pp. 325–328.

[12] Y. Koizumi and K. Itou, "Expressive oriented time-scale adjustment for mis-played musical signals based on tempo curve estimation," in *Proceedings of International Conference on Digital Audio Effects (DAFx)*, 2013.

[13] J.J. Burred and A. Robel, "A segmental spectro-temporal model of musical timbre," in *Proceedings of International Conference on Digital Audio Effects (DAFx)*, 2010.

[14] E. Maestre and E. Gomez, "Automatic characterization of dynamics and articulation of expressive monophonic recordings," in *Proceedings of the 118th Audio Eng. Society Convention*, 2005.

[15] K. Jensen, "Envelope model of isolated musical sounds," in *Proceedings of International Conference on Digital Audio Effects (DAFx)*, 1999.

[16] E. B. Fox, E. B. Sudderth, M. I. Jordan, and A. S. Willsky, "The sticky hdp-hmm: Bayesian nonparametric hidden markov models with persistent states," Tech. Rep., MIT Laboratory for Information and Decision Systems, 2007.

[17] J.P. Bello, L. Daudet, S. Abdallah, C. Duxbury, M. Davies, and M. Sandler, "A tutorial on onset detection in music signals," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 5, pp. 1035–1047, 2005.

[18] M. Caetano, J.J. Burred, and X. Rodet, "Automatic segmentation of the temporal evolution of isolated acoustic musical instrument sounds using spectro-temporal cues," in *Proceedings of International Conference on Digital Audio Effects (DAFx)*, 2010.

[19] P. Renevey and A. Drygajlo, "Entropy based voice activity detection in very noisy conditions," in *Proceedings of 7th European Conference on Speech Communication and Technology (EUROSPEECH)*, 2001.

[20] G. Peeters, "A large set of audio features for sound description (similarity and classification) in the cuidado project," Tech. Rep., `http://www.ircam.fr/anasyn/peeters/ARTICLES/Peeters_2003_cuidadoaudiofeatures.pdf.`, 2004.

[21] David M. Blei and Michael I. Jordan, "Variational inference for dirichlet process mixtures," *Bayesian Analysis*, vol. 1, pp. 121–144, 2005.

[22] A. Rodrigueza, D. B. Dunsona, and A. E. Gelfand, "The nested dirichlet process," *the American Statistical Association*, pp. 1131–1154, 2008.

[23] E. B. Fox, E. B. Sudderth, M. I. Jordan, and A. S. Willsky, "An hdp-hmm for systems with state persistence," in *Proceeding on International Conference on Machine Learning (ICML)*, 2008, pp. 312–319.

[24] C. E. Rasmussen, "The infinite gaussian mixture model," in *In Advances in Neural Information Processing Systems*, 2000, pp. 554–560.

[25] P. Leveau, L. Daudet, and G. Richard, "Methodology and tools for the evaluation of automatic onset detection algorithms in music," in *Int. Symp. on Music Information Retrieval (ISMIR)*, 2004, pp. 72–75.

[26] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka, "Rwc music database: Popular, classical, and jazz music databases," in *Int. Symp. on Music Information Retrieval (ISMIR)*, 2002.

[27] K. Jensen, *Timbre Models of Musical Sounds*, Ph.D. thesis, University of Copenhagen, 1999.