PROBABILISTIC APPROACH TO AUTOMATIC MUSIC TRANSCRIPTION FROM AUDIO SIGNALS

Kenichi Miyamoto, Hirokazu Kameoka, Haruto Takeda, Takuya Nishimoto and Shigeki Sagayama

Graduate School of Information Science and Technology
The University of Tokyo
Hongo, Bunkyo-ku, Tokyo, 113-8656, Japan
{miyamoto, kameoka, takeda, nishi, sagayama}@hil.t.u-tokyo.ac.jp

ABSTRACT

We discuss automatic music transcription from audio input to music score by integrating our probabilistic approaches to multipitch spectral analysis, rhythm recognition and tempo estimation. In spectral analysis, acoustic energies in spectrogram are clustered into acoustic objects (i.e., music notes) with our method called Harmonic-Temporal-structured Clustering (HTC) utilizing EM algorithm over a structured Gaussian mixture with constraints of harmonic structure and temporal smoothness. After onset and offset timings are found from separated energies of music notes through note power envelope modeling to obtain the piano-roll representation, the rhythm and tempo are simultaneously recognized and estimated in terms of maximum posterior probability given a probabilistic note duration models with HMM (Hidden Markov Model) and probabilistic "rhythm vocabulary." Variable tempo is also modeled by a smooth analytic curve. Rhythm recognition and tempo estimation is alternately performed to iteratively maximize the joint posterior probability. Experimental results are also shown.

Index Terms— music transcription, harmonic-temporal-structured clustering, rhythm estimation, HMM

1. INTRODUCTION

Automatic music transcription has been one of ultimate goals of music information processing that converts audio signal of performed music into symbolic representation of music score similarly as in automatic speech recognition which converts speech into text. It has a wide range of potential applications including score display/printing, music analysis, input for automatic arrangement, music manipulation (e.g., changing the timbre) and music information retrieval (MIR) both in building music database and in transcribing the query input.

In this paper, we focus on retrieving music notes in polyphonic music audio signals from a single instrument such as piano, leaving out the rest of diverse problems related to music transcription including key signature, measure, music expressions (e.g., forte, piano), articulations (e.g., staccato, marcato), tempo (e.g., allegro, largo), tempo changes (e.g., ritardando, accelerando), style (e.g., a la marcia) and many other constituents of music score.

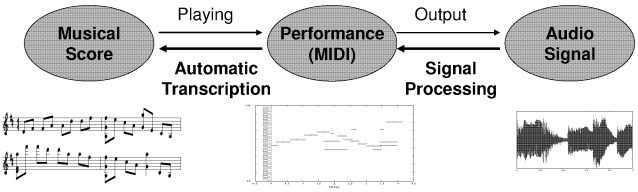


Fig. 1. Two stages for music transcription from audio signals

Music transcription can be divided into two subproblems as shown in Fig.1, i.e., signal processing to convert audio signal into piano-roll representation (nearly equivalent to the MIDI format) and automatic transcription to convert the piano-roll representation into music score.

The first stage is multipitch analysis and onset detection from audio signals, i.e., conversion of music audio signals into music performance data similar as MIDI (Musical Instrument Digital Interface) containing the pitch and onset time of each sound event. Most of previous works on multipitch estimation such as predominant F0 estimation method [4], *PreFEst* [1], Specmurt analysis [2] and so forth start by dealing with pitch extraction at each short-time frame using frequency-domain models and then try to find smooth pitch contours by interpolating or extrapolating the pitch features based on time evolution models. Contrary to such a strategy, Harmonic-Temporal structured Clustering (HTC) [9], that will be used as a front-end of the proposed method in this paper, tries to estimate simultaneously the spectral structure in both the time and frequency directions. Onset detection of each sound event is another problem to solve, and some methods use hierarchical approach [5] and Markov Chain Monte Carlo method [6].

The second stage is the rhythm and tempo estimation from estimated onset timings. To realize this stage, it can be effective the approach estimating rhythm pattern from MIDI or piano-roll data. Many methods for rhythm and tempo estimation that have been so far reported are based on a rule-based AI approach or graphical model (for example, see [6, 7]). In our previous work described in [10], we dealt with rhythm and tempo estimation as a simultaneous problem, whereas these previous methods treat them separately.

We propose in this paper a method for automatic music

transcription from audio signals to music score. The proposed method is an integration of the HTC multipitch estimation method [9] followed by the onset time reestimation, and the rhythm and tempo estimation method using HMM [10].

2. AUTOMATIC MUSIC TRANSCRIPTION

2.1. A Probabilistic Approach

Automatic transcription from audio signals can be considered as an inverse problem to estimate the source music score from the audio signal of performed music. In this paper, we focus on polyphonic sound from a single instrument such as piano, onset rhythm, and tempo curve as mentioned in the previous section. In this situation, given the audio spectrogram W of the input signal, maximum *a posteriori* estimation of the source score $\boldsymbol{S}$ consisting of note number $\boldsymbol{U}$, onset rhythm $\boldsymbol{Q}$, and tempo $\boldsymbol{R}$ is represented by the following probabilistic model:

$$\begin{aligned}&\underset{\boldsymbol{S}}{\operatorname{argmax}}\, P(\boldsymbol{S}|W) = \underset{\boldsymbol{U},\boldsymbol{Q},\boldsymbol{R}}{\operatorname{argmax}}\, P(\boldsymbol{U},\boldsymbol{Q},\boldsymbol{R}|W)\\ &= \underset{\boldsymbol{U},\boldsymbol{Q},\boldsymbol{R}}{\operatorname{argmax}} \int P(\boldsymbol{Q},\boldsymbol{R}|\boldsymbol{X})P(\boldsymbol{X},\boldsymbol{U}|\boldsymbol{\Theta},W)P(\boldsymbol{\Theta}|W)d\boldsymbol{\Theta}d\boldsymbol{X}\end{aligned}$$

where $\boldsymbol{X}$ denotes the onset time data and $\boldsymbol{\Theta}$ represents the set of the acoustic object parameters in HTC as intermediate parameters. To avoid the integral calculation over all possible ranges of intermediate parameters in the above equation this time, we adopt the following approximation in this paper:

$$\begin{aligned}P(\boldsymbol{\Theta}|W) &\approx \delta(\boldsymbol{\Theta} - \hat{\boldsymbol{\Theta}}(W))\\ P(\boldsymbol{X},\boldsymbol{U}|\boldsymbol{\Theta},W) &\approx \delta(X - \hat{X}(\boldsymbol{\Theta},W))\delta(U - \hat{U}(\boldsymbol{\Theta})).\end{aligned}$$

where $\delta(\cdot)$ denotes Dirac's delta function to yield the following approximation without using integration:

$$\underset{\boldsymbol{S}}{\operatorname{argmax}}\, P(\boldsymbol{S}|W) \approx \underset{\boldsymbol{Q},\boldsymbol{R}}{\operatorname{argmax}}\, P(\boldsymbol{Q},\boldsymbol{R}|\hat{X}(\hat{\boldsymbol{\Theta}},W),\hat{\boldsymbol{\Theta}}(W)). \quad (1)$$

2.2. Proposed Approach with a Multi-Step Process

By the approximated transcription model in Eq.(1), the automatic music transcription is realized by a multi-step process is shown in Fig.2 and described below:

(1) Calculate the spectrogram $W(x,t)$ from audio signal $f(t)$ by wavelet transform,
(2) Estimate the acoustic object parameter $\boldsymbol{\Theta}$ and obtain power envelope $Q_n(t)$ at each pitch using HTC,
(3) Estimate onset time dataset $\boldsymbol{X}$ at each acoustic object by the method proposed in following section
(4) Estimate onset rhythm pattern $\boldsymbol{Q}$ and tempo curve $\boldsymbol{R}$ by HMM-based rhythm and tempo estimation,
(5) Generate music score $\boldsymbol{S}$ from derived pitch and rhythm patterns.

3. PROPOSED PROBABILISTIC APPROACH

3.1. Decomposition of Spectral Energy using HTC

We will describe here in brief the multipitch estimation method, HTC [9]. Let the wavelet power spectrum of a music acoustic signal be $W(x,t)$, where x is log-frequency and t is time.

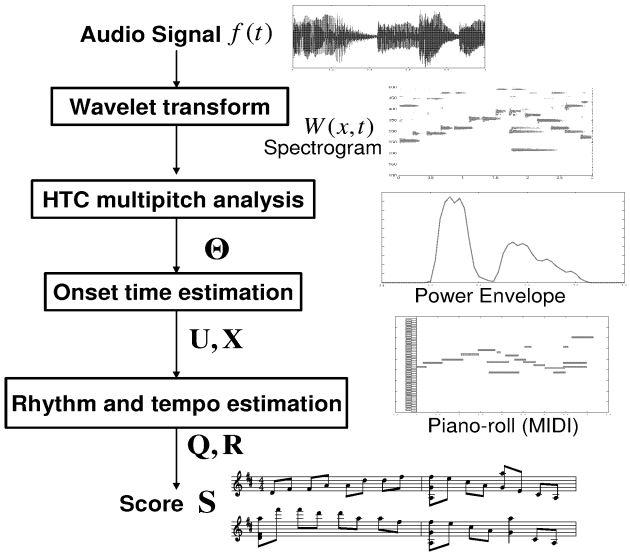


Fig. 2. Our proposed approach to automatic music transcription

The problem is to approximate it as well as possible as the sum of K parametric source models $q_k(x,t;\boldsymbol{\Theta})$, where $\boldsymbol{\Theta}$ is the set of model parameters, modeling the power spectrum of K "objects" each with a flat pitch contour μ_k.

As described in [9], these source models are expressed as a structured Gaussian mixture model with constraints on the characteristics of the kernel distributions: supposing that there is harmonicity with N partials modeled in the frequency direction, and that the power envelope is described using Y kernel functions in the time direction, we can write each source model as

$$q_k(x,t;\boldsymbol{\Theta}) = \sum_{n=1}^{N}\sum_{y=0}^{Y-1} S_{k,n,y}(x,t;\boldsymbol{\Theta}), \quad (2)$$

with kernel densities $S_{k,n,y}(x,t;\boldsymbol{\Theta})$ which are supposed to have the following shape:

$$S_{k,n,y}(x,t;\boldsymbol{\Theta}) \triangleq \frac{w_k v_{k,n} u_{k,n,y}}{2\pi\sigma_k\phi_k} e^{-\frac{(x-\mu_k-\log n)^2}{2\sigma_k^2} - \frac{(t-\tau_k-y\phi_k)^2}{2\phi_k^2}}$$

where the parameters w_k, $v_{k,n}$ and $u_{k,n,y}$ are normalized to unity. τ_k is supposed to correspond to the onset time, $v_{k,n}$ and $u_{k,n,y}$ the shapes of the spectral envelope and the power envelope of the kth source, respectively. A graphical representation of a HTC source model $q_k(x,t;\boldsymbol{\Theta})$ is shown in Fig. 3. HTC method iteratively searches for the optimal parameter set $\boldsymbol{\Theta}$ (μ_k, τ_k, ϕ_k, ψ_k, $v_{k,n}$ and $u_{k,y}$) that minimizes the sum over k of the KL-divergence between the kth source model $q_k(x,t;\boldsymbol{\Theta})$ and the corresponding spectral cluster $m_k(x,t)W(x,t)$:

$$J = \sum_k \iint m_k(x,t)W(x,t)\log\frac{m_k(x,t)W(x,t)}{q_k(x,t)}dxdt.$$

$m_k(x,t)$ is a spectral masking function, that we also want to estimate along with $\boldsymbol{\Theta}$. It is shown in [9] that minimizing J

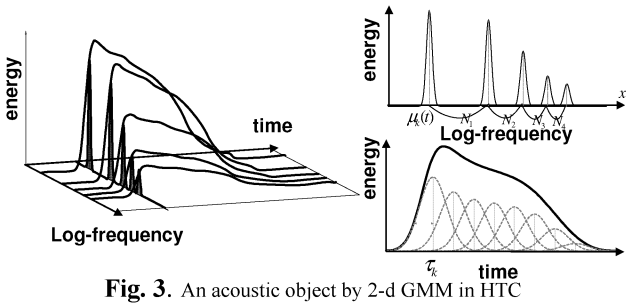


Fig. 3. An acoustic object by 2-d GMM in HTC

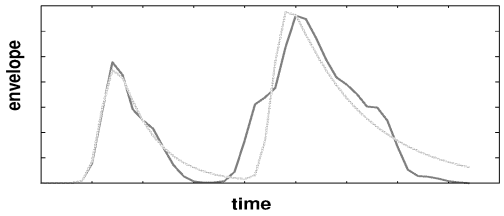


Fig. 4. An example of model fitting : the red curve represents obtained power envelope and the green curve represents two fitted model envelopes.

iteratively with respect to $m_k(x,t)$ and $\mathbf{\Theta}$ amounts to minimizing the KL divergence between the whole spectrogram $W(x,t)$ and $\sum_k q_k(x,t;\mathbf{\Theta})$. It is also shown that minimizing the KL divergence between $W(x,t)$ and $\sum_k q_k(x,t;\mathbf{\Theta})$ could be understood as maximizing the likelihood defined as the joint distribution of multinomial-like distributions and the problem can then be considered as a statistical maximum likelihood estimation. This leads us to introduce prior distributions for the parameters we wish to enforce constraints on. We will use here a Dirichlet prior distribution for $P(v_{k,1},\cdots,v_{k,N})$ and $P(u_{k,0},\cdots,u_{k,Y-1})$ whose maxima are taken when the spectral and power envelopes of the source model have particular shapes. These prior distributions are in particular very helpful for avoiding overfitting the source models to $W(x,t)$.

Letting $\widehat{m}_k(x,t)$ be the optimal spectral masking function and $\widehat{\mathbf{\Theta}}$ be the optimal model parameters estimated through HTC method, the spectral portion corresponding to the kth source is, according to [9], given by:

$$\tilde{q}_k(x,t;\widehat{\mathbf{\Theta}}) = \hat{m}_k(x,t)W(x,t) = \frac{q_k(x,t;\widehat{\mathbf{\Theta}})}{\sum_k q_k(x,t;\widehat{\mathbf{\Theta}})}W(x,t). \quad (3)$$

Thus, we obtain the power envelope of each of sound component of different fundamental frequencies.

3.2. Reestimation of Onsets

In order to have the following rhythm estimation work well, the onset time of each note X needs to be estimated as precisely as possible. Although τ_k in the HTC source model has been roughly considered as an onset time estimate, modeling the power envelope with GMM does not necessarily give sufficiently precise onset time of each note (though it should be emphasized that such a modeling has been advantageous in optimization). We will thus try here to refine the onset time estimate of each note using the spectral portion $\tilde{q}_k(x,t;\widehat{\mathbf{\Theta}})$ obtained with the HTC method.

First, the spectral portion $\tilde{q}_k(x,t;\widehat{\mathbf{\Theta}})$ can be combined at each note number obtained from the estimated fundamental frequency $\widehat{\mu}_k$, and the power envelope

$$Q_n(t) = \sum_{k\in C_n}\int \tilde{q}_k(x,t;\widehat{\mu}_k)dx \quad (4)$$

$$C_n = \left\{k \middle| A\left(n-\tfrac{1}{2}\right) \leqq \widehat{\mu}_k < A\left(n+\tfrac{1}{2}\right),\ k,n\in\mathbb{N}\right\}$$

can be obtained at each note number (pitch) n where $A = 100$[cents], i.e., energy summation over one semi-tone interval. Next, we discuss separation of power envelope $Q_n(t)$ into individual sound events and estimation of onset times. In this paper, we assume ideal-onset signal $g(t;\omega_0,\alpha,\tau,c)$ which has an ideal onset (onset time τ) and exponential decay (decay coefficient α), and single frequency ω_0:

$$g(t) = cu(t-\tau)e^{j\omega_0(t-\tau)}e^{-\alpha(t-\tau)}.$$

Its model envelope $\Psi(t;\tau,\alpha,C)$ can be obtained from wavelet transform of $g(t;\omega_0,\alpha,\tau,c)$ at frequency ω_0:

$$\Psi(t;\tau,\alpha,C) = Ce^{-2\alpha(t-\tau)}\left(\int_{-\infty}^{t-\tau-\frac{d\alpha}{2\omega_0^2}} e^{-\frac{\omega_0^2}{d}s^2}ds\right)^2,$$

and the n-th envelope $Q_n(t)$ in Eq.(4) can be approximated by a summation of L models. Parameters of the model envelopes can be obtained by minimizing the objective function

$$\int_{-\infty}^{\infty}\left|Q_n(t) - \sum_{l=1}^{L}\Psi(\alpha_l,\tau_l,C_l,t)\right|^2 dt. \quad (5)$$

The parameters α_l, τ_l and C_l is estimated by the following iteration:

1. optimizing $\mathbf{C} = (C_1,\cdots,C_L)^T$ with α_l,τ_l fixed
2. Updating α_l , τ_l by steepest decent with $\mathbf{C}$ fixed (determining the step size by linear search).

Its convergence to a local optimum is guaranteed as the objective function is non-increasing at each step. An example of the result of the model fitting can be seen in Fig.4. We finally obtain pitch frequency parameter $\widehat{\mu}_k$ and the refined estimate of the onset time τ_l.

As the result of this step, combining of note events whose pitch is given by HTC and onset time $\boldsymbol{X}$ given by estimated τ_l, a MIDI data can be obtained.

3.3. Estimation of Rhythm and Tempo using HMM

Since pitch of each note is already identified in the preceding steps, the music score can be obtained by estimating rhythm information including note values, time signature, and position of bar lines. We have been proposed a probabilistic top-down approach for this estimation[10], which estimates the rhythm score of polyphonic music from an IOI (Inter-onset interval) sequence of MIDI performance data as shown in Fig.5. Rhythm estimation can be formulated as MAP estimation, i.e., estimating the most likely rhythm score for the given IOI sequence. In our method, all of the aspects related to rhythm estimation such as fluctuation of tempo, deviation of onset timings, grammar and vocabulary of rhythm patterns are probabilistically modeled and integrated in the framework of HMMs (hidden Markov models) in the same manner as continuous speech recognition (CSR). Alternately executing

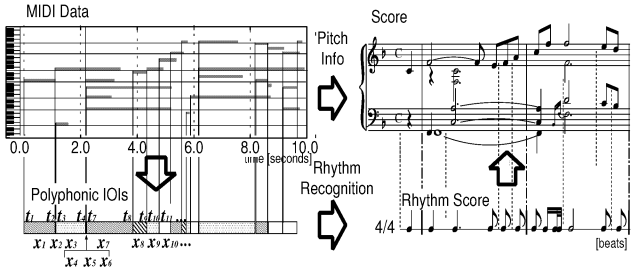


Fig. 5. Procedure of polyphonic rhythm and tempo estimation from IOIs

these optimizations converges to the simultaneously optimal rhythm pattern and tempo curve though not guaranteed to be the global optimum.

Finally, combining the informations of the note number estimated by the HTC and the rhythm pattern estimated by the HMM-based method, the most likely score for the given audio signals can be obtained. That is, automatic transcription is performed.

4. EXPERIMENTS AND DISCUSSION

4.1. Experimental Setup

We experimentally tested our approach to automatic music transcription from an audio input executing our proposed system. A waveform data of a piano piece by Bürgmuller performed by a human was sampled at 16kHz and fed to the system. In the experimental setup, 3-rd order polynomials were chosen for tempo curves, and 40 acoustic models per 6.4 secs in HTC. The number of envelope models in onset time estimation was determined automatically from the change in the power envelope. HMM parameters in rhythm and tempo estimation were trained with 137 MIDI data of piano pieces.

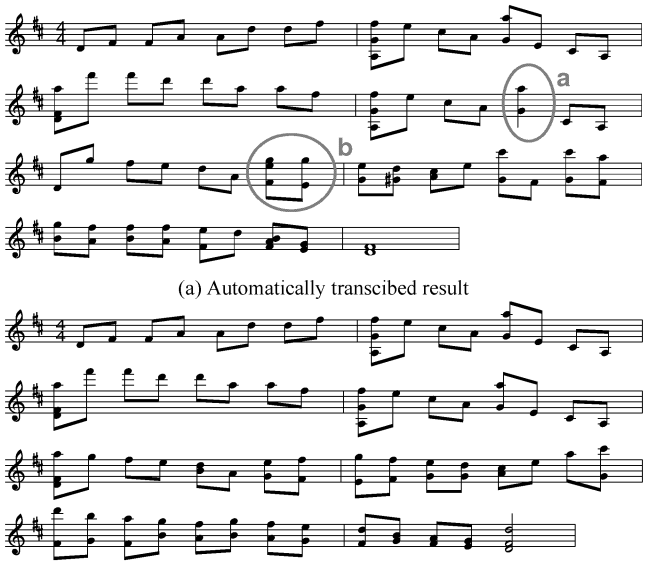
4.2. Experimental Results and Discussion

The yielded score from automatic transcription is shown in Fig. 6 with the key signature manually given. A nearly correct score was estimated successfully for the top two lines including chords. At point 'a' in Fig.6, the rhythm IOI was successfully estimated as a quarter note, while the following note was missing due to errors in approximation in HTC and onset detection. On the other hand, at point 'b' in the same figure, missing eighth notes caused errors in the bar line positions.

Note deletions are considered to be caused by fixed number of notes in HTC and onset time estimation, and may be improved by applying an information criterion such as Akaike's Information Criterion (AIC). Some of errors in HMM-based rhythm estimation are considered to be caused by mismatches between the erroneous input and the "rhythm vocabulary," and may be improved in the future taking into account the presence of 'noise,' i.e., deletions and insertions in the input onset sequence.

5. CONCLUSION AND FUTURE WORK

In this paper, we discussed automatic music transcription of audio signal into music score by integrating our approahces



(a) Automatically transcibed result

(b) Correct score (Burgmüller's piano studies Op.100-10.)

Fig. 6. A sample result of automatic transcription from audio input

of Harmonic-Temporal-structured Clustering (HTC) for multipitch analysis, HMM-based rhythm recognition and tempo estimation, and newly proposed onset time estimation. We experimentally evaluated the proposed system with audio inputs performed by piano. Future work will include thoroughly integrated framework for HTC incorporating musical constraints concerning rhythm and chord.

This research was partly supported by MEXT Grant-in-Aid #17300054 and CrestMuse Project of JST.

6. REFERENCES

[1] M. Goto *et al.*, "A real-time music-scene-description system: predoinant-F0 estimation for detecting melody and bass lines in real-world audio signals," *ICSA Journal*, Vol. 43, No. 4, pp. 311-329,2004.

[2] S. Sagayama *et al.*, "Specmurt Analysis: A Piano-Roll-Visualization of Polyphonic Music Signal by Deconvolusion of Log-Frequency Spectrum," *SAPA2004*, pages in CD-ROM, 2004.

[3] S. Godsill and M.Davy, "Bayesian Harmonic Models for Musical Pitch Estimation and Analysis," *Proc. ICASSP2002*, Vol. 2, pp. 1769-1772, 2002.

[4] A. Klapuri, "Multiple fundamental frequency estimation based on harmonicity and spectral smoothness," *IEEE Trans. Speech and Audio Proc.*, 11(6), 804-816, 2003.

[5] E. Kapanci and A. Pfeffer, "A Hierarchical Approach to Onset Detection," *Proceedings of the International Computer Music Conerence*, Miami, Florida, USA, 2004.

[6] E. Kapanci and A. Pfeffer, "Signal-to-Score Music Transcription using Graphical Models," *Proceedings of the 19th International Joint Conference on Artificial Intelligence (IJCAI)*, Edinburgh, UK, 2005.

[7] C. Raphael, "Automated Rhythm Transcription," *Proc. ISMIR*, pp. 99-107, 2001.

[8] A.T. Cemgil,*et al.*, "On Tempo Tracking: Tempogram Representaion and Kalman Filtering," JNMR, 28(4), pp.259-273, 2001.

[9] H. Kameoka *et al.*, "A Multipitch Analyzer Based on Harmonic Temporal Structured Clustering," *IEEE Trans. on Audio, Speech and Language Processing*, in press.

[10] H. Takeda *et al.*, "Rhythm and Tempo Analysis toward Automatic Music Transcription," submitted to ICASSP2007.