HARMONIC-TEMPORAL-TIMBRAL CLUSTERING (HTTC) FOR THE ANALYSIS OF MULTI-INSTRUMENT POLYPHONIC MUSIC SIGNALS

Kenichi Miyamoto, Hirokazu Kameoka†, Takuya Nishimoto, Nobutaka Ono and Shigeki Sagayama

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

ABSTRACT

In this paper, we discuss a new approach named Harmonic-Temporal-Timbral Clustering (HTTC) for the analysis of single-channel audio signal of multi-instrument polyphonic music to estimate the pitch, onset timing, power and duration of all the acoustic events and to classify them into timbre categories simultaneously. Each acoustic event is modeled by a harmonic structure and a smooth envelope both represented by Gaussian mixtures. Based on the similarity between these spectro-temporal structures, timbres are clustered to form timbre categories. The entire process is mathematically formulated as a minimization problem for the I-divergence between the HTTC parametric model and the observed spectrogram of the music audio signal to simultaneously update harmonic, temporal and timbral model parameters through the EM algorithm. Some experimental results are presented to discuss the performance of the algorithm.

Index Terms— analysis of multi-instrument music, EM algorithm, Harmonic-Temporal-Timbral Clustering (HTTC)

1. INTRODUCTION

Analysis of single channel multi-instrument music signal has been one of the ultimate goals of music signal processing, with the ambition of obtaining a total estimation of each acoustic event, including the information on the instrument or the timbre. This problem has a wide range of potential applications including music transcription with part division, part tracking, and music information retrieval (MIR). However, it has also been one of the most intricate problems, composed of several difficult sub-problems such as multipitch analysis and instrument recognition.

So far, motivated by the psychological theory of auditory scene analysis [1], we have developed a method for multipitch analysis called Harmonic-Temporal Clustering (HTC) [2]. HTC decomposes the spectral energy of the signal in the time-frequency domain into acoustic events, which are modeled using acoustic object models with a harmonic and temporal 2-dimensional structure. It is thus able to estimate information such as F_0 frequency, onset timing, etc., for each acoustic event. Unlike conventional frame-wise approaches such as [3, 4], HTC deals with the structures in both time and frequency directions simultaneously and shows high performance. In this paper, we present a new approach for multi-instrument music analysis which simultaneously realizes a clustering of the spectral energy into acoustic events and a classification of each acoustic event into timbre categories according to their similarity with regard to the timbre feature. This approach is developed as an extension of HTC.

Although instrument recognition, modeling of musical instrument sounds and multipitch analysis have been considered as difficult problems, some approaches dealing with polyphonic signal have recently been developed [6]. These methods usually handle the problem of instrument information analysis separately from multipitch analysis, resulting in a a two-step approach which first extracts the features for each acoustic event from the result of multipitch analysis, conducted as a prior processing, then performs classification in the timbre space [7, 8], training of the instrument sounds model [9] or identification of the source instrument through matching with learned timbre features [10, 11]. In another approach, a system visualizing frame-by-frame the probability of existence of each instrument frame-by-frame without estimating the F_0 information has also been developed [12].

However, we consider that multipitch analysis and timbre clustering should be realized simultaneously. From the estimation of the F_0 frequency of polyphonic music, one can derive a timbre clustering of the acoustic events. Conversely, estimation of the timbre structure for each acoustic event can give clues for the separation of overlapped harmonics from several events and reduce errors such as half-pitch or double-pitch errors, thus enabling higher performance of the multipitch analysis.

From another standpoint, when we listen to music performed with multiple instruments, we can usually naturally classify acoustic events into timbre categories according to similarity of their timbre feature, even if the music is played with unknown instruments. In this paper, our goal is to propose a computational method which realizes a learning system similar to what humans perform.

We name the unified analysis we propose for multi-instrument music Harmonic-Temporal-Timbral Clustering (HTTC).

†Now with NTT Communication Science Laboratories

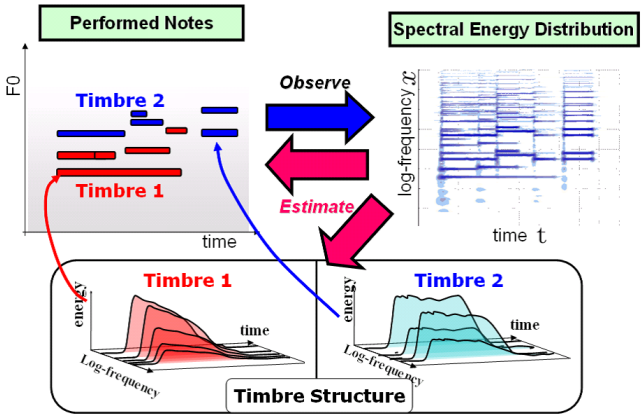

Fig. 1. Generative Model in HTTC

2. APPROACH FOR HTTC

2.1. Generative Model for the Spectral Energy

Consider the observed power spectrum time series $W(x,t)$ of a music acoustic signal, where x is log-frequency and t is time. This spectrum $W(x,t)$ is assumed to be generated as the sum of the spectral energy corresponding to acoustic source events performed with different onset time, pitch, power and duration and which belong to an unknown timbre category, as is shown in Fig.1.

Therefore, the problem we have to solve is an inverse problem to estimate the parameter set Θ which best approximates $W(x,t)$ as the sum of K parametric acoustic object models $q_k(x,t;\Theta)$ corresponding to spectral energy patterns originated from a single source. The parameter set Θ includes all the parameters of each acoustic event, such as F_0 frequency, onset timing and the timbre category the source belongs to, and the parameters of each timbre feature, which we shall define in the following section. Estimation is performed simultaneously on all the parameters.

2.2. Definition of Timbre

Although many definitions of timbre or instrument have been considered in previous works, it is still ambiguous. Some definitions treating three elements of music sounds, i.e., loudness, pitch and timbre, say that timbre is the total feature which does not depend on F_0 frequency and power. In addition, timbre feature can also be considered independent from the duration according to our general knowledge.

In this paper, timbre feature can be defined as the shape of spectral energy in time and log-frequency space which does not depend on F_0, spectral power, onset, and duration. For example, spectral energy shapes for the piano and the violin are shown in Fig.2. There are many differences between these two shapes, both in the spectral power of the partials and the temporal structure. Therefore we can consider that the difference in timbre is derived from the difference of shape of the spectral energy, and that the shapes of acoustic events classified into the same timbre category should look alike regardless of the pitch, power, onset timing and duration.

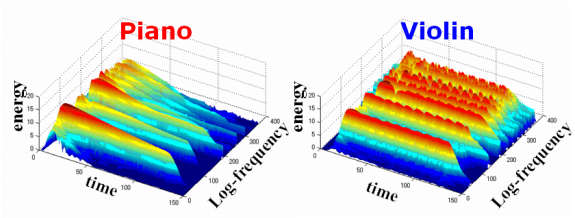

Fig. 2. Example of Timbre Structure. left: piano, right: violin

We shall note that this definition of timbre is simple and does not aim at expressing all the features of timbre in a single instrument. Timbre clustering in HTTC might thus face some limitations as an instrument recognition method, and make some mistakes as we humans sometimes do too.

2.3. Justification of our approach

A very interesting point in music is that acoustic events from the same instrument but with different durations are recognized as belonging to the same timbre category. This shows that a simple pattern clustering in the time-frequency domain can not account for timbre, and that the clustering of the timbre categories and the estimation of the audio object parameters (including duration) need to be performed simultaneously. This justifies our approach to perform jointly the clustering of the timbre categories and the estimation of the audio object parameters.

3. HTTC MODEL

3.1. Parametric Model for the Timbre Structure

First, as we defined in the previous section, the timbre feature is expressed in time-frequency space and can be modeled as a parametric structure corresponding to the energy distribution originated from a single source. Assuming that we are dealing with pitched instruments in this paper, the source energy is characterized by its harmonic structure and its time duration. We model this structure with a 2-dimensional Gaussian Mixture Model (GMM) distribution $T_c(x,t;\Theta)$ described as:

$$T_c(x,t;\Theta) = \sum_{n,y} T_{c,n,y}(x,t;\Theta) \quad (1)$$

$$T_{c,n,y}(x,t;\Theta) = \frac{v_{c,n}u_{c,n,y}}{2\pi\sigma\phi} e^{-\frac{(x-\log n)^2}{2\sigma^2}-\frac{(t-y\phi)^2}{2\phi^2}} \quad (2)$$

$$\forall c, \sum_n v_{c,n} = 1, \quad \forall c, \forall n, \sum_y u_{c,n,y} = 1, \quad (3)$$

where c denotes the index of the timbre category, n is the index of the partial, y is the index of the GMM for time duration, $v_{c,n}$ and $u_{c,y}$ denote variables corresponding to the relative energy of the partials and the time envelope respectively in the c-th timbre structure, and ϕ and ψ denote the frequency and time spreads of every 2-dimensional Gaussian respectively, which are considered constant in the HTTC model. The model's shape is shown in Fig. 3,

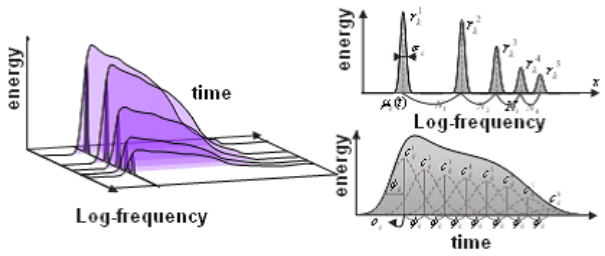

Fig. 3. Timbre Structure. left: the entire shape, upper right: GMM in the log-frequency direction, lower right: GMM for the time duration

Table 1. Parameters of the HTTC Model

denotation	physical meanings
w_k	total energy of k-th source
μ_k	pitch contour of k-th source
τ_k	onset time of k-th source
γ_k	duration of k-th source
c_k	timbre category which k-th source belongs to
$v_{c,n}$	relative energy of n-th harmonic in c-th timbre
$u_{c,y}$	coefficient of the power envelope of c-th timbre

3.2. Acoustic Object Model using Timbre Structure

Next we describe the acoustic object model corresponding to the energy of a single acoustic source. The 2-dimensional shape of each object model can be expressed with the timbre structure $T_c(x,t;\Theta)$ modeled in the previous section. Therefore, according to the definition of timbre, in addition to the timbre category c_k deciding the shape of the object model, the k-th acoustic object has the pitch contour μ_k, the onset time τ_k, the total energy w_k, and the duration γ_k as parameters which are independent from the timbre structure.

Finally, in order to control the time duration of each object independently of the time envelope of the timbre structure, the model can multiply the parameters $u_{c_k,y}$ of the time envelope by a time muting function $R(t;\gamma_k)$ which is equal to 1 before the duration γ_k and then drops to 0 as the sigmoid function $\frac{1}{1+e^{p(t-\gamma_k)}}$.

Altogether, the k-th acoustic object model can be described as

$$q_k(x,t;\Theta) = w_k \sum_{n,y} \frac{1}{1+e^{p(y\phi-\gamma_k)}} \frac{v_{c_k,n}u_{c_k,y}}{2\pi\sigma\phi} e^{-\frac{(x-\mu_k-\log n)^2}{2\sigma^2}-\frac{(t-\tau_k-y\phi)}{2\phi^2}}. \quad (4)$$

4. UPDATING PARAMETERS

4.1. Decomposition of Spectral Energy

As we explained in the previous section, the problem we have to solve is to estimate the model parameters Θ of each acoustic source and each timbre structure listed in Table 1 from the observed energy pattern $W(x,t)$. It can be solved through an EM-like algorithm, introducing spectral masking functions $m_k(x,t)$ which decompose $W(x,t)$ into acoustic objects at each coordinate (x,t) ($0 \le m_k(x,t) \le 1, \sum_k m_k(x,t) = 1$), and optimizing the model parameters and the mask functions iteratively as proposed in our previous work [2]. Since the partitioned cluster $m_k(x,t)W(x,t)$ is expected to be the spectral energy distribution of a single acoustic source, this decomposion of energy is beneficial not only for the estimation of the model parameters depending on each acoustic source but also for timbre clustering and estimation of the timbre structure which needs the shape of a single acoustic source.

4.2. Minimization of the Objective Function Using the EM Algorithm

In order to estimate the model parameters, we introduce as objective function the sum of the distances, measured using the I-divergence, between the partitioned clusters $m_k(x,t)W(x,t)$ and the acoustic source models $q_k(x,t;\Theta)$, including the timbre structure $T_{c_k}(x,t;\Theta)$:

$$J = \sum_k J_k \quad (5)$$

$$J_k = \iint_D m_k(x,t)W(x,t)\log\frac{m_k(x,t)W(x,t)}{q_k(x,t;\Theta)} - (m_k(x,t)W(x,t) - q_k(x,t;\Theta))dxdt. \quad (6)$$

Our problem can be regarded as the minimization of (5).

The minimization of J can be realized by the iteration of the update rule shown in Fig.4 and described below:

(1) Update the mask functions $m_k(x,t)$
(2) Estimate the parameters depending on each acoustic object such as w_k, μ_k, τ_k, and γ_k
(3) Estimate the timbre category c_k of each object from J_k as the discriminant function
(4) Estimate the GMM parameters $v_{c,n}$, $u_{c,y}$ representing the timbre structure.

Since each step of this update rule can reduce the objective function (5) successfully, the iteration of these update steps can yield to locally optimal parameters. For length purposes, we skip here the description of the details of the update equations for each parameters, which can be obtained analytically by the combination of an undetermined multipliers Lagrange's method and a k-means-like algorithm for the update of c_k.

5. EXPERIMENTS AND DISCUSSION

5.1. Experimental Setup

We experimentally tested our approach, HTTC, from an audio input. An input waveform data of about 10 seconds in a violin sonata by Frank performed with violin and piano was obtained by converting the MIDI data into WAV data sampled at 16kHz. In addition, the initial pitches and onset times of each acoustic source model was set correctly in order this time to discuss the performance of timbre clustering and learned timbre shape specifically. The input piano-roll data and the converted wavelet spectral energy are shown in Fig.5.

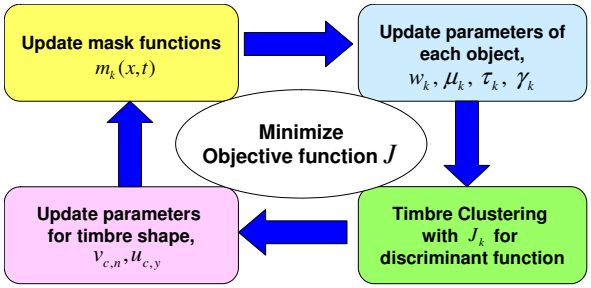

Fig. 4. 4-step Algorithm Minimizing the Objective Function

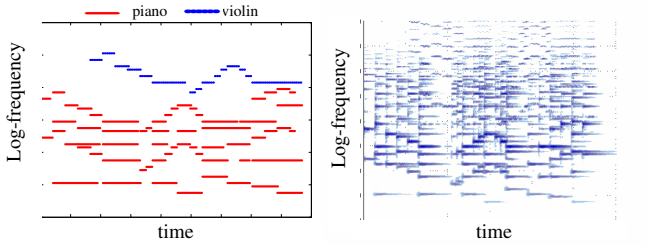

Fig. 5. Input Data. left: input piano-roll data, right: obtained spectral energy distribution

5.2. Experimental Results and Discussion

5.2.1. Result with the number of timbre categories set to 2

First, we tested our algorithm with the number of timbre set to 2. The obtained piano-roll data and timbre shape are shown in Fig.6. We can see in the figure that many notes of the violin and some notes of the piano in the low frequencies were classified into the same timbre category. We think that the reason for this is that the harmonic structures are different between notes with different F_0 within a single instrument.

5.2.2. Result with the number of timbre categories set to 3

We tested next the same input signal with the number of timbre categories set to 3. The obtained piano-roll data and timbre shape are shown in Fig.7. In this result, most of the events belonging to the violin were classified into the third category (blue line) and few notes of the piano in the low frequencies were classified into it. Therefore, we think that notes of the violin and the piano were well classified into different timbre categories.

6. CONCLUSION AND FUTURE WORK

In this paper, we proposed a new approach for multi-instrument musical analysis, called Harmonic-Temporal-Timbral Clustering (HTTC), and we experimentally evaluated the performance of HTTC with a music signal consisting of piano and violin. Future work will include the design of a timbre structure model depending on F_0 for a better instrument recognition.

This research was partly supported by MEXT Grant-in-Aid #17300054.

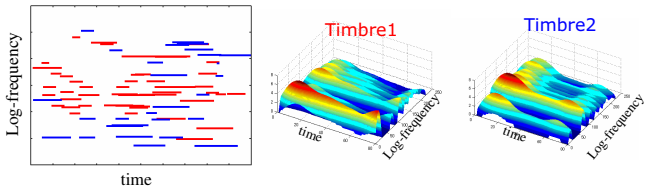

Fig. 6. Experimental result with 2 timbre categories

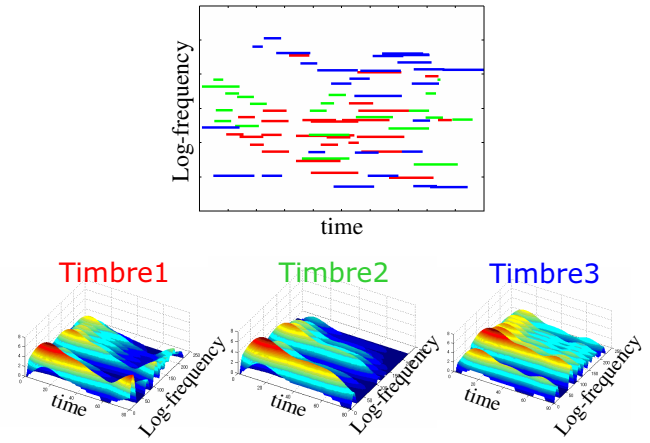

Fig. 7. Experimental result with 3 timbre categories

7. REFERENCES

[1] A. S. Bregman, Auditory Scene Analysis, MIT Press, Cambridge, 1990.

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

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

[4] M.Goto, "A Robust Predominant-F_0 Estimation Method for Real-time Detection of Melody and and bass lines in CD recordings," in *Proc. ICASSP*,2000.

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

[6] T. Kitahara, "Computational Musical Instrument Recognition and Its Application to Content-based Music Information Retrieval," Ph. D. Thesis, Kyoto University, 2007.

[7] T. Kitahara, *et al.* ,"Musical Instrument Identification based on F_0-dependent Multivariate Normal Distribution," *Proc. ICASSP2003*, Vol. 5, pp421-424, 2003.

[8] T. Kitahara, *et al.* ,"Category-level Identification of Non-registered Musical Instrument Sounds," *Proc. ICASSP2004*, Vol. 4, pp253-256, 2004.

[9] A. Klapuri, "Analysis of Musical Instrument Sounds by Source-Filter-Decay Model," *Proc, ICASSP*, 2007.

[10] J. Eggink, G. J. Brown, "A Missing Feature Approach to Instrument Identification in Polyphonic Music," *Proc, ICASSP*, 2003.

[11] K. Kashino, Hiroshi Murase,"A sound source identification system for ensemble music based on template adaptation and music stream extraction," *Speech Communication*,vol.27, 1999.

[12] T. Kitahara, M. Goto, K. Komatani, T. Ogata, and H. G. Okuno, "Instrogram: Probabilistic Representation of Instrument Existence for Polyphonic Music", *IPSJ Journal*, Vol. 48, No. 1, pp. 214–226, 2007.