Flexible Harmonic Temporal Structure for Modeling Musical Instrument

Jun Wu, Yu Kitano, Takuya Nishimoto, Nobutaka Ono, Shigeki Sagayama

Flexible Harmonic Temporal Structure for modeling musical instrument

Jun Wu, Yu Kitano, Takuya Nishimoto, Nobutaka Ono and Shigeki Sagayama
The Graduate School of Information Science and Technology, University of Tokyo
Tokyo 113–8656, Japan
{wu,kitano,nishi,onono,sagayama}@hil.t.u-tokyo.ac.jp

Abstract. Multipitch estimation is an important and difficult problem in entertainment computing. In this paper a flexible harmonic temporal structure for modeling musical instrument was proposed for estimating pitch in real music. Unlike the previous research, the proposed model does multipitch estimation according to the specific characteristics of specific musical instrument and uses EM algorithm to estimate the parameters in the model. Through choosing parameters suitable for its own characters for specific instrument, the proposed model preponderated over the common model.

1 Introduction

Multipitch estimation is one of the fast growing topic in MIR in which most of the research going on now is using signal processing method. A multipitch tracking method in noisy environment by filter bank process and pitch tracking using HMM was proposed by Wu et al. [1]. Goto presented a method to track F0 of objective single sound from polyphonic musical signals without restriction of the number of simultaneous sounds [2]. Some other multipitch analyzers such as graphical model-based [3], filterbank-based [4], nonparametric Kalman filtering-based [5], [6]. Then a method for multipitch analysis called Harmonic-Temporal Clustering (HTC) was proposed [7] to deal with the harmonic and temporal structures in both time and frequency directions and shows high performance. However, all of these algorithms cannot do multipitch estimation according to the specific characteristics of specific musical instrument, which is actually important and meaningful for real music. In this paper, a flexible harmonic temporal structure for modeling musical instrument is proposed in section 2. The proposed model is based on the clustering principle and uses EM algorithm to estimate each mean parameter. In section 3, the experimental results were demonstrated.

2 Flexible harmonic temporal structure

We propose model $q_k(x,t;\theta)$ to be the model for a single note in the music, where x is log-frequency, t is time and θ is the parameter in the model. It is composed of the fundamental partial and harmonic partials. The normalized energy density of the nth

partial in the *k*th source model can be assumed to be a multiplication of the power envelope of the *n*th partial $U_{k,n}(t)$ and the Gaussian distribution centered at $\mu_k\,(\mathrm{t}) + \log(n)$, $U_{k,n}(t) \times \frac{v_{k,n}}{\sqrt{2\pi}\sigma_k} e^{-(x-\mu_k(t)-logn)^2/2\sigma_k^2}$ $n = 1, \dots, N$ (1)

$\mu_k\,(\mathrm{t})$ is pitch contour of the *k*th source, $v_{k,n}$ is relative energy of *n*th partial in *k*th source. Let the frequency spread of each harmonic component be approximated by a Gaussian distribution function when the spectra are obtained by the wavelet transform (constant Q transform) using Gabor wavelet basis function. Denote $U_{k,n}(t)$ as the power envelope of the *n*th partial.

$$U_{k,n}(t) = \sum_{\forall y} \frac{u_{k,n,y}}{\sqrt{2\pi\phi_{k,n}{}^2}} exp\left\{-\frac{(t-\tau_k-y\phi_{k,n,y})^2}{2\phi_{k,n}{}^2}\right\} \quad (2)$$

τ_k is the center of the forefront Gaussian, which is considered as an onset time estimate, $u_{k,n,y}$ is the weight parameter for each kernel, which allows the function to have variable shapes for each harmonic partial.

$u_{k,n,y}$should be normalized to satisfy $\forall k, \forall y: \sum_y u_{k,n,y}\,(x,t) = 1$. $\phi_{k,n,y}$ is the distance between the centers of the Gaussian function kernels. The power spectrogram structures of different instruments are very different. So we set $\phi_{k,n,y-1} = \alpha\phi_{k,n,y-1}$ to give flexibility to the distance between the Gaussian function kernels. If the instrument's power spectrogram structure is relatively steep at the beginning part, the model is able to choose larger α which means more Gaussian function kernels at the steep beginning part.

The source models $q_k(x,t;\theta)$ are expressed as a mixture of Gaussian mixture model (GMM) with constraints on the kernel distributions: supposing that there is harmonicity with *N* partials modeled in the frequency direction, and the power envelope is described using *Y* kernel distribution in the time direction. The source model can be written in the form

$$q_k(x,t;\theta) = \sum_n \sum_y S_{k,n,y}(x,t;\theta) \quad (3)$$

And the Kernel distribution can be written in the form

$$S_{k,n,y}(x,t;\theta) = \frac{w_k v_{k,n} u_{k,n,y}}{2\pi\delta_k\emptyset_k} e^{-\frac{(x-\mu_k(t)-logn)^2}{2\sigma_k^2} - \frac{(t-\tau_k-y\emptyset_{k,n})^2}{2\phi_{k,n,y}^2}} \quad (4)$$

w_k is the energy of the *k*th source. Therefore the source model $q_k(x,t;\theta)$ is the mixture of mixture of Gaussian distribution $S_{k,n,y}(x,t;\theta)$. And the whole model is the mixture of the source model $q_k(x,t;\theta)$.

The proposed the algorithm uses EM procedure for the parameter estimation procedure. We assume that the energy density *W(x;t)* has an unknown fuzzy membership to the *k*th source, introduced as a spectral masking function $m_k(x,t)$. To minimize the difference between the observed power spectrogram time series *W(x;t)* and the model $\sum_k q_k(x,t;\theta)$, we use the Kullback–Leibler (KL) divergence as the global cost function.

$$J = \sum_k \iint_D m_k(x,t) W(x;t) log \frac{m_k(x,t)W(x;t)}{q_k(x,t;\theta)} \quad (5)$$

Satisfying with:

$\forall x, \forall t, \sum_k m_k\,(x,t) = 1, 0 < m_k\,(x,t) < 1.$

Then the problem is regarded as the minimization of (5).

The membership degree $m_k(x,t)$(spectral masking function) of *k*th source/stream

can be considered to be the weight of the *k*th source model in the whole spectrogram model. It is unknown at the begging and need to be estimated. On the other hand, the spectrogram of the *k*th source can be modeled by a function $q_k(x,t;\theta)$, where θ is the set of model parameters. They are also unknown variables. The proposed model works by using EM algorithm for iteratively updating of: E-step: $m_k(x,t)$ with θ fixed and M-step: θ with $m_k(x,t)$ fixed. $m_{k,n,y}(x,t)$ is masking function.

The E-step is realized by the following equation.

$$m_k(x,t)m_{k,n,y}(x,t) = \frac{s_{k,n,y}(x,t;\theta)}{\sum_k \sum_n \sum_y s_{k,n,y}(x,t;\theta)} \quad (6)$$

The update of parameters can be obtained analytically by undetermined multipliers Lagrange's method.

3 Experiments

To evaluate the proposed flexible harmonic temporal structure, we tested it with 151 music instrument pieces (including 4 instruments: 36 guitar pieces, 45 violin pieces, 36 flute pieces and 34 oboe pieces) chosen from the RWC music database. [8] Since the RWC database also includes the MIDI files associated with each real-performed music signal data, we evaluated the accuracy by comparing the estimated fundamental frequency and the MIDI files. The proposed algorithm preponderates over the HTC [7] approach for 23.3% for guitar signal, 2.8 % for oboe signal, 12.5 % for flute signal and 3.1% for violin signal.

References

1. M.Wu, D. Wang and G. J. Brown, "A Multi-pitch Tracking Algorithm for Noisy Speech," ICASSP2002, Vol. 1, pp. 369–372, 1995.
2. M.Goto, "A Predominant-F0 Estimation Method for CD Recordings: MAP Estimation Using EM Algorithm for Adaptive Tone Models," Proc. ICASSP2001, Vol. 5, pp. 3365–3368, Sep 2001.
3. K.Kashino, K.Nakadai, and H.Tanaka, "Organization of hierarchical perceptual sounds: Music scene analysis with autonomous processing modules and a quantitive information integration mechanism," in Proc. IJCAI , 1995, vol. 1, pp. 158–164.
4. A.Klapuri, T.Virtanen, and J.Holm, "Robust multipitch estimation for the analysis and manipulation of polyphonic musical signals," in Proc. COST-G6 Conf. Digital Audio Effects, 2000, pp. 233–236.
5. K.Nishi and S.Ando, "Optimum harmonics tracking filter for auditory scene analysis," in Proc. IEEE ICASSP '96, 1996, vol. 1, pp. 573–576.
6. M.Abe and S.Ando, "Auditory scene analysis based on time-frequency integration of shared FM and AM (II): Optimum time-domain integration and stream sound reconstruction," (in Japanese) Trans. IEICE, vol. J83-D-II, no. 2, pp. 468–477, 2000.
7. Hirokazu Kameoka, Takuya Nishimoto, Shigeki Sagayama, ``A Multipitch Analyzer Based on Harmonic Temporal Structured Clustering," IEEE Trans. on Audio, Speech and Language Processing, Vol. 15, No. 3, pp.982-994, Mar., 2007.
8. M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka, "RWC music database: Popular, classical, and jazz music database," in Proc. ISMIR, 2002, pp. 287–288