Harmonic Temporal Model for Multiple Pitch Estimation

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

ABSTRACT

This paper describes a system for the Multiple Fundamental Frequency Estimation and Tracking task in MIREX (Music Information Retrieval Evaluation eXchange) 2010. It decomposes the spectral energy of the signal in the time-frequency domain into pitch models. Each pitch is modelled by 3-dimensional Gaussian Mixture structure. EM algorithm is used to estimate the parameters in the model.

1. INTRODUCTION

The goal of multiple pitch estimation is to estimate fundamental frequencies of multiple harmonic signals simultaneously present in an input musical signal. It is considered a difficult problem mainly due to large overlap between the overtones of different pitches—a phenomenon common in Western music, where combinations of sounds that share partials are preferred for their pleasant sound.

HTC is a multipitch analyzer proposed in [1]. It decomposes the energy patterns of observed power spectrum into clusters such that each of them represents a single source and then can extract the note events such as fundamental frequency, intensity, onset and duration of notes from polyphonic audio signals. The sources are modeled by superimposed HTC source models, which is a harmonically constrained Gaussian mixture. HTC try to fit mixture of the source models to observed power spectrum by updating model parameters and clustering the energy patterns using EM algorithm.

2. MODEL DESCRIPTION

HTC model approximates the observed power spectrogram of input music signal *W(x;t)* (where *x* is the log-frequency and *t* is the time) with a sum of *K* parametric models, each of which represents a single continuous pitch in the input signal. Every pitch model is composed of a fundamental partial (F0) and *N* harmonic partials. Since we do not know in advance what the signal sources are, it is important for the pitch model to be as flexible as possible. It should also be nonnegative since we model a nonnegative power spectra. We have chosen to use a mixture of Gaussian functions due to its simplicity and flexibility. The list of model's parameters is shown in Table 1.

Table 1. Parameters of Harmonic Temporal Model

parameter	Physical meaning
$\mu_k(t)$	Pitch contour of the *k*th pitch
w_k	Energy of the *k*th pitch
$v_{k,n}$	Relative energy of *n*th partial in *k*th pitch
$u_{k,n,y}$	Coefficient of the power envelop function of *k*th model, *n*th partial, *y*th kernel
τ_k	Onset time
$Y\phi_{k,n,y}$	duration(Y is constant)
σ_k	Diffusion in the frequency direction of the harmonics

We have employed the EM algorithm to estimate all of the model's parameters. We assume that the energy density W(x;t) has an unknown fuzzy membership to the kth model, introduced as a spectral masking function . To minimize the difference between the observed power spectrogram time series W(x;t) and the pitch model , we use the Kullback–Leibler (KL) divergence as the global cost function;

$$J = \sum_k \iint_D m_k(x,t) W(x;t) log \frac{m_k(x,t)W(x;t)}{q_k(x,t;\theta)} \quad (1)$$

under the constraint;

$$\sum_k m_k(x,t) = 1, 0 < m_k(x,t) < 1, \forall x, \forall t, \quad (2)$$

The M-step can be realized by the iteration of the update the parameters depending on each acoustic object, which can be obtained analytically by the combination of an undetermined multipliers Lagrange's method.

3. IMPLEMENTATION

The system is implemented in C with standard C library. Performance of this system varies depending on the complexity of audio input, especially on the number of sources. The more the number of sources exists in an input, the longer time to estimate model parameters tends to become. It is because usually one source model is needed to fit one source on the input power spectrum. This system is originally intended to be an converter from audio signal to MIDI data. Therefore it assumes that the fundamental frequency of a source is constant while the

note is active. Fluctuation of the frequency is ignored. And also output of fundamental frequency is quantized at the frequency of each note number of MIDI. Frame length of spectrum produced with Gabor Wavelet Transform is 10ms and frequency range is between 50Hz and about 2.5kHz.

4. REFERENCES

[1] 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.