HARMONIC TEMPORAL STRUCTURED CLUSTERING FOR MULTIPLE FUNDAMENTAL FREQUENCY ESTIMATION

Koji Egashira
Graduate School of
Information Science and Technology,
University of Tokyo, Japan
egashira@hil.t.u-tokyo.ac.jp

Hirokazu Kameoka
Media Information Laboratory,
NTT Communication
Science Laboratories, Japan
kameoka@eye.brl.ntt.co.jp

Shigeki Sagayama
Graduate School of
Information Science and Technology,
University of Tokyo, Japan
sagayama@hil.t.u-tokyo.ac.jp

ABSTRACT

This abstract describes a method for the Multiple Fundamental Frequency (F0) Estimation and Tracking task in the Music Information Retrieval Evaluation eXchange (MIREX) 2007. The method is called Harmonic Temporal Structured Clustering (HTC), which is a kind of constrained Gaussian Mixture Model (GMM) estimation using EM algorithm. It can jointly extract F0, intensity, onset, duration of each underlying source in a monaural polyphonic audio signal with superimposed HTC source models.

1 INTRODUCTION

HTC is a multiple-F0 analyzer proposed in [1]. Most of other analyzers start estimation with feature extraction of input audio signal at each short-time segment and then find the most likely F0 trajectories along time. In contrast, HTC does both the feature extraction and the F0 trajectories estimation not independently but in a cooperative way for more reliable results.

HTC decomposes the energy patterns of observed power spectrum into clusters such that each of them represents a single source and then extract the note events such as F0, intensity, onset and duration of notes from polyphonic audio signals. The sources are modeled by HTC source models, which is a mixture of 2-d Gaussian constrained harmonically at frequency and continuously at time. HTC tries to fit mixture of the source models to observed power spectrum by updating model parameters and clustering the energy patterns using EM algorithm.

2 METHOD DESCRIPTION

The outline of the algorithm is as follows (For details about formulation and derivation, please see [1]). First, monaural audio signal is taken as input and the power spectrum of it is obtained using Gabor Wavelet Transform.

Second, model parameters of GMM are estimated using EM algorithm. A spectral masking function is introduced for clustering of energy patterns of the spectrum,

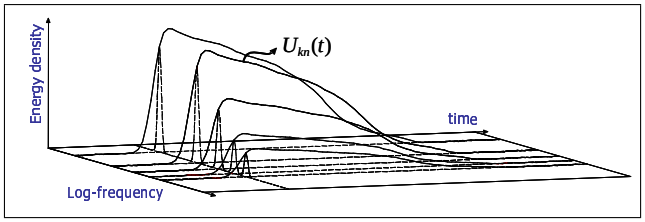

Figure 1. Profile of the k-th HTC source model $S_k(x,t)$

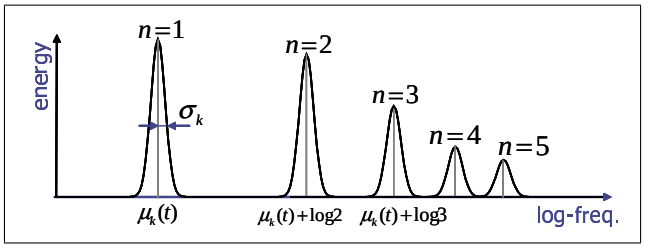

Figure 2. A cutting plane of $S_k(x,t)$ at time t

indicating active area of each source in the whole spectrum, as a hidden variable. The decomposed spectrum by the masking function is then modeled by one or more HTC source models (Fig. 1). Objective function of the algorithm to be minimized is sum of the Kullback-Leibler (KL) divergence between each masked spectrum and source models. The model parameters are shown in Table 1 (k, n, y are index of the source model, index of harmonic and index of Gaussian kernel in time series, respectively). Unknown variables in the objective function to be estimated are the spectral masking functions, which are unobservable, and the set of parameters of HTC source model. In E-step of the EM algorithm, the masking function is estimated with model parameters fixed. In M-step, on the other hand, the model parameters are updated with the masking functions fixed. When these variables are converged to a stationary point, an optimal result will be obtained.

3 IMPLEMENTATION

The submitted system to the task is implemented in C with standard C library (no other libraries). This system is orig-

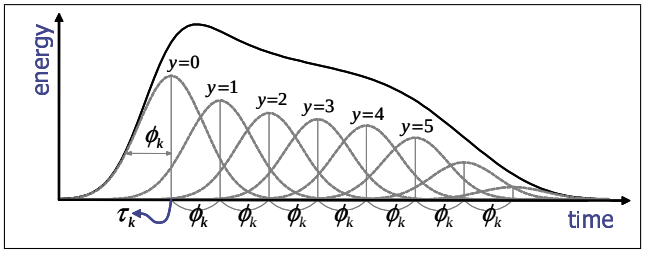

Figure 3. Approximation of temporal power envelope with mixture of Gaussian

parameter	meaning
w_k	total energy of the *k*-th source
v_{kn}	relative energy of *n*-th harmonic
u_{kny}	*y*-th weight of Gaussian in time direction
τ_k	onset time
μ_k	F0 of the source
$Y\phi_k$	duration (Y is constant)
σ_k	diffusion in the frequency direction

Table 1. The meaning of HTC model parameters

inally intended to be an converter from audio signal to MIDI data. Therefore it assumes that F0 of a source is constant while the note is active. Fluctuation of the frequency is ignored. And also output of F0 is quantized at the frequency of each note number of MIDI.

Frame shift of the spectrum produced with Gabor Wavelet Transform is about 10ms and frequency range is between 50Hz and about 4kHz.

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, 2007.