HARMONIC TEMPORAL STRUCTURED CLUSTERING WITH UNSUPERVISED MODEL LEARNING FOR MULTIPITCH ESTIMATION

Masahiro Nakano, Koji Egashira, Nobutaka Ono, Shigeki Sagayama
Graduate School of Information Science and Technology,
University of Tokyo, Tokyo 113-8656, Japan
{ mnakano, egashira, 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) 2009. The method is a modification of Harmonic Temporal Structured Clustering(HTC), which is a kind of constrained Gaussian Mixture Model estimation using EM algorithm. In the modification one term of a function of model power is added to objective function to fit input power spectrogram with possible fewest Gaussian kernels.We submitted this system to the same task in MIREX 2007.

1 INTRODUCTION

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.

The algorithm described in [1] has no care on how to initialize source models. The number of source models, which is difficult to assume before starting estimation process, has to be provided for it. Fewer models than actually needed tend to increase estimation error.

Then the algorithm of HTC was modified so that much many source models are given at initialization and unnecessary models are deleted during the updating of estimates. To realize the idea, one term of a function of model power was added to the objective function of EM algorithm. This modification become very similar to what is described in [2]. The modified algorithm no longer needs to know the number of necessary models.

variable	meaning
x	log-frequency
t	time
$W(x,t)$	observed power spectrum
k	index of source model
$m(k;x,t)$	spectral masking function
$S_k(x,t)$	HTC source model
n	index of harmonic
y	index of Gaussian kernel in time series

Table 1. The meaning of variables

2 METHOD DESCRIPTION

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

Second, model parameters of Gaussian mixture models are estimated using EM algorithm. A spectral masking function is introduced for clustering of energy patterns of the spectrum, indicating active area of each source in the spectrum, as a hidden variable. Objective function of the algorithm to be minimized is the Kullback-Leibler (KL) divergence of input power spectrum and mixture of source models (1).

$$J = \sum_k \iint m(k;x,t)W(x,t)\log \frac{m(k;x,t)W(x,t)}{S_k(x,t)}dxdt \\ +\chi \sum_k \log w_k \tag{1}$$

HTC source model $S_k(x,t)$ is denoted as (2).

$$S_k(x,t) = \sum_n \sum_y \frac{w_k v_{kn} u_{kny}}{2\pi\sigma_k\phi_k} e^{\left(-\frac{(x-\mu_k(t)-\log n)^2}{2\sigma_k^2} - \frac{(t-\tau_k-y\phi_k)^2}{2\phi_k^2}\right)} \tag{2}$$

The meaning of variables and parameters later is shown in Table 1 and Table 2.
The last term (sum of log of each source model power) is the added one in the modification, which works to reduce source models. χ is the weight of the term relative to the rest.

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 power envelope function
τ_k	onset time
$\mu_k(t)$	fundamental frequency of the source
$Y\phi_k$	duration (Y is constant)
σ_k	diffusion in the frequency direction

Table 2. The meaning of parameters

In E-step of the EM algorithm, spectral masking function, which is unobservable, is estimated with model parameters fixed. In M-step, on the other hand, the set of model parameters of HTC source model are updated with spectral masking function fixed.

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, 2007.

[2] Mario A.T. Figueiredo, Anil K.Jain, ”Unsupervised Learning of Finite Mixture Models”, *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 24, No. 3, 2002.