# SEQUENTIAL ESTIMATION OF MULTIPLE FUNDAMENTAL FREQUENCIES THROUGH HARMONIC-TEMPORAL-STRUCTURED CLUSTERING

**Koji Egashira, Nobutaka Ono, Shigeki Sagayama**
Graduate School of Information Science and Technology, University of Tokyo, Japan
{egashira, onono, sagayama}@hil.t.u-tokyo.ac.jp

**ABSTRACT**

This paper describes a system for the Multiple Fundamental Frequency (F0) Estimation and Tracking task in MIREX (Music Information Retrieval Evaluation eXchange) 2008. The system is based on modeling energy distribution of a sound source on the time-frequency plane (frequency domain is log-frequency axis) and estimates F0's through representing the spectrogram of input signals by mixture of the sound source models. We submitted a similar system to the same task in MIREX 2007. The system submitted this time has many improvements including sequential F0 estimation and adjustment of the number of sound source models.

## 1 ALGORITHM

### 1.1 Harmonic-Temporal-structured Clustering

The core of our system is a multiple-F0 estimation technique called Harmonic-Temporal-structured Clustering (HTC) [1], as well as the system submitted last year. It uses a 3-dimensional parametric model of energy distribution of a harmonic sound source, which has harmonically constrained and temporally smooth structure and is governed by parameters meaning the intensity, F0, onset time, offset time, relative weight of each harmonics, etc. of each model. HTC tries to represent the spectrogram of input signals via minimization of dissimilarity between the spectrogram and the mixture model. Difference of two distributions can be measured using I-divergence, therefore HTC minimize the I-divergence of the two energy distributions. Minimization is achieved using auxiliary variables and auxiliary functions through iterative update of model parameters like EM algorithm. F0, onset time and other information of estimated each sound source is obtained after convergence of the minimization.

### 1.2 adjusting the number of models

One of the improvements in this year's system is an adjustment function of the number of sound source models. This is necessary because the amount of active sounds in input signals is unknown. To estimate the number of sound sources, their F0 and onset time simultaneously, an approach that extra models are deleted after much many models are given is used. We realize this by adding a penalty term to the I-divergence, choosing summation of log of each model energy as the penalty term.

### 1.3 sequential estimation

Second improvement is to add the sequential estimation technique. Instead of scattering many sound source models over the whole time-frequency plane, a window with short time range, set at the beginning of the spectrogram at first and moving to the end little by little, is prepared to focus on a portion of the whole spectrogram, only in which the models are put and fit into the input spectrogram. After the onset of each sound is found using the method explained in previous section, each model is successively extended until the corresponding sound trails off and its energy decreases, considering temporal smoothness of energy. This method can reduce computational cost and waiting time until beginning of the estimation results are obtained.

The process of our system is as follows: First, a spectrogram of input signals is obtained by wavelet analysis using Gabor-wavelet basis functions. Second, set the window at the beginning of the spectrogram, put many sound source models in the window and fit mixture of the models into the input spectrogram by iteratively updating model parameters to decrease an objective function consisting of summation of I-divergence and the penalty term. Then, move the window a little and execute the model fitting procedure again. Model parameters are fixed after the window passes, which are the information obtained as the estimation result. Repeat fitting the models and moving the window until the end of the spectrogram.

## 2 REFERENCES

[1] H. Kameoka, T. Nishimoto, and S. Sagayama, "A Multipitch Analyzer Based on Harmonic Temporal Structured Clustering," *IEEE Transactions on Speech and Audio Processing*, vol. 15, no. 3, pp. 982–994, 2007.