SEQUENTIAL ESTIMATION OF MULTIPLE FUNDAMENTAL FREQUENCY THROUGH HARMONIC-TEMPORAL CLUSTERING

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 system is based on modeling energy distribution of a sound source on the time-frequency plane and estimates F0's through representing the spectrogram of input signals by mixture of sound source models. The method successively models the energy from one sound source at the next frame on the spectrogram of input signal, concerning simultaneously both harmonic features and temporal smoothness of energy based on Harmonic-Temporal Clustering. To recognize the onset of a new note and assign the corresponding sound source model to it, we introduced the idea of sparseness where the spectrogram is represented by some excessively prepared sound source models. We submitted this system to the same task in MIREX 2008.

1 METHOD DESCRIPTION

1.1 Harmonic-Temporal-structured Clustering

Our system is based on a multiple-F0 estimation technique called Harmonic-Temporal-structured Clustering (HTC)[1]. The harmonic structure is modeled by a constrained Gaussian Mixture Model. $q_k[x,t;\Theta_k]$, the energy of the k-th source model governed by parameter vector Θ_k at each coordinate (x,t) in the spectrogram (x and t are log-frequency and time), can be written as

$$q_k[x,t;\Theta_k] \stackrel{\text{def}}{=} w_k[t]\sum_n \frac{v_{k,n}}{\sqrt{2\pi}\sigma_k}\exp\left(-\frac{(x-\mu_k-\log n)^2}{2\sigma_k^2}\right) \quad (1)$$

where the parameter $w_k[t]$ is the energy of the source models at t-th frame of time, v_{kn} the relative energy of n-th harmonic ($\forall k\colon \sum_n v_{kn}=1$), σ_k the diffusion in the frequency direction of the harmonics, μ_k the F_0 of the sound source, respectively.

HTC tries to represent the spectrogram of input signals via minimization of dissimilarity between the spectrogram and the mixture model.Differnce of two distributions can be measured using I-divergence.

$$I \stackrel{\text{def}}{=} \sum_{x,t}\{W[x,t]\log\frac{W[x,t]}{\sum_k q_k[x,t;\Theta_k]} -(W[x,t]-\sum_k q_k[x,t;\Theta_k])\} \quad (2)$$

1.2 adjusting the number of models

The amount of active sounds in input signals is unknown. To estimate the number of sound source, their F0 and onset time simultaneously, an approach that extra models are deleted after much many models are given is used. We achieve this by adding a penalty term to the I-divergence, choosing summation of log of each model energy as the penalty term.

$$J_{sp} \stackrel{\text{def}}{=} \sum_{\substack{\forall\ k,t\ s.t.\\ w_k[t]>0}} \log w_k[t] \quad (3)$$

1.3 sequential estimation

Our system uses the sequential method of HTC. This method can reduce computational cost and waiting time until beginning of the estimation results are obtained. Instead of scattering many sound source models over the whole time-frequency plane, a window with short time range, set the beginning of 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 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. We assume that the gradient of energy between adjoining two frames in time domain is normally distributed. Consequently the cost of the smoothness can be

$$J_{sm} \stackrel{\text{def}}{=} \frac{1}{2\phi^2}\sum_{k,t}\left(w_k[t+1]-w_k[t]\right)^2 \quad (4)$$

where ϕ is the variance of the gradient. Finally, the objective function to be minimized in this algorithm is obtained

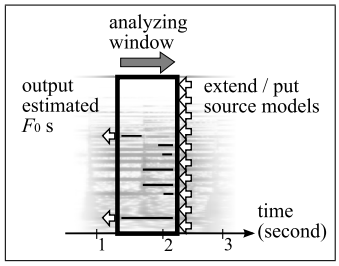
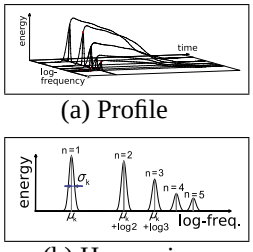


Figure 1. Analyzing window moving on an input spectrogram



(a) Profile

(b) Harmonic constraints

Figure 2. Source model

as

$$
\begin{aligned}
J &\stackrel{\text{def}}{=} I + \chi J_{sp} + J_{sm} \\
&= \sum_{x,t} \{ W[x,t] \log \frac{W[x,t]}{\sum_k q_k[x,t]} \\
&-(W[x,t] - \sum_k q_k[x,t;\Theta_k]) \} \\
&+\chi \sum_{\substack{\forall\ k,t\ s.t. \\ w_k[t]>0}} \log w_k[t] \\
&+\frac{1}{2\phi^2} \sum_{k,t} \left(w_k[t+1] - w_k[t] \right)^2
\end{aligned}
\tag{5}
$$

where χ is a constant determining weight of J_{sp} compared to the other terms.

The outline of the algorithm is as follows.

1. Mono audio is taken as input and the spectrum of it is obtained using Gabor Wavelet Transform.
2. Input the first several frame of the spectrogram into beginning of the analyzing windows (Figure 1).
3. Put the source models (Figure 2) at every scale on area in the analyzing window where the spectrogram is newly input.
4. Update model parameters iteratively to decrease the objective function in the analyzing window.
5. Output estimation result at the last of the analyzing window.
6. End if the analyzing window has scanned all the spectrogram, otherwise input the next several frames into the analyzing window and back to 3.

2 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.