Extraction of Multiple Fundamental Frequencies from Polyphonic Music Using Harmonic Clustering

Hirokazu Kameoka, Takuya Nishimoto, Shigeki Sagayama

Graduate School of Information Science and Technology,
The University of Tokyo, Japan
{kameoka,nishi,sagayama}@hil.t.u-tokyo.ac.jp

Abstract

In this paper, a method for extracting fundamental frequencies (F_0s) from single channel input signal of concurrent sounds is described. By considering that an observed spectral density distribution is a statistical distribution of (imaginary) micro-energies, we attempt to classify them into each sound by the use of clustering principle. We call this approach a "Harmonic Clustering." One of the formulation of this clustering can be expressed in same way as a maximum likelihood of Gaussian mixture model (GMM) using EM algorithm. Our algorithm enables to estimate not only F_0s but also a number and each spectral envelope of underlying harmonic structure on the basis of an information criterion. It operates without restriction of a number of mixed sounds and a variety of sound sources, and extracts F_0s as accurate values with spectral domain procedures. Experimental results showed high performance of our algorithm.

1. Introduction

Multi-pitch estimation (MPE) technique of a single channel input is surely practicable in many applications such as automatic transcription of music, sound source identification, signal-to-MIDI converter, audio coding, audio enhancement and robust speech recognition.

In early attempts for MPE, the aim was to materialize automatic transcription of music but were firmly restricted in regard to a variety of instruments, a number of simultaneous sounds, and also a range and a resolution of extractable pitch. Lately, Kashino et al. proposed a method which can transcribe polyphonic music even if there were several kinds of instruments included [1].

Recently, MPE which enables to detect various information (such as number of simultaneous sounds, spectral envelopes, etc.) at the same time accompanied with F_0s has been put stress for multi purposes, and numerous methods have been reported mainly in musical signal processing [2, 3], speech signal processing [5, 6] and auditory scene analysis [7, 8]. Goto presented a method for extracting objective single sound from polyphonic musical signals without restriction of the number of simultaneous sounds [2]. This method offers an optimal spectral envelope of the single sound by introducing a priori distrubution. Chazan addressed a speech separation method by introducing a time warped signal model which allows a continuous pitch variations within a long analysis frame [6] and Klapuri described a robust MPE method [3] and Virtanen constructed a sophisticated sound separation system by implicating it [4]. These two methods are prominent in respect of a capability of extracting not only the amplitudes of the partials but even the phases due to a parameter estimation of time domain signal model.

However, most of the previous methods still have not included a specific process of detecting the number of simultaneous sounds. Our objective is to develop a new algorithm which detects not only F_0s but also a number of simultaneous sounds and spectral envelopes respectively as the solutions of an optimization problem.

2. Harmonic Clustering

2.1. General Formulation of Spectral Clustering

An influence of a window function and a varying pitch within the short time single analysis frame inevitably cause widening of the spectral harmonics which makes it difficult to extract the precise value of F_0s and partial energies. We consider each widened partial as a statistical distribution of micro-energies and attempt to classify them into several harmonic structures by use of clustering principle. This principle is based on a formation of a harmonic cluster which consists of several tied clusters, constrained by a representative centroid. In log-frequency scale, cluster centroids contained in the kth harmonic cluster are denoted here as $\boldsymbol{\mu}_k = \{\mu_k, \cdots, \mu_k + \log n, \cdots, \mu_k + \log N_k\}$ where μ_k is the representative centroid and is expected to be the F_0 of kth sound and n denotes the index of partials. Therefore, classifying micro-energies with K harmonic clusters is identical with separating spectral components into K harmonic structures. If we denote a distance between one centroid $\mu_k + \log n$ and one micro-energy positioned in log-frequency x as $d(x, \mu_k + \log n)$, a membership degree as $p_n^k(x)$, and the number of micro-energies as $f(x)$, i.e., a

spectral density, the total distance function, which should be minimized, is represented as

$$J = \sum_{k=1}^{K} \sum_{n=1}^{N_k} \int_{-\infty}^{\infty} f(x) p_n^k(x) d(x, \mu_k + \log n) dx \qquad (1)$$

where N_k denotes the number of clusters in the kth harmonic cluster. This function can be minimized by k-means algorithm, for instance, if $d(x, \mu_k + \log n)$ is the square of euclidean distance and $p_n^k(x)$ is the rectangularly divided uniform distribution.

2.2. Correspondence with Q-function

The objective function in equation (1) can also be interpreted as a Q-function of EM algorithm. First we model a kth harmonic structure with several Gaussian distributions, whose means are constrained by the representative mean μ_k and denoted as $\boldsymbol{\mu}_k = \{\mu_k, \cdots, \mu_k + \log n, \cdots, \mu_k + \log N_k\}$. We briefly call this model as harmonic-GMM (Gaussian mixture model). We then introduce a model of multiple harmonic structures $P_\theta(x)$ which is a mixture of K harmonic-GMMs whose model parameter $\boldsymbol{\theta}$ is denoted as

$$\{\theta\} = \{\boldsymbol{\mu}_k, \boldsymbol{w}_k, \sigma \mid k = 1, \cdots, K\}, \qquad (2)$$

where $\boldsymbol{w}_k = \{w_1^k, \cdots, w_n^k, \cdots, w_{N_k}^k\}$ and σ indicate the weights and variance (which is assumed here as a constant) of the respective Gaussian distributions.

As we consider that the normalized spectral density function $f(x)$ is a probability distribution of frequencies (events) which are generated from the model of multiple harmonic structures, the log-likelihood difference in accordance with an update of the model parameter $\boldsymbol{\theta}$ to $\bar{\boldsymbol{\theta}}$ is

$$f(x) \log P_{\bar{\theta}}(x) - f(x) \log P_\theta(x) = f(x) \log \frac{P_{\bar{\theta}}(x)}{P_\theta(x)}. \quad (3)$$

Dempster formulated EM algorithm [9] in order to maximize the mean log-likelihood by taking expectation of both sides with respect to $P_\theta(n, k|x)$ which represents the probability of the $\{n, k\}$-labeled Gaussian distribution from which x is generated. The Q-function will be derived in the right-hand side and given as

$$Q(\theta, \bar{\theta}) = \sum_{k=1}^{K} \sum_{n=1}^{N_k} \int_{-\infty}^{\infty} P_\theta(n, k|x) f(x) \log P_{\bar{\theta}}(x, n, k) dx. \quad (4)$$

In comparison with equation (1), it can be considered to be one of the objective function formulated above.

2.3. Model Parameter Estimation by EM Algorithm

Since an inequality is derived as

$$\int_{-\infty}^{\infty} \left\{ f(x) \log P_{\bar{\theta}}(x) - f(x) \log P_\theta(x) \right\} dx \geq Q(\theta, \bar{\theta}) - Q(\theta, \theta), \qquad (5)$$

the log-likelihood of the model of multiple harmonic structures with respect to every x will be monotonously increased by obtaining $\bar{\theta}$ which maximizes the Q-function. A posteriori probability $P_\theta(n, k|x)$ in equation (4) is given as

$$P_\theta(n, k|x) = \frac{P_\theta(x, n, k)}{P_\theta(x)}, \qquad (6)$$

$$= \frac{w_n^k \cdot g(x|\mu_k + \log n, \sigma^2)}{\sum_n \sum_k w_n^k \cdot g(x|\mu_k + \log n, \sigma^2)}, \qquad (7)$$

$$g(x|x_0, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{(x - x_0)^2}{2\sigma^2} \right\}, \qquad (8)$$

where $g(x|x_0, \sigma^2)$ is a Gaussian distribution. By the iterative procedure of E (*Expectation*)-step and M (*Maximization*)-step, the model parameter $\boldsymbol{\theta}$ locally converges to ML estimates.

3. F_0 Extraction Algorithm

The F_0 extraction scheme as a whole consists of two processes. In 3.1, we adopt one of the most widely used information criterion on which both processes described in 3.2 and 3.3, are based.

3.1. Criterion of Model Selection

Provided multiple different model candidates exist, the optimal model must somehow be judged. Here we introduce Akaike Information Criterion (AIC) which was proposed by Akaike in 1973 [10]. AIC is given by

$$\begin{aligned} \text{AIC} = &-2 \times (\text{maximum log-likelihood of model}) \\ &+2 \times (\text{number of free parameters of model}), \end{aligned} \quad (9)$$

whose minimum offers a proper estimate of the number of free parameters.

3.2. Estimation of a Number of Harmonic Structures

It is generally known that ML estimates firmly depend on initial values and may often converge to undesirable values. To avoid this, we first prepare extra amount of harmonic-GMMs in the model in order to raise possibility of obtaining the true values. Then, obviously, the model may overfit the given observed specrum. If one Gaussian is enough for approximating the shape of one partial, the same number of underlying harmonic structures must be enough with the harmonic-GMMs. And this number can be estimated by reducing harmonic-GMM one after another until they become the proper number on the basis of AIC. Although there may be exception, we assume the number of harmonic structures as a number of simultaneous sounds. The specific operation is as follows:

1. Set initial values of $\{\mu_1, \cdots, \mu_K\}$ in the limited frequency range.
2. Estimate the ML model parameters by EM algorithm. However, w_n^k is constrained here as

$$w_1^k = w_2^k = \cdots = w_{N_k}^k (= w^k). \qquad (10)$$

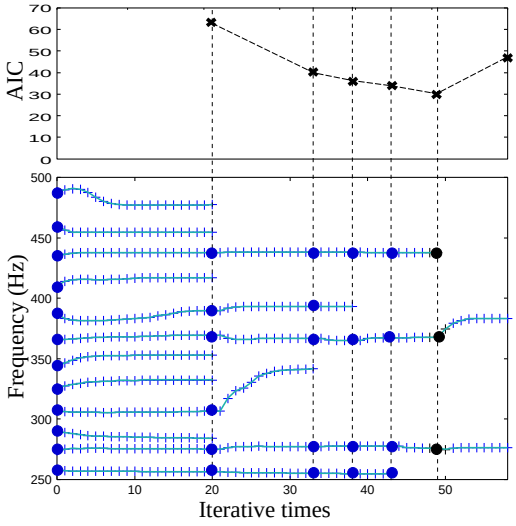

Figure 1: Updates of representative means μ_k

This w^k represents the degree of predominance of kth harmonic-GMM. In the M-step, model parameters μ_k and w^k should be updated to

$$\bar{\mu}_k = \frac{\sum_{n=1}^{N_k} \int_{-\infty}^{\infty} (x - \log n) P_\theta(n,k|x) f(x) dx}{\sum_{n=1}^{N_k} \int_{-\infty}^{\infty} P_\theta(n,k|x) f(x) dx} \quad (11)$$

$$\bar{w}^k = \frac{1}{F N_k} \sum_{n=1}^{N_k} \int_{-\infty}^{\infty} P_\theta(n,k|x) dx, \quad (12)$$

where F is an integral of $f(x)$ with respect to x.

3. Calculate AIC. Since there are two free parameters for each harmonic-GMM, the model has $2 \times K$ free parameters altogether. The number of harmonic-GMMs when AIC takes minimum corresponds to the number of simultaneous sounds.

4. Remove the harmonic-GMM(s) which conforms either of the two conditions as below and repeat from step 2.

   - The one whose w^k is the minimum among all. Since the contribution to the maximum log-likelihood must be the least.
   - The one whose w^k is smaller if the two adjacent representative means become closer than a certain distance (threshold). Since the two representative means are presumed to converge to the same optimal solution.

An example of how this process actually operates is shown in Fig.1. The broken line represents the point where the model parameters were judged to be converged and the line graph indicates the value of AIC at each point. Since AIC takes minimum when three harmonic-GMMs remain, the estimate number here is 3.

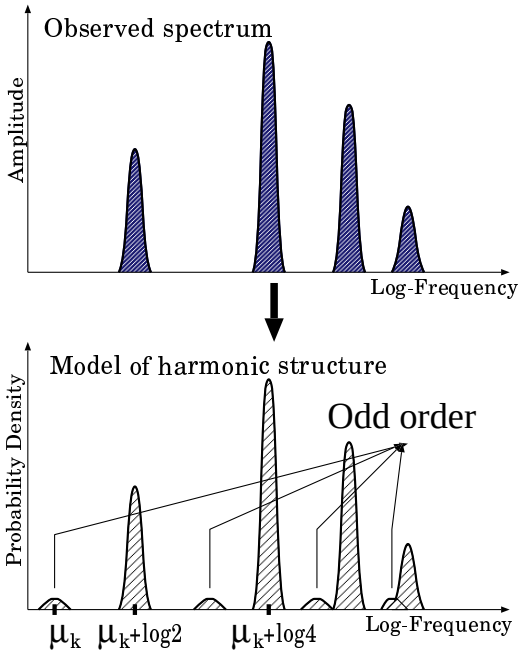

Figure 2: A harmonic-GMM when μ_k is $1/2$ of true F_0

3.3. Extraction of F_0 and Spectral Envelope

In the previous process, the ML procedure allows to aqcuire local optimal solutions of μ_k without distinction of the true F_0s, harmonics or subharmonics. Therefore, the true F_0s must somehow be discovered by replacing μ_k each by each to their multiples. Consider now that a degree of freedom is given to every w_n^k and consequently allows to extract the spectral envelope, i.e., the relative amplitudes of the partials. If μ_k corresponds to subharmonics, the model must overfit (Fig 2). From this point of view, the problem of obtaining the true F_0s and the spectral envelope can also be handled with the information criterion. The process shown below is done with all remaining harmonic-GMMs after the previous process.

1. Replace the representative means to $\mu_k + \log t$ where t is an integer number whose initial value is 1. The number of Gaussians limited below the Nyquist log-frequency is denoted as N_k^t.

2. Estimate the ML model parameters by EM algorithm. Here we only update w_n^k and should be updated to

$$\bar{w}_n^k = \frac{1}{F} \int_{-\infty}^{\infty} P_\theta(n,k|x) dx. \quad (13)$$

3. Calculate AIC. The number of free parameters here is N_k^t. The place of the representative mean when AIC takes minimum is considered as the F_0 estimate, and if not, add 1 to t and return to step1.

4. Operation Experiments

Experiments were carried out to validate the performance of our algorithm with polyphonic music by evaluating the

Table 1: Results for polyphonic music

Expermiental data		Accuracy(%)
Composer & Title	Instruments	
J. Pahelbel: “Kanon”	Violins	92.7
J. S. Bach: “Menuet”	Piano	74.9

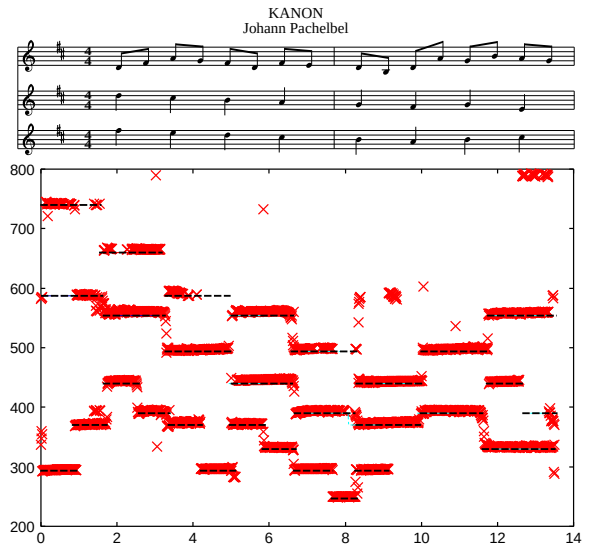

Figure 3: A F_0 detection result and the corresponding score of “Kanon”

accuracy of F_0 extraction.

4.1. Experimental Condition

The algorithm was tested on 2 pieces of music which were performed by three violinists and a pianist, respectively. All musical signals were digitized at 44.1 kHz and analyzed with Hamming window where frame length and frame shift were 46 ms and 10 ms. Reference F_0s were hand-labeled according to the notes and the durations transcribed in the musical score. Detection accuracy was calculated as a percentage of frames at which F_0s were correctly detected. The initial number of the harmonic-GMMs was set to 5, the frequency range was from 108 Hz to 215 Hz, and σ was assigned to 0.53.

4.2. Results for Recorded Polyphonic Music

The accuracy rates are shown in table 2 and the example of F_0 detection result is shown in Fig 3. The results showed that the algorithm worked well with the violin performance. As for the piano performance, though fast decay of piano sound made the detection difficult, F_0s before its decay were mostly extracted properly.

5. Conclusions

We proposed an algorithm which enables to detect a number of underlying harmonic structures and respective F_0s and spectral envelopes as the solutions of an optimal problem. It showed a high performance for recorded polyphonic music. Still, several improvements are prospective by applying temporal information available, incorporating variance into the model parameters also as a variable or by introducing a priori probability distribution of the model parameters, etc.

6. References

[1] K. Kashino, K. Nakadai, T. Kinoshita and H. Tanaka, “Organization of Hierarchical Perceptual Sounds: Music Scene Analysis with Autonomous Processing Modules and a Quantitive Information Integration Mechanism,” *Proc. IJCAI*, Vol. 1, pp. 158–164, 1995.

[2] M. Goto, “A Predominant-F0 Estimation Method for CD Recordings: MAP Estimation Using EM Algorithm for Adaptive Tone Models,” *Proc. ICASSP2001*, Vol. 5, pp. 3365–3368, Sep 2001.

[3] A. Klapuri, T. Virtanen and J. Holm, “Robust Multipitch Estimation for the Analysis and Manipulation of Polyphonic Musical Signals,” *In Proc. COST-G6 Conference on Digital Audio Effects*, pp. 233–236, 2000.

[4] T. Virtanen and A. Klapuri, “Separation of Harmonic Sounds Using Linear Models for the Overtone Series,” *Proc. ICASSP2002*, Vol. 2, pp. 1757–1760, 2002.

[5] M. Wu, D. Wang and G. J. Brown, “A Multipitch Tracking Algorithm for Noisy Speech,” *ICASSP2002*, Vol. 1, pp. 369–372, 1995.

[6] D. Chazan, Y. Stettiner and D. Malah, “Optimal Multi-pitch Estimation Using the EM Algorithm for Co-channel Speech Separation,” *Proc. ICASSP93*, Vol. 2, pp. 728–731, 1993.

[7] K. Nishi, M. Abe and S. Ando, “Multiple Pitch Tracking and Harmonic Segregation Algorithm for Auditory Scene Analysis,” *Trans. SICE*, Vol. 34, No. 6, pp. 483–490, 1998, (in Japanese).

[8] M. Abe and S. Ando, “Auditory Scene Analysis Based on Time-Frequency Integration of Shared FM and AM (II): Optimum Time-Domain Integration and Stream Sound Reconstruction,” *Trans. IEICE*, Vol. J83-D-II, No. 2, pp. 468–477, 2000, (in Japanese).

[9] A. P. Dempster, N. M. Laird and D. B. Rubin, “Maximum Likelihood from Incomplete Data via the EM Algorithm,” *J. of Royal StatisticalSociety Series B*, Vol. 39, pp. 1-38, 1977.

[10] H. Akaike, “Information Theory and an Extension of the Maximum Likelihood Principle,” *2nd Inter. Symp. on Information Theory*, Akademia Kiado, Budapest, pp. 267–281, 1973.