SEPARATION OF HARMONIC STRUCTURES BASED ON TIED GAUSSIAN MIXTURE MODEL AND INFORMATION CRITERION FOR CONCURRENT SOUNDS

Hirokazu Kameoka, Takuya Nishimoto and Shigeki Sagayama

Graduate School of Information Science and Technology, The University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo, 113-8656, Japan
{kameoka,nishi,sagayama}@hil.t.u-tokyo.ac.jp

ABSTRACT

In this paper, a method for separation of harmonic structures of co-channel input concurrent sounds is described. A model for multiple harmonic structures is constructed with a mixture of tied Gaussian mixtures, from which a single harmonic structure is modeled. Our algorithm enables estimation of both the number and the shape of the underlying harmonic structures, based on a maximum likelihood estimation of the model parameters using EM algorithm and an information criterion. It operates without restriction on the number of mixed sounds and varieties of sound sources, and extracts accurate fundamental frequencies continuously with simple procedures in spectral domain. Experiments showed high performance of the algorithm for both simultaneous speech and polyphonic music.

1. INTRODUCTION

Co-channel sound separation technique plays an important role in many applications such as automatic transcription of music, sound source identification, audio coding, audio enhancement and robust speech recognition. However, separation from a single input is a complicated inverse problem that is difficult to be solved analytically. Multi-pitch estimation (MPE) technique has been generally taken as one of the most useful approach for it.

In early attempts for MPE, the aim was automatic transcription of music but most of them were limited in regard to varieties of instruments, number of simultaneous sounds, and range and resolution of extractable pitch. Lately, however, Kashino et al. proposed a method that enables transcription of polyphonic music even if there are several kinds of instruments included [1]. Goto presented a method for extracting objective single sound from polyphonic musical signals without restriction on the number of simultaneous sounds [2]. Although both methods are superior for their respective purposes, they are not suited for separating every individual sound.

Meanwhile, numerous methods of MPE aiming especially at sound separation have been reported, mainly in musical signal processing [3], speech signal processing [4] and auditory scene analysis [5, 6]. Chazan addressed a speech separation method by introducing a time warped signal model that allows a continuous pitch variation within a long analysis frame [7]. The robustness of this method is limited to the continuous pitch variation, and it can not cope with discontinuous pitch variation such as in piano performance. Klapuri described a robust MPE method [8] and Virtanen constructed a sound separation system by sophisticating it [9]. While this system is prominent in respect that it enables extraction not only of the amplitudes of the partials but even of the phases by parameter estimation of time domain signal model. However, it requires a number of stages such as bandwise processing, partial removing, spectral smoothing, and sinusoidal modeling analysis.

Our objective is to develop a method that estimates the number of underlying harmonic structures and separates them without restriction on the number of concurrent sounds and a variety of sound sources, and is also able to extract pitches accurately as continuous values, with much more simple procedure in the spectral domain.

2. A MAXIMUM LIKELIHOOD FORMULATION

2.1. Model of Harmonic Structures

The use of window function and the varying pitch within a short time single analysis frame inevitably cause widening of the spectral harmonics which makes it difficult to extract the precise value of fundamental frequencies (F_0s) and to separate close partials. First, we assume that each widened partial is a probability distribution of frequencies, approximated by a Gaussian distribution model. Therefore, a single harmonic structure can then be modeled by a tied Gassian mixture model (tied-GMM), in which their means have only 1 degree of freedom. In log-frequency scale, means of tied-GMM are denoted here as $\boldsymbol{\mu}_k = \{\mu_k, \cdots, \mu_k + \log n, \cdots, \mu_k + \log N_k\}$ where μ_k ideally corresponds to the F_0 of kth sound and n denotes the index of partials. We then introduce a model of multiple harmonic structures $P_\theta(x)$ which is a mixture of K tied-GMMs whose model parameter $\boldsymbol{\theta}$ is denoted as

$$\{\theta\} = \{\boldsymbol{\mu}_k, \boldsymbol{w}_k, \sigma \mid k = 1, \cdots, K\}, \tag{1}$$

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

2.2. Model Parameter Estimation using EM Algorithm

Since the observed spectral density function $f(x)$, where x denotes log-frequency, is considered to be 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

$$f(x)\log P_{\bar{\theta}}(x) - f(x)\log P_{\theta}(x) = f(x)\log\frac{P_{\bar{\theta}}(x)}{P_{\theta}(x)}. \quad (2)$$

Although Dempster formulated EM algorithm [10] in order to maximize the mean log-likelihood considering $f(x)$ as a probabilistic density function, it can also be formulated in the same way even if $f(x)$ is replaced with spectral density function. By taking expectation of both sides with respect to $P_\theta(n,k|x)$, representing the probability of $\{n,k\}$-labeled Gaussian distribution from which x is generated, Q function is derived in the right-hand side. Q function is by

$$Q(\theta,\bar{\theta}) = \sum_{k=1}^{K}\sum_{n}^{N_k}\int_{-\infty}^{\infty} P_\theta(n,k|x) f(x)\log P_{\bar{\theta}}(x,n,k)dx, \quad (3)$$

thus it yields

$$\int_{-\infty}^{\infty}\left\{ f(x)\log P_{\bar{\theta}}(x) - f(x)\log P_{\theta}(x)\right\}dx \geq Q(\theta,\bar{\theta}) - Q(\theta,\theta). \quad (4)$$

By obtaining $\bar{\theta}$ that maximizes the Q function, the log-likelihood of the model of multiple harmonic structures with respect to every x will monotonically increase. A posteriori probability $P_\theta(n,k|x)$ in equation (3) is given as

$$P_\theta(n,k|x) = \frac{P_\theta(x,n,k)}{P_\theta(x)}, \quad (5)$$

$$= \frac{w_n^k \cdot g(x|\mu_k + \log n, \sigma^2)}{\displaystyle\sum_n\sum_k w_n^k \cdot g(x|\mu_k + \log n, \sigma^2)}, \quad (6)$$

$$g(x|x_0,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{(x-x_0)^2}{2\sigma^2}\right\}, \quad (7)$$

where $g(x|x_0,\sigma^2)$ is a Gaussian distribution. By iterating in two steps as follows, the model parameter $\boldsymbol{\theta}$ locally converges to ML estimates.

Initial-step

Initialize the model parameter $\boldsymbol{\theta}$.

Expectaion-step

Calculate $Q(\theta,\bar{\theta})$ with equation (3).

Maximization-step

Maximize $Q(\theta,\bar{\theta})$ to obtain the next estimate

$$\boldsymbol{\theta} = \underset{\bar{\theta}}{\operatorname{argmax}}\, Q(\theta,\bar{\theta}). \quad (8)$$

Replace $\bar{\boldsymbol{\theta}}$ with $\boldsymbol{\theta}$ and repeat from the Expectation-step.

2.3. Physical Interpretation as Clustering

From another viewpoint, this ML procedure can be understood as a clustering method under a harmonic constraint between Gaussian mixture components where spectral density function is considered as a statistical distribution of micro-energies along frequency axis. As we regard μ_k as cluster centroids, the posteriori probability in equation (6) as a membership degree of each micro-energy and the log-likelihood $P_{\bar{\theta}}(x,n,k)$ as a distance function between centroid μ_k and a micro-energy, thus the Q function in equation (3) turns out to be the objective function for fuzzy clustering. We call this concept "Harmonic Clustering."

3. SEPARATION OF HARMONIC STRUCTURES

The separation scheme as a whole consists of only 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 selected. Here we introduce Akaike Information Criterion (AIC) proposed by Akaike in 1973 [11]. AIC is given by

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

and is known to offers proper estimate of the number of free parameters.

3.2. Estimation of the number of harmonic structures

It is generally known that ML estimates depend highly on initial values and may often converge to undesirable values. To avoid this, we first prepare extra amount of tied-GMMs in the model in order to raise possibility of obtaining the true values. Then, obviously, the model may over-fit 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 tied-GMMs. And this number can be estimated by reducing tied-GMM one after another until AIC takes minimum value. 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

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

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

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

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

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

3. Calculate AIC with equation (9). Since there are two free parameters for each tied-GMM, the model has $2 \times K$ free parameters altogether. If the AIC increases, the number of tied-GMMs just before they are reduced in step4 will be the estimate of the number of harmonic structures.

4. Remove tied-GMM(s) that conforms to either of the two conditions given 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 works is shown in Fig.1 and the input spectrum used in it is depicted in Fig.2. The broken line represents the point where the model parameters were judged to be converged and the line graph indicates the value of AIC. Since AIC takes minimum value when three tied GMMs are left, the estimate number here is 3.

3.3. Estimation of F_0 and Spectral Shape

In the previous process, the ML procedure allows to aqcuire local optimal solutions of μ_k without distinction of true F_0s, harmonics or subharmonics. Therefore, true F_0 can be estimated by replacing μ_k sequentially to its harmonics and subharmonics. Consider now that a degree of freedom is given to every w_n^k (except one) and allows to extract spectral shape, i.e., the relative amplitudes of partials. If μ_k corresponds to a subharmonic of true F_0, the model must overfit the given spectrum. From this point of view, searching true F_0s and extraction of the spectral shape can also be handled with information criterion. The process shown below is done with all remaining tied-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.

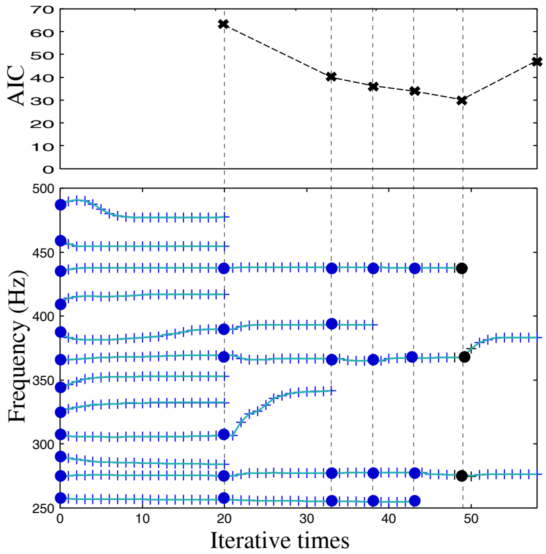

Fig. 1. Example of convergence to the true values

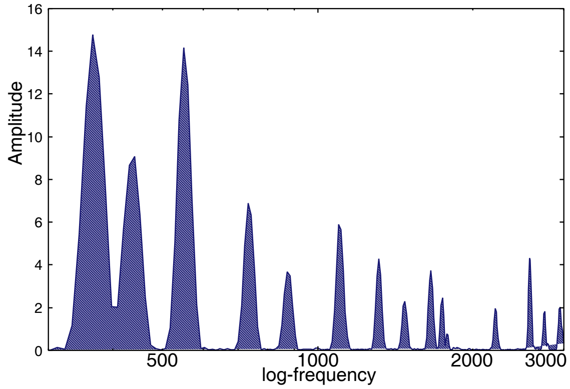

Fig. 2. Input spectrum for Fig. 1

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 with equation (9). The number of free parameters here is N_k^t. When AIC takes minimum value, $\mu_k + \log(n-1)$ is the F_0 estimate, and if not, add 1 to t and return to step1.

4. EXPERIMENTS

Experiments were carried out to evaluate the performance of our algorithm against mixed sounds, both continuous and discontinuous pitch variations by determining the accuracy of F_0 detection.

4.1. Results for Simultaneous Speech

The algorithm was first tested on co-channel simultaneous speech signals spoken by two speakers, from which continuous pitch variations were expected. Each speech signal file

Table 1. Results for simultaneous speech

Speech signals		Accuracy(%)	
Speaker 1	Speaker 2	Speaker 1	Speaker 2
‘myi’	‘myi’	90.1	83.0
‘myi’	‘fym’	74.8	92.8
‘fym’	‘fym’	86.2	92.6

Table 2. Results for polyphonic music

Expermiental data		Accuracy(%)
Composer & Title	Instrument	
J. Pahelbel: “Kanon”	Violin	92.7
J. S. Bach: “Ricercare à 6”	Violin	87.7
J. S. Bach: “BWV 1046 no.1, mov.4”	Oboe	89.2
J. S. Bach: “Menuet”	Piano	84.2

was artificially created by mixing two independent speech signals of the ATR Speech Database at 0 dB signal-to-signal ratio. All signals were digitized at 12 kHz sampling rate and analyzed with Hamming window where frame length and shift were 64 ms and 10 ms, respectively. Hand-labeled F_0 contours, also included in the database, were used as references. The accuracy rates for the respective speakers are shown in table 1. Label ‘myi’ and ‘fym’ stand for a male and a female speaker. Deviations over 5% from the references were deemed as gross errors. The initial number of the tied GMMs was set to 3 and the frequency range was from 70 Hz to 140 Hz, and σ was assigned to 0.45. As a result, as for the concurrent speech with male and female speakers, the accuracy for male speaker was relatively low. At the second process stated in Section 3, AIC rather prefers μ_k to be positioned at as higher frequency as it can, since the number of free parameters can be decreased. Accordingly, if both the pitch and the amplitude of one utterence were specificially lower than another, it was disregarded. Some other gross errors were found at the first process mainly due to uttered consonants. Since we focused only on harmonic structure, gross errors caused by ‘noise’, including unvoiced consonants, were difficult to avoid.

4.2. Results for Recorded Polyphonic Music

The algorithm was next tested on 4 pieces of music in the RWC Music Database and CD recordings. The musical signals were sampled at 44.1 kHz and analyzed with Hamming window where frame length and frame shift were 50 ms and 10 ms, respectively. Reference F_0s were hand-labeled according to the notes and the durations transcribed in the musical score. The accuracy rates for the music pieces are shown in table 2. The initial number of tied GMMs was set to 5, the frequency range was from 108 Hz to 215 Hz, and σ was assigned to 0.53. The results show the algorithm worked well with the violin performance. As for the piano performance, though fast decay of piano sound made detection difficult, F_0s before its decay were extracted properly.

5. CONCLUSIONS

We proposed an algorithm that enables estimation of the number of underlying harmonic structures and multiple F_0s and separation of mixed harmonic structures with spectral domain procedure. It showed high performance for both simultaneous speech and polyphonic music. Still, improvement is expected 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] S. Godsill and M. Davy, “Baysian Harmonic Models for Musical Pitch Estimation and Analysis,” *Proc. ICASSP2002*, Vol. 2, pp. 1769–1772, 2002.

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

[5] 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).

[6] 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).

[7] 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.

[8] 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.

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

[10] 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.

[11] 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.