Multi-pitch Detection Algorithm Using Constrained Gaussian Mixture Model and Information Criterion for Simultaneous Speech

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 co-channel multi-pitch detection algorithm is described. We suggest the importance of this when prosodic information is need to be extracted separately from respective F_0 patterns of concurrent utterances. Though temporal continuity of speech prosody should be considered, we discuss a process done independently on each single frame as the first step. A model of multiple harmonic structures is constructed with a mixture of tied Gaussian mixtures with which a single harmonic structure is modeled. Our algorithm enables to detect both a number of concurrent speakers, and each spectral envelope of underlying harmonic structure based on a maximum likelihood estimation of the model parameters using EM algorithm and an information criterion. It operates without a priori information of F_0 contours and a restriction of a number of speakers, and it also extracts accurate F_0s as continuous values with simple procedures in spectral domain. Experiments showed our algorithm outperformed well-known cepstrum for both speech signals of a single speaker and simultaneous two speakers.

1. Introduction

It is known that prosodic information offers many useful clues for speech recognition, such as location of important words and phrases, topic segment boundaries, location of disfluencies, identification of languages and others. The process of extracting prosodic information is generally conducted on the assumption that F_0 pattern is already (roughly) extracted. Yet F_0 patterns can not always be extracted simply in spontaneous dialogue speech in which simultaneous utterances by two or more speakers often occur. Thus, in order to incorporate proper prosodic information into spontaneous dialogue speech recognition, a number of simultaneous speakers and respective F_0 patterns are desired to be extracted precisely. However, the multi-pitch detection problem is hardly simple and is difficult to be solved analytically.

Until now, numerous multi-pitch detection methods have been reported not only in speech signal processing [1, 2] but also in musical signal processing[3, 4, 5] and auditory scene analysis [6, 7]. Chazan et al. addressed a speech separation method by introducing a time warped signal model which allows a continuous pitch variations within a long analysis frame [1]. Wu et al. described a multi-pitch tracking method in noisy environment by filter bank process and pitch tracking using HMM [2]. Although these methods actualize an accurate detection of F_0s, either of them does not include specific process of determining the number of speakers.

Our objective is to develop a multi-pitch detection algorithm which enables to detect the number of simultaneous speakers, the accurate F_0s as a continuous values, and moreover, respective spectral envelopes with spectral domain procedure. The basic approach is stated in Section 2, and the detection algorithm is described in Section 3. And the results of operation experiments are reported in Section 4.

2. A Maximum Likelihood Formulation

2.1. Model of Harmonic Structures

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 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 $\log F_0$ 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 variances (which are 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

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

Although Dempster formulated EM algorithm [8] in order to maximize the mean log-likelihood considering $f(x)$ as a probabilistic density function, it can also be formulated in a same way even if $f(x)$ is replaced with spectral density function. By taking expectation of both sides with respect to $P_\theta(n, k|x)$ which represents the probability of the $\{n, k\}$-labeled Gaussian distribution from which x is generated, Q-function will be derived in the right-hand side. Given Q-function as

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

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

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

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

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

where $g(x|x_0,\sigma^2)$ is a Gaussian distribution. By the iterative procedure of the 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. Another 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 a 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. Multi-pitch Detection Algorithm

The detection algorithm 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 [9]. AIC is given by

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

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

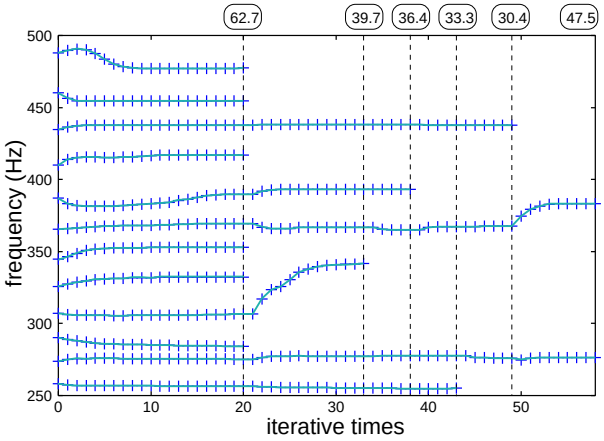

Figure 1: *An example of convergence to the true values*

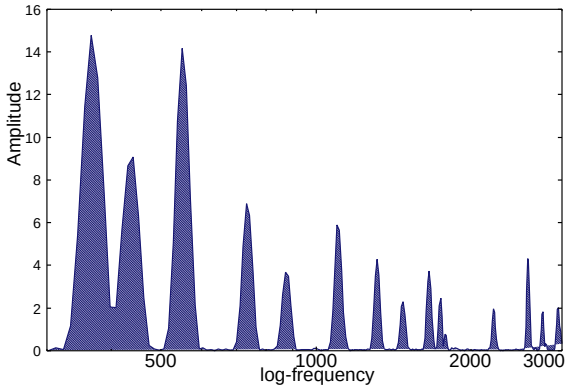

Figure 2: *Input spectrum for Figure 1*

3.2. Detection of the number of speakers

It is generally known that ML estimates obtained by EM algorithm firmly depend 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 the tied-GMMs. And this number can be detected by reducing tied-GMM one after another until they become the proper number on the basis of AIC. 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 to

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

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

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 the tied-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 works is shown in Fig.1 where the observed spectrum used is depicted in Fig.2. The broken line represents the point where the model parameters were judged to be converged and the circled value indicates the value of AIC at each point. Since AIC takes minimum when 3 tied-GMMs remain, the detected number here is 3.

3.3. Detection of F_0s and Spectral Envelopes

In the previous process, the ML procedure allows to aqcuire local optimal solutions of μ_k without distinction of the true F_0s or the multiples of the true F_0s. 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 is lower than the true F_0, the model must be over-fit. 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 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.
2. Estimate the ML model parameters by EM algorithm. Here we only update w_n^k and should be updated to

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

3. Calculate AIC with equation (9). The number of free parameters here is N_k^t. If the AIC increases, the process should be interrupted and the $\mu_k + \log(n-1)$ is considered as the detected F_0, and if not, add 1 to t and return to step1.

4. Experiments

Experiments were carried out to validate our algorithm by evaluating the accuracy of F_0 detection in comparison with well-known cepstrum. A database of every speech file and reference F_0 contour are constructed from the ATR Speech Database. 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. The initial number of the tied-GMMs was set to 4 and the frequency range was from 70 Hz to 140 Hz, and σ was assigned to 0.45. Speech files begin with 'myi-' and 'fym-' stand for speech signals of a male and a female speakers. Deviations over 5% from the references were deemed as gross errors. Every accuracy shown in table 1, 2 and 3 is a percentage of frames at which F_0s are correctly detected.

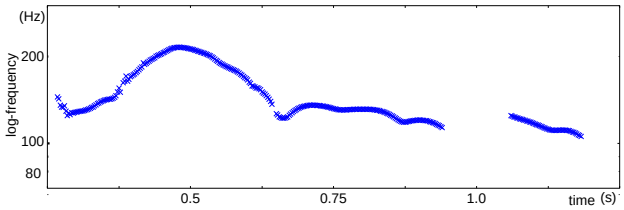

Figure 3: *Detected F_0 contour of a single speaker*

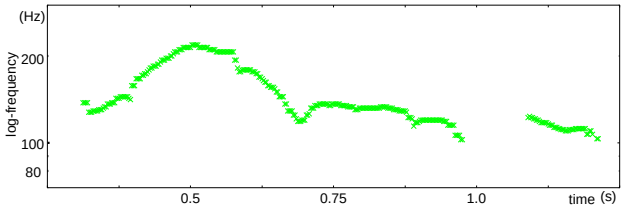

Figure 4: *Reference F_0 contour corresponding to Figure 3*

Table 1: *Results for a single speaker*

Speech file	Accuracy(%)	
	Cepstrum	Proposed
'myisda01'	88.2	98.0
'myisda02'	88.4	99.0
'myisda03'	84.8	98.1
'myisda04'	85.1	92.4
'myisda05'	76.8	93.7
'fymsda01'	86.3	98.5
'fymsda02'	87.1	97.5
'fymsda03'	83.3	95.8
'fymsda04'	86.7	96.8
'fymsda05'	85.2	96.0

4.1. Results for Speech Signals of a single speaker

The algorithm was first tested on single-channel speech signals of a single speaker. A comparison of accuracies between cepstrum and proposed method for each speaker are shown in table 1. As the results, our algorithm significantly outperforms cepstrum. An example of detected F_0 contour is depicted in figure 3 where the reference is shown in figure 4.

4.2. Results for Simultaneous Speech Signals

The algorithm was next tested on co-channel simultaneous speech signals spoken by two speakers. Each speech signal file was artificially created by mixing two independent speech signals with 0 dB signal-to-signal ratio. To evaluate our algorithm objectively, we also applied cepstrum for simultaneous speech signals which is not generally designed as a multi-pitch detector. Results with cepstrum are shown in table 2 and results with our algorithm are shown in table 3. An example of detected F_0 contours is depicted in figure 5 where the reference is shown in figure 6. Pairs of speech files by which concurrent speech signals are created are shown in the first and second columns in table 2 and 3. As the results, our algorithm significantly outperformed cepstrum as well and showed high performance.

Some of the gross errors were found at the first process

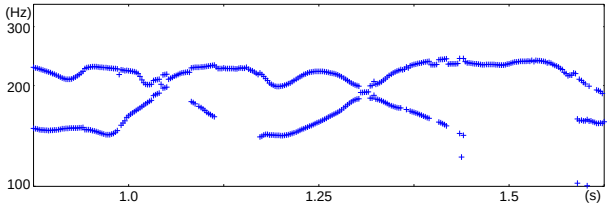

Figure 5: *Detected F_0 contours of two concurrent speakers*

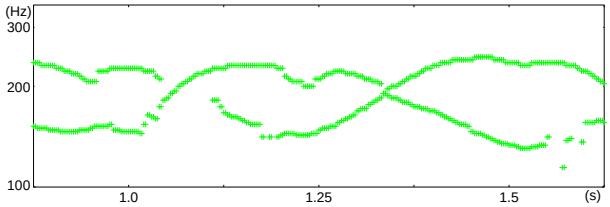

Figure 6: *Reference F_0 contours corresponding to Figure 5*

Table 2: *Results for two speakers (Cepstrum)*

Speech files		Accuracy(%)	
File 1	File 2	Speaker 1	Speaker 2
'myisda01'	'myisda03'	63.7	63.1
'myisda01'	'myisda04'	45.7	51.6
'myisda02'	'myisda03'	63.3	50.1
'myisda02'	'myisda04'	59.4	42.1
'fymsda01'	'fymsda02'	57.7	54.0
'fymsda01'	'fymsda04'	53.1	41.0
'fymsda02'	'fymsda03'	52.9	59.6
'fymsda02'	'fymsda04'	64.9	64.7
'myisda01'	'fymsda03'	45.7	43.0
'myisda02'	'fymsda05'	55.0	44.5
'myisda03'	'fymsda04'	41.4	59.9
'myisda04'	'fymsda02'	64.9	50.6
'myisda05'	'fymsda03'	59.4	62.8
'myisda04'	'fymsda01'	62.0	71.7

Table 3: *Results for two speakers (Proposed)*

Speech files		Accuracy(%)	
File 1	File 2	Speaker 1	Speaker 2
'myisda01'	'myisda03'	90.1	83.0
'myisda01'	'myisda04'	92.8	81.3
'myisda02'	'myisda03'	88.2	85.7
'myisda02'	'myisda04'	84.4	87.6
'fymsda01'	'fymsda02'	90.7	84.3
'fymsda01'	'fymsda04'	85.3	82.6
'fymsda02'	'fymsda03'	79.2	90.3
'fymsda02'	'fymsda04'	86.2	92.6
'myisda01'	'fymsda03'	76.1	84.9
'myisda02'	'fymsda05'	74.8	92.8
'myisda03'	'fymsda04'	72.6	88.4
'myisda04'	'fymsda02'	86.3	85.5
'myisda05'	'fymsda03'	78.0	86.6
'myisda04'	'fymsda01'	79.0	86.6

mainly because of unvoiced consonants. Since we focused only on harmonic structure, the gross errors caused by them were difficult to avoid. Meanwhile, when the two simultaneous speakers were male and female, male rather resulted worse. At the second process stated in Section 3, AIC rather prefers μ_k to be positioned in as higher frequency as it can because the number of free parameters can be lessen. Accordingly, if both pitch and amplitude of one utterence was specificially lower than another, it tended to be ignored.

5. Conclusions

We proposed an algorithm which enables to detect the number of speakers, accurate F_0s and spectral envelopes from co-channel input simultaneous speech signals with spectral domain procedure. It showed a high performance for speech signals of both single speaker and two speakers. Still, several improvements are prospective by considering temporal continuity of F_0 contour (e.g., introducing Fujisaki model), 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] Chazan, D.; Stettiner, Y.; Malah, D., 1993. Optimal Multi-pitch Estimation Using the EM Algorithm for Co-channel Speech Separation. *Proc. ICASSP93*, Vol. 2, 728–731.

[2] Wu, M.; Wang, D.; Brown, G. J., 2002. A Multi-pitch Tracking Algorithm for Noisy Speech. *ICASSP2002*, Vol. 1, 369–372.

[3] Godsill, S.; Davy, M., 2002. Baysian Harmonic Models for Musical Pitch Estimation and Analysis. *Proc. ICASSP2002*, Vol. 2, 1769–1772.

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

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

[6] Abe, M.; Ando, S., 2000. 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, 468–477, (in Japanese).

[7] Karjalainen, M.; Tolonen, T., 2001. Multi-pitch and Periodicity Analysis Model for Sound Separation and Auditory Scene Analysis. *IEEE Trans. on Speech and Audio Processing*, vol. 9, no. 2, 127–140.

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

[9] Akaike, H., 1973. Information Theory and an Extension of the Maximum Likelihood Principle. *2nd Inter. Symp. on Information Theory*, 267–281.