Multi-Pitch Trajectory Estimation of Concurrent Speech Based on Harmonic GMM and Nonlinear Kalman Filtering

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

This paper describes a multi-pitch tracking algorithm of 1-channel simultaneous multiple speech. The algorithm selectively carries out the two alternative processes at each frame: frame-independent-process and frame-dependent-process. The former is the one we have previously proposed[6], that gives good estimates of the number of speakers and F_0s with a single-frame-processing. The latter corresponds to the topic mainly described in this paper, that recursively tracks F_0s using nonlinear Kalman filtering. We tested our algorithm on simultaneous speech signal data and showed higher performance than when the frame-independent-process was only used.

1. Introduction

1-channel multi-pitch estimation technique may contribute to various applications, such as spontaneous dialogue speech recognition, that allows competitive speech, noise robust speech recognition, especially where noise is a harmonic signal(e.g., telephone ring, back ground music, etc.), and also many music applications. However, multi-pitch estimation of non-stationary signals is hardly simple due to the complex factors such as spectral overlap, poor frequency resolution and spectral widening in short-time analysis, etc. Various approaches concerning to this problem have conventionally been attempted [1, 2, 3], while two important tasks have been left unsolved. Firstly, there has been no robust way of estimating the number of speakers, and most of the methods were obliged to assume for simplicity that the number is known *a priori*. Secondly, the double/half(harmonics/subharmonics) pitch error has still been one of the most critical problem where convincing solutions are not yet proposed. One may say both problems share the same difficulty of defining physically or mathematically proper criteria. Until now, we have proposed a GMM(Gaussian mixture model)-based multi-pitch estimation algorithm that works as a single-frame-processing and gives solutions to the two tasks stated above according to the information criterion[6]. This algorithm has not yet taken into account any time dependency property, that would ensure the improvements in performance. In this paper we incorporate a procedure based on nonlinear Kalman filtering, that effectively reflects the local time dependency property of F_0 pattern.

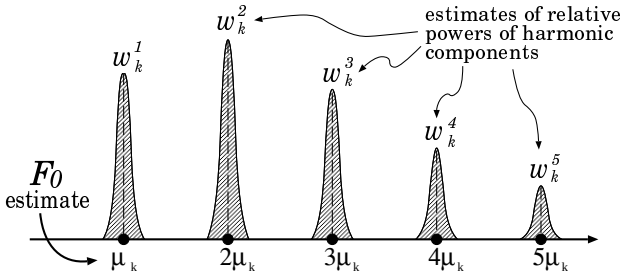

Figure 1: *Model parameters of the Harmonic GMM*

2. Formulation of the Harmonic GMM

2.1. Gaussian Modeling of Multi-Pitch Spectrum

Our approach is fundamentally based on modeling of the observed spectrum by a conditional function model, that characterizes the "harmonicity" of periodic sounds. For instance, a single harmonic structure is modeled by a tied Gaussian mixture model (harmonic GMM), of which their means are constrained by the F_0 estimate parameter (see Fig1). The means of the kth harmonic GMM are denoted as $\boldsymbol{\mu}_k = \{\mu_k, \cdots, n\mu_k, \cdots, N_k\mu_k\}$ where μ_k ideally corresponds to the F_0 of kth sound and n and N_k are the index and the number of Gaussians (partials). We can model multi-pitch spectrum by a mixture of multiple harmonic GMMs (multi-pitch spectral model). Let the multi-pitch spectral model be a probabilistic density function $p(\omega|\theta)$ whose model parameter $\boldsymbol{\theta}$ is

$$\{\theta\} = \{\mu_k, \boldsymbol{w}_k, \boldsymbol{\sigma}_k \mid k \in \mathbf{N}, k=1, \cdots, K\}, \qquad (1)$$

where $\boldsymbol{w}_k$ and $\boldsymbol{\sigma}_k$ indicate the weight and variance parameters of the respective Gaussians, that represent the relative powers of frequency components and the degrees of spectral widening.

2.2. Model Parameter Estimation

By considering the (normalized) observed spectrum $y(\omega)$ being a statistical distribution of imaginary micro-energy particle, minimization of Kullback-Leibler (KL) information between $y(\omega)$ and the multi-pitch spectral model

$p(\omega|\theta)$ is mathematically equal to maximum likelihood (ML). The mean log-likelihood function is given as

$$\log p(y|\theta) = \int_{-\infty}^{\infty} f(\omega) \log p(\omega|\theta) d\omega, \quad (2)$$

and the local maximization can effectively be operated using EM algorithm.

3. Multi-Pitch Trajectory Estimation Using Nonlinear Kalman Filtering

We apply a recursive solution to the multi-pitch temporal trajectory estimation within a certain length of time segment, in which the number of speakers is already known or estimated and also assumed to be constant.

3.1. Nonlinear Kalman Filtering

Nonlinear Kalman filtering[5], refered to here, is a sub-optimal solution to the state estimation of a discrete-time controlled process that has a state vector $\boldsymbol{\Theta} \in \mathbb{R}^K$ and governed by the nonlinear stochastic differential equation,

$$\boldsymbol{\Theta}_{i+1} = f(\boldsymbol{\Theta}_i, \boldsymbol{\xi}_i) \quad (3)$$

$$\boldsymbol{Y}_{i+1} = h(\boldsymbol{\Theta}_{i+1}, \boldsymbol{v}_{i+1}), \quad (4)$$

where the random variables $\boldsymbol{\xi}_i$ and $\boldsymbol{v}_i$ represent the process and observation noise. The nonlinear function f in the differential equation relates the state at the current time step i to the state at the future time step $i+1$, and h relates the state $\boldsymbol{\Theta}_i$ to the observation $\boldsymbol{Y}_i$.

3.2. Stochastic Modeling of Multi-Pitch Trajectory

Suppose F_0 trajectory in a certain segment of speech phrase can be modeled as a state process that follows the differential equation (3), F_0 can be recursively estimated along time steps by using nonlinear Kalman filtering principle. To begin with, we define a state $\boldsymbol{\Theta}_i$ as a vector of F_0 estimates and their derivatives,

$$\boldsymbol{\Theta}_i = \begin{pmatrix} \boldsymbol{\mu}_i \\ \dot{\boldsymbol{\mu}}_i \end{pmatrix} = \begin{pmatrix} \mu_1^{(i)} & \cdots & \mu_k^{(i)} & \cdots & \mu_K^{(i)} \\ \dot{\mu}_1^{(i)} & \cdots & \dot{\mu}_k^{(i)} & \cdots & \dot{\mu}_K^{(i)} \end{pmatrix} \quad (5)$$

where the number of speakers is K. Note that the weight parameters should also be included in the state vector to make the estiamtion more robust, for simplicity, we focus only on the behaviour of F_0s. Each element of the observation vector $\boldsymbol{Y}_i = \{y_i, y_{i-1}, y_{i-2} \cdots\}$ signifies an observed spectrum distribution at each time step.

3.3. Formulation

A linearised approximation of equation (3) together with statistical estimation approach offers a reliable solution to the nonlinear problem. Now we have a new differential equation notation as below.

$$\boldsymbol{\Theta}_{i+1} = \boldsymbol{A}\boldsymbol{\Theta}_i + \boldsymbol{B}\xi_i \quad (6)$$

Let $\boldsymbol{A}$ be a first order dynamics, given as

$$\boldsymbol{A} = \begin{pmatrix} 1 & \Delta t \\ 0 & 1 \end{pmatrix}, \quad (7)$$

and $\boldsymbol{B}\xi_i$ be a zero-mean Gaussian noise, whose variances corresponding to $\boldsymbol{\mu}_i$ and $\dot{\boldsymbol{\mu}}_i$ are ν^2 and η^2. The observation equation (4) can be interpreted, in a probabilistic sense, as a likelihood function $p(y_{i+1}|\boldsymbol{\Theta}_{i+1})$, that is obtained by a given spectrum with the multi-pitch spectral model (harmonic GMM mixture) as stated in section 2.

The *a posteriori* density function $p(\boldsymbol{\Theta}_{i+1}|\boldsymbol{Y}_{i+1})$ can be estimated by $p(\boldsymbol{\Theta}_i|\boldsymbol{Y}_i)$, via two update equations: *time update ("predictor")* equations and *observation update ("corrector")* equations.

3.3.1. Time update step

In this step, the *a posteriori* density function $p(\boldsymbol{\Theta}_i|\boldsymbol{Y}_i)$ at the current time step is modified into the *a priori* density function $p(\boldsymbol{\Theta}_{i+1}|\boldsymbol{Y}_i)$ at the future time step $i+1$.

If the state sequence $\boldsymbol{\Theta}_0, \cdots, \boldsymbol{\Theta}_i$ follows a first order Markov process, we obtain the Chapman-Kolmogorov equation, which is given by

$$p(\boldsymbol{\Theta}_{i+1}|\boldsymbol{Y}_i) = \int p(\boldsymbol{\Theta}_{i+1}|\boldsymbol{\Theta}_i) p(\boldsymbol{\Theta}_i|\boldsymbol{Y}_i) d\boldsymbol{\Theta}_i. \quad (8)$$

F_0 and its derivative of the kth speaker $\boldsymbol{\mu}_k^{(i)} = (\mu_k^{(i)}, \dot{\mu}_k^{(i)})^T$ is usually independent of those of other speakers $\boldsymbol{\mu}_{k'}^{(i)} = (\mu_{k'}^{(i)}, \dot{\mu}_{k'}^{(i)})^T$ $(k' \neq k)$, so that $p(\boldsymbol{\Theta}_{i+1}|\boldsymbol{\Theta}_i)$ is expressed as a joint distribution of every $p(\boldsymbol{\mu}_k^{(i+1)}|\boldsymbol{\mu}_k^{(i)})$ with respect to k. The transition density function $p(\boldsymbol{\mu}_k^{(i+1)}|\boldsymbol{\mu}_k^{(i)})$ can be further separated into

$$p(\boldsymbol{\mu}_k^{(i+1)}|\boldsymbol{\mu}_k^{(i)}) = p(\mu_k^{(i+1)}|\mu_k^{(i)}, \dot{\mu}_k^{(i)}) p(\dot{\mu}_k^{(i+1)}|\dot{\mu}_k^{(i)}) \quad (9)$$

where each term is explicitly given as

$$p(\mu_k^{(i+1)}|\mu_k^{(i)}, \dot{\mu}_k^{(i)}) = \frac{1}{\sqrt{2\pi\nu^2}} e^{-\frac{(\mu_k^{(i+1)} - \mu_k^{(i)} - \dot{\mu}_k^{(i)} \Delta t)^2}{2\nu^2}} \quad (10)$$

$$p(\dot{\mu}_k^{(i+1)}|\dot{\mu}_k^{(i)}) = \frac{1}{\sqrt{2\pi\eta^2}} e^{-\frac{(\mu_k^{(i+1)} - \mu_k^{(i)})}{2\eta^2}} \quad (11)$$

from the equation (6) and (7).

We can simply reduce the computation of equation (8) , that requires 2-dimensional convolution operation, without losing its essence by the decomposition

$$\begin{aligned} p(\boldsymbol{\Theta}_{i+1}|\boldsymbol{Y}_i) &= p(\boldsymbol{\mu}_{i+1}|\boldsymbol{Y}_i) p(\dot{\boldsymbol{\mu}}_{i+1}|\boldsymbol{\mu}_{i+1}, \boldsymbol{Y}_i) \\ &= \prod_{k=1}^{K} p(\mu_k^{(i+1)}|\boldsymbol{Y}_i) p(\dot{\mu}_k^{(i+1)}|\mu_k^{(i+1)}, \boldsymbol{Y}_i) \end{aligned} \quad (12)$$

and the independent computation of each element. The state prediction of the next time step $p(\boldsymbol{\Theta}_{i+1}|\boldsymbol{Y}_i)$, i.e., the *a priori* density at the future time step $i+1$, can consequently be computed.

3.3.2. Observation update step

In this step, the *a priori* density function $p(\boldsymbol{\Theta}_{i+1}|\boldsymbol{Y}_i)$ is modified into the *a posteriori* density function $p(\boldsymbol{\Theta}_{i+1}|\boldsymbol{Y}_{i+1})$ at the future time step $i+1$ according to the appearance of a new observation y_{i+1}. $p(\boldsymbol{\Theta}_{i+1}|\boldsymbol{Y}_{i+1})$ can be derived from Bayes' theorem:

$$p(\boldsymbol{\Theta}_{i+1}|\boldsymbol{Y}_{i+1}) = \frac{p(y_{i+1}|\boldsymbol{\Theta}_{i+1}, \boldsymbol{Y}_i)p(\boldsymbol{\Theta}_{i+1}|\boldsymbol{Y}_i)}{\int p(y_{i+1}|\boldsymbol{\Theta}_{i+1}, \boldsymbol{Y}_i)p(\boldsymbol{\Theta}_{i+1}|\boldsymbol{Y}_i)d\boldsymbol{\Theta}_{i+1}} \quad (13)$$

F_0s of K speakers in the future time step $i+1$ can be estimated by Maximum *a posteriori* (MAP) estimation:

$$\bar{\boldsymbol{\Theta}}_{i+1} = \underset{\boldsymbol{\Theta}_{i+1}}{\operatorname{argmax}}\, p(y_{i+1}|\boldsymbol{\Theta}_{i+1}, \boldsymbol{Y}_i)p(\boldsymbol{\Theta}_{i+1}|\boldsymbol{Y}_i), \quad (14)$$

where $p(y_{i+1}|\boldsymbol{\Theta}_{i+1}, \boldsymbol{Y}_i)$ is the likelihood function of the multi-pitch spectral model given observed spectrum y_{i+1}.

The recursive procedure of updating $p(\boldsymbol{\Theta}_i|\boldsymbol{Y}_i)$ to $p(\boldsymbol{\Theta}_{i+1}|\boldsymbol{Y}_{i+1})$, via two steps described above, suboptimally estimates the overall F_0 trajectories of a certain segment of multiple concurrent speech.

4. The Proposed Algorithm

The algorithm broadly consists of the frame-independent-process[6] (Akaike Information Criterion[4] based approach), and the frame-dependent-process (the Kalman filtering based approach stated in section 3).

4.1. AIC Based Frame-Independent-Process

At a discontinuous part of F_0 trajectory or a time step where a competetive speaker appears/disappears (a phrase boundary or a speech onset/offset), F_0s and the number of speakers should be estimated without information of previous time step. Let us give a brief review of our previous work, that provides good estimation of the number of concurrent sounds and F_0s with a single frame processing[6] through 2-phase procedure: estimation of *the number of speakers* and estimation of F_0 *positions*.

4.1.1. Phase 1: Estimation of the number of speakers

An underfit spectral model that has smaller number of harmonic GMMs than the actual number of concurrent sounds(speakers), obviously gives extremely small maximum likelihood in comparison with the 'bestfit' model that has exactly the same number of harmonic GMMs as the actual number. On the other hand, an overfit model, that has larger number of harmonic GMMs than the actual number, gives somewhat larger but almost the same maximum likelihood, despite of the advantage in terms of the number of free parameters(harmonic GMMs) (see Fig 2). We can estimate the number of speakers by the iteration; *AIC calculation after the EM algorithm convergence* and *deletion of the least-weighted harmonic GMM*, that begins with a large number of harmonic GMMs and stops when AIC takes local minimum value.

4.1.2. Phase 2: Estimation of the proper F_0 positions

Since not only the true F_0 but also its harmonics/ subharmonics are the local optimal solution of μ_k, the F_0 estimate via the previous phase $\hat{\mu}_k$ is not guaranteed to be the true F_0. Let us now estimate the optimal weight parameters (spectral envelope), where $\hat{\mu}_k$ is fixed, to confirm whether or not $\hat{\mu}_k$ has been correctly estimated. If $\hat{\mu}_k$ has been harmonics/subharmonics, the kth harmonic GMM

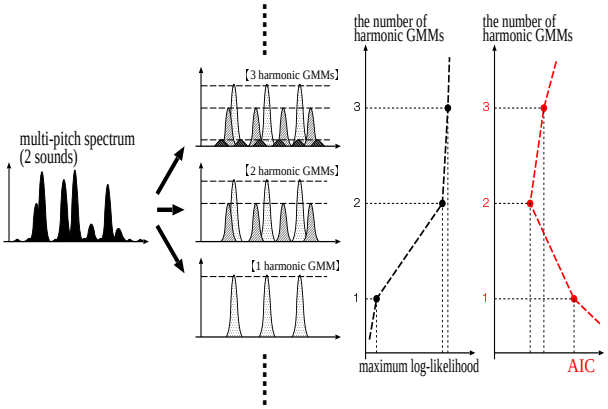

Figure 2: *Comparison of maximum likelihood and AIC with overfit, 'bestfit' and underfit models (phase 1).*

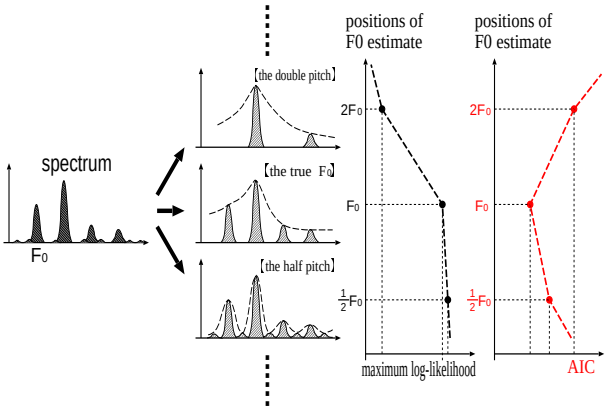

Figure 3: *Comparison of maximum likelihood and AIC with overfit, 'bestfit' and underfit models (phase 2).*

appears to be an underfit/overfit model (see Fig 3). Thus we can estimate the true F_0 position by the iteration: *AIC calculation after the EM algorithm convergence* and *replacing $\hat{\mu}_k$ to its multiples/divisors*, and stops when AIC takes local minimum value.

4.2. Kalman Filter Based Frame-Dependent-Process

If the process carried out at the previous analysis frame is the time-independent-process, the initial *a posteriori* density function $p(\boldsymbol{\Theta}_0|y_0)$ can tentatively be set according to the ML estimate of μ_k.

4.3. Process Selection at Each Analysis Frame

The frame-independent-process is carried out at the initial analysis frame and after that, the alternative processes are selectively carried out according to KL divergence between the multi-pitch spectral model and the observed spectrum. The process is simply selected with a threshold procedure (if the KL divergence is smaller than a threshold, frame-dependent-process will be selected, and if larger, frame-independent-process will be selected to reexamine whether or not a new speaker has appeared). Note that this is no more than an *ad hoc* way and an appropriate decision criteria for the process selection should further be investigated.

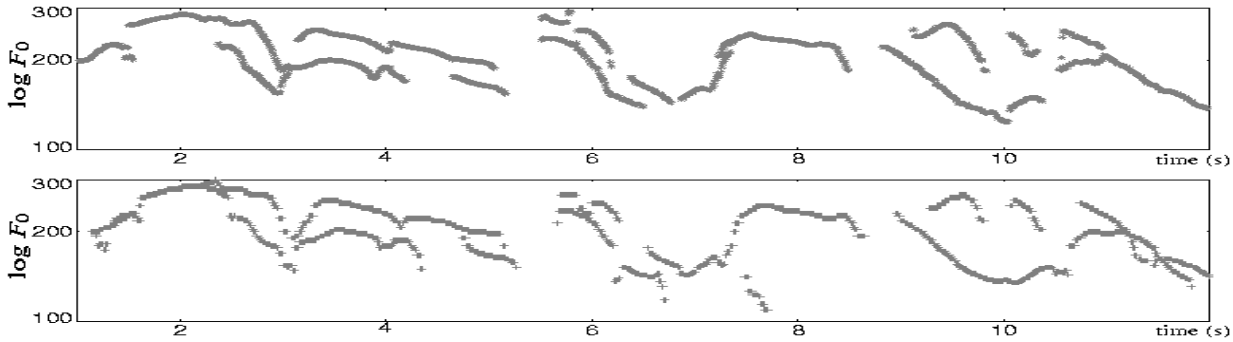

Figure 4: *A Multi-pitch estimation result(top) and the corresponding reference F_0 patterns(bottom)*

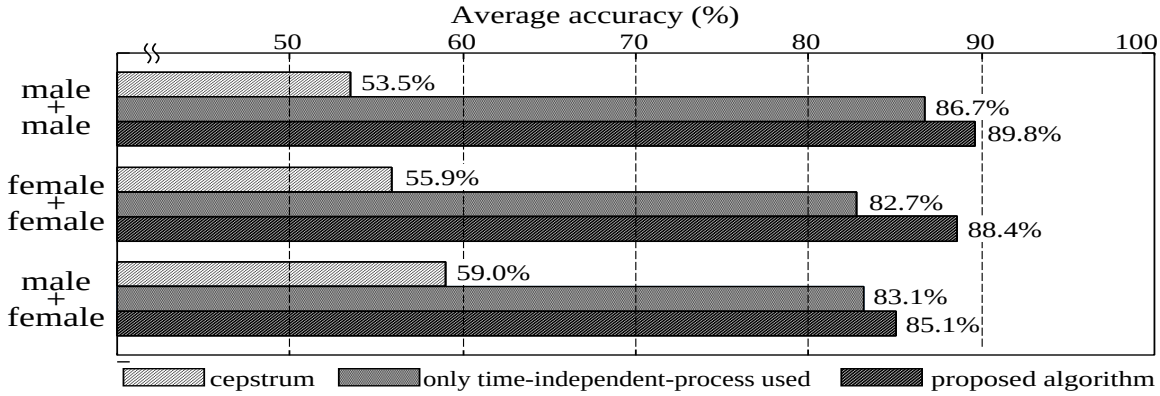

Figure 5: *Accuracy comparison of the proposed method and the time-independent-process only applied*

5. Experiments

To evaluate the performance of our algorithm, it was tested on 23 simultaneous speech data. Each of them was artificially created by mixing two Japanese sentence signals with 0 dB signal-to-signal ratio, excerpted from set-A of ATR speech database. All signals were digitized at 12 kHz sampling rate and analyzed with Gaussian window where frame length and shift were 64 ms and 10 ms, respectively. The initial number of harmonic GMMs at the time-independent-process was set to 6. Deviations over 5% from the references were defined as gross errors.

Fig 5 shows a comparison of the proposed algorithm and the case in which the frame-independent-process was only used at every frame. A preliminary comparison with cepstrum is also shown. Fig 4 shows an example of the estimated multi-pitch trajectories and the corresponding reference F_0 patterns. The Kalman filter based procedure effectively improved the estimation accuracy especially on continuous F_0 trajectory segments, while some errors were found due to the process selection mistakes such that even when passing through a discontinuous parts of F_0 or speech offsets, the frame-dependent-process were continually carried out.

6. Conclusions

In this paper, we proposed a multi-pitch trajectory estimation procedure using harmonic GMM and nonlinear Kalman filtering that reflects the local time dependency property of F_0 pattern. This procedure was experimentally shown to be effective for multi-pitch tracking of simltaneous(competitive) speech. Our future work includes investigation of a decision criteria of phrase boundary or speech onset/offset in order to select appropriate alternative processes at each analysis frame. Furthermore, the evaluation in noisy environments, the possibility of noise reduction purpose, and more sophisticated spectral modeling (e.g., 3 demensional spectrogram modeling) will also be investigated.

7. References

[1] Chazan, D., Stettiner, Y., and Malah, D., "Optimal Multi-pitch Estimation Using the EM Algorithm for Co-channel Speech Separation," Proc. IEEE, ICASSP 93, Vol. 2, pp. 728–731, 1993.

[2] Nishi, K., Abe, M., and Ando, S.,"Optimum Harmonics Tracking Filter for Auditory Scene Analysis," Proc. IEEE, ICASSP 96, pp. 573-576, 1996.

[3] Godsill, S., and Davy, M., "Baysian Harmonic Models for Musical Pitch Estimation and Analysis," Proc. IEEE, ICASSP 2002, Vol. 2, pp. 1769–1772, 2002.

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

[5] Jazwinski, A., "Stochastic Process and Filtering Theory," Academic Press, New York, chap. 6, pp. 162–193, 1970.

[6] Kameoka, H., Nishimoto, T., and Sagayama, S., "Multi-Pitch Detection Algorithm Using Constrained Gaussian Mixture Model and Information Criterion for Simultaneous Speech," Proc. Speech Prosody 2004 (SP2004), pp. 533–536, 2004.