Proceedings of the 18th International Congress on Acoustics (ICA 2004), pp.II-1085-1088, April 2004. (Invited Paper)

A Predominant-F0 Estimation Method for Polyphonic Musical Audio Signals

Masataka Goto

National Institute of Advanced Industrial Science and Technology (AIST)

m.goto@aist.go.jp

Abstract

In this paper I introduce a method, called *PreFEst*, for estimating the fundamental frequency (F0) of simultaneous sounds in monaural polyphonic audio signals. Most previous F0-estimation methods have had difficulty dealing with such complex audio signals because these methods were designed to deal with mixtures of only a few sounds. Without assuming the number of sound sources, PreFEst can estimate the relative dominance of every possible harmonic structure in the input mixture. It treats the mixture as if it contains all possible harmonic structures with different weights, and estimates their weights by MAP estimation. PreFEst can obtain the melody and bass lines by regarding the most predominant F0 in middle- and high-frequency regions as the melody line and the one in a low-frequency region as the bass line. Experimental results with compact-disc recordings showed that a real-time system implementing this method was able to detect melody and bass lines about 80% of the time these existed.

1. Introduction

The estimation of the fundamental frequency (F0) of simultaneous sounds in monaural polyphonic mixtures is important to build a computational model that can understand audio signals in a human-like fashion. Moreover, it is useful in various practical applications, such as automatic indexing for information retrieval and intelligent editing of audio signals. It has, however, been considered difficult to estimate the F0 in such real-world audio signals; this is because the number of sound sources in them generally cannot be assumed, because the frequency components of one sound often overlap the frequency components of simultaneous sounds, and because the F0's frequency component (the frequency component corresponding to the F0) is sometimes very weak or missing (*missing fundamental*). Most previous F0 estimation methods have been premised upon the input audio signal containing just a single-pitch sound with aperiodic noise. Although several methods for dealing with multiple-pitch mixtures have been proposed [1, 2, 3, 4], these require that the number of simultaneous sounds be assumed and have difficulty estimating the F0 in complex audio signals like those sampled from music compact discs.

This paper introduces a Predominant-F0 Estimation method, called *PreFEst*, that I developed during 1999 and 2001 [5, 6, 7, 8]. PreFEst can estimate the fundamental frequency (F0) of melody and bass lines in monaural audio signals containing simultaneous sounds of various musical instruments. Unlike previous methods, PreFEst does not assume the number of sound sources, locally trace frequency components, or even rely on the existence of the F0's frequency component. PreFEst basically estimates the F0 of the most predominant harmonic structure — the most predominant F0 corresponding to the melody or bass line — within an intentionally limited frequency range of the input mixture. It simultaneously takes into consideration all possibilities for the F0 and treats the input mixture as if it contains all possible harmonic structures with different weights (amplitudes). It regards a probability density function (PDF) of the input frequency components as a weighted mixture of harmonic-structure tone models (represented by PDFs) of all possible F0s and simultaneously estimates both their weights corresponding to the relative dominance of every possible harmonic structure and the shape of the tone models by MAP (Maximum *A Posteriori* Probability) estimation considering their prior distribution. It then considers the maximum-weight model as the most predominant harmonic structure and obtains its F0. The method also considers the F0's temporal continuity by using a multiple-agent architecture.

2. Predominant-F0 estimation method: PreFEst

Figure 1 shows an overview of PreFEst. PreFEst consists of three components, the *PreFEst-front-end* for frequency analysis, the *PreFEst-core* to estimate the predominant F0, and the *PreFEst-back-end* to evaluate the temporal continuity of the F0. Since the melody line tends to have the most predominant harmonic structure in middle- and

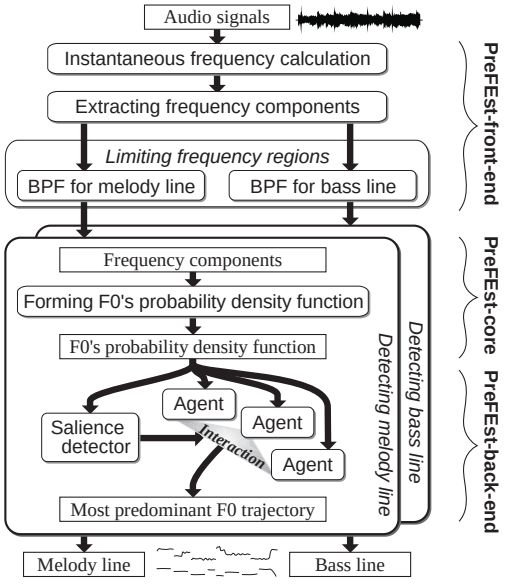

Figure 1: Overview of PreFEst.

high-frequency regions and the bass line tends to have the most predominant harmonic structure in a low-frequency region, the F0s of the melody and bass lines can be estimated by applying the PreFEst-core with appropriate frequency-range limitation.

2.1. PreFEst-front-end: Forming the observed probability density functions

The PreFEst-front-end first uses an STFT-based multirate filter bank in order to obtain adequate time-frequency resolution under the real-time constraint. It then extracts frequency components by using an instantaneous-frequency-related measure [5, 8] and obtains two sets of bandpass-filtered frequency components, one for the melody line (261.6 - 4186 Hz) and the other for the bass line (32.7 - 261.6 Hz). To enable the application of statistical methods, each set of the bandpass-filtered components is represented as a probability density function (PDF), called an *observed PDF*, $p_{\Psi}^{(t)}(x)$, where t is the time measured in units of frame-shifts (10 ms), and x is the log-scale frequency denoted in units of *cents* [1].

2.2. PreFEst-core: Estimating the F0's probability density function

For each set of filtered frequency components represented as an observed PDF $p_{\Psi}^{(t)}(x)$, the PreFEst-core forms a probability density function of the F0, called the *F0's PDF*, $p_{F0}^{(t)}(F)$, where F is the log-scale frequency in cents. We consider each observed PDF to have been generated from a weighted-mixture model of the tone models of all the possible F0s; a tone model is the PDF corresponding to a typical harmonic structure and indicates where the harmonics of the F0 tend to occur. Because the weights of tone models represent the relative dominance of every possible harmonic structure, these weights can be regarded as the F0's PDF: the more dominant a tone model is in the mixture, the higher the probability of the F0 of its model.

2.2.1. Weighted-mixture model of adaptive tone models

To deal with diversity of the harmonic structure, the PreFEst-core can use several types of harmonic-structure tone models. The PDF of the m-th tone model for each F0 F is denoted by $p(x|F, m, \mu^{(t)}(F, m))$ (Figure 2), where the model parameter $\mu^{(t)}(F, m)$ represents the shape of the tone model. The number of tone models is M_i ($1 \leq m \leq \mathrm{M}_i$) where i denotes the melody line (i = m) or the bass line (i = b). Each tone model is defined by

$$p(x|F, m, \mu^{(t)}(F, m)) = \sum_{h=1}^{\mathrm{H}_i} p(x, h|F, m, \mu^{(t)}(F, m)), \quad (1)$$

$$p(x, h|F, m, \mu^{(t)}(F, m)) = c^{(t)}(h|F, m)\, G(x; F + 1200 \log_2 h, \mathrm{W}_i), \quad (2)$$

$$\mu^{(t)}(F, m) = \{c^{(t)}(h|F, m) \mid h = 1, ..., \mathrm{H}_i\}, \quad (3)$$

$$G(x; x_0, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{(x-x_0)^2}{2\sigma^2}}, \quad (4)$$

[1] In this paper I define that frequency f_{Hz} in hertz is converted to frequency f_{cent} in cents so that there are 100 cents to a tempered semitone and 1200 to an octave: $f_{\mathrm{cent}} = 1200 \log_2(f_{\mathrm{Hz}} \,/\, (440 \times 2^{\frac{3}{12}-5}))$.

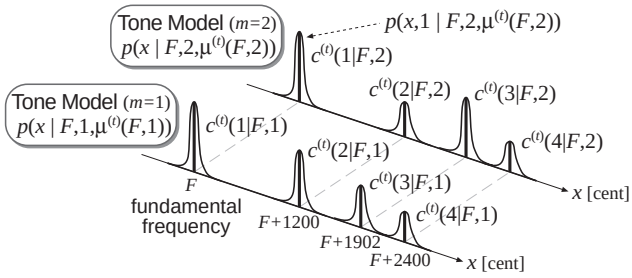

Figure 2: Model parameters of multiple adaptive tone models.

where H_i is the number of harmonics considered, W_i is the standard deviation σ of the Gaussian distribution $G(x; x_0, \sigma)$, and $c^{(t)}(h|F, m)$ determines the relative amplitude of the h-th harmonic component (the shape of the tone model) and satisfies

$$\sum_{h=1}^{\mathrm{H}_i} c^{(t)}(h|F, m) = 1. \quad (5)$$

In short, this tone model places a weighted Gaussian distribution at the position of each harmonic component.

We then consider the observed PDF $p_{\Psi}^{(t)}(x)$ to have been generated from the following model $p(x|\theta^{(t)})$, which is a weighted mixture of all possible tone models $p(x|F, m, \mu^{(t)}(F, m))$:

$$p(x|\theta^{(t)}) = \int_{\mathrm{Fl}_i}^{\mathrm{Fh}_i} \sum_{m=1}^{\mathrm{M}_i} w^{(t)}(F, m)\; p(x|F, m, \mu^{(t)}(F, m))\; dF, \quad (6)$$

$$\theta^{(t)} = \{w^{(t)}, \mu^{(t)}\}, \quad (7)$$

$$w^{(t)} = \{w^{(t)}(F, m) \mid \mathrm{Fl}_i \leq F \leq \mathrm{Fh}_i, m = 1, ..., \mathrm{M}_i\}, \quad (8)$$

$$\mu^{(t)} = \{\mu^{(t)}(F, m) \mid \mathrm{Fl}_i \leq F \leq \mathrm{Fh}_i, m = 1, ..., \mathrm{M}_i\}, \quad (9)$$

where Fl_i and Fh_i denote the lower and upper limits of the possible (allowable) F0 range and $w^{(t)}(F, m)$ is the weight of a tone model $p(x|F, m, \mu^{(t)}(F, m))$ that satisfies

$$\int_{\mathrm{Fl}_i}^{\mathrm{Fh}_i} \sum_{m=1}^{\mathrm{M}_i} w^{(t)}(F, m)\; dF = 1. \quad (10)$$

Because we cannot know *a priori* the number of sound sources, it is important that we simultaneously take into consideration all F0 possibilities as expressed in Equation (6). If we can estimate the model parameter $\theta^{(t)}$ such that the observed PDF $p_{\Psi}^{(t)}(x)$ is likely to have been generated from the model $p(x|\theta^{(t)})$, the weight $w^{(t)}(F, m)$ can be interpreted as the F0's PDF $p_{F0}^{(t)}(F)$:

$$p_{F0}^{(t)}(F) = \sum_{m=1}^{\mathrm{M}_i} w^{(t)}(F, m) \;\; (\mathrm{Fl}_i \leq F \leq \mathrm{Fh}_i). \quad (11)$$

2.2.2. Introducing a prior distribution

To use prior knowledge about F0 estimates and the tone-model shapes, we define a prior distribution $p_{0i}(\theta^{(t)})$ of $\theta^{(t)}$ as follows:

$$p_{0i}(\theta^{(t)}) = p_{0i}(w^{(t)})\; p_{0i}(\mu^{(t)}), \quad (12)$$

$$p_{0i}(w^{(t)}) = \frac{1}{Z_w} e^{-\beta_{wi}^{(t)}\, D_w(w_{0i}^{(t)}; w^{(t)})}, \quad (13)$$

$$p_{0i}(\mu^{(t)}) = \frac{1}{Z_\mu} e^{-\int_{\mathrm{Fl}_i}^{\mathrm{Fh}_i} \sum_{m=1}^{\mathrm{M}_i} \beta_{\mu i}^{(t)}(F,m)\, D_\mu(\mu_{0i}^{(t)}(F,m);\mu^{(t)}(F,m))\, dF}. \quad (14)$$

Here $p_{0i}(w^{(t)})$ and $p_{0i}(\mu^{(t)})$ are unimodal distributions: $p_{0i}(w^{(t)})$ takes its maximum value at $w_{0i}^{(t)}(F,m)$ and $p_{0i}(\mu^{(t)})$ takes its maximum value at $\mu_{0i}^{(t)}(F,m)$, where $w_{0i}^{(t)}(F,m)$ and $\mu_{0i}^{(t)}(F,m)$ ($c_{0i}^{(t)}(h|F,m)$) are the most probable parameters. Z_w and Z_μ are normalization factors, and $\beta_{wi}^{(t)}$ and $\beta_{\mu i}^{(t)}(F,m)$ are parameters determining how much emphasis is put on the maximum value. The prior distribution is not informative (i.e., it is uniform) when $\beta_{wi}^{(t)}$ and $\beta_{\mu i}^{(t)}(F,m)$ are 0, corresponding to the case when no prior knowledge is available. In Equations (13) and (14), $D_w(w_{0i}^{(t)}; w^{(t)})$ and $D_\mu(\mu_{0i}^{(t)}(F,m); \mu^{(t)}(F,m))$ are the following Kullback-Leibler information:

$$D_w(w_{0i}^{(t)}; w^{(t)}) = \int_{\mathrm{Fl}_i}^{\mathrm{Fh}_i} \sum_{m=1}^{\mathrm{M}_i} w_{0i}^{(t)}(F,m)\, log \frac{w_{0i}^{(t)}(F,m)}{w^{(t)}(F,m)} dF, \quad (15)$$

$$D_\mu(\mu_{0i}^{(t)}(F,m); \mu^{(t)}(F,m)) = \sum_{h=1}^{\mathrm{H}_i} c_{0i}^{(t)}(h|F,m)\, log \frac{c_{0i}^{(t)}(h|F,m)}{c^{(t)}(h|F,m)}. \quad (16)$$

2.2.3. MAP estimation using the EM algorithm

The problem to be solved is to estimate the model parameter $\theta^{(t)}$, taking into account the prior distribution $p_{0i}(\theta^{(t)})$, when we observe $p_\Psi^{(t)}(x)$. The MAP estimator of $\theta^{(t)}$ is obtained by maximizing

$$\int_{-\infty}^{\infty} p_\Psi^{(t)}(x)\, (\log p(x|\theta^{(t)}) + \log p_{0i}(\theta^{(t)}))\, dx. \quad (17)$$

Because this maximization problem is too difficult to solve analytically, we use the Expectation-Maximization (EM) algorithm [9], which is an algorithm iteratively applying two steps — the *expectation step (E-step)* and the *maximization step (M-step)* — to compute MAP estimates from incomplete observed data (i.e., from $p_\Psi^{(t)}(x)$). With respect to $\theta^{(t)}$, each iteration updates the old estimate $\theta'^{(t)} = \{w'^{(t)}, \mu'^{(t)}\}$ to obtain the new (improved) estimate $\overline{\theta^{(t)}} = \{\overline{w^{(t)}}, \overline{\mu^{(t)}}\}$. For each frame t, $w'^{(t)}$ is initialized with the final estimate $\overline{w^{(t-1)}}$ after iterations at the previous frame $t-1$; $\mu'^{(t)}$ is initialized with the most probable parameter $\mu_{0i}^{(t)}$ in the current implementation.

By introducing the hidden (unobservable) variables F, m, and h, which, respectively, describe which F0, which tone model, and which harmonic component were responsible for generating each observed frequency component at x, we can specify the two steps as follows:

1. (E-step)
   Compute the following $Q_{\mathrm{MAP}}(\theta^{(t)}|\theta'^{(t)})$ for the MAP estimation:
   $$Q_{\mathrm{MAP}}(\theta^{(t)}|\theta'^{(t)}) = Q(\theta^{(t)}|\theta'^{(t)}) + \log p_{0i}(\theta^{(t)}), \quad (18)$$
   $$Q(\theta^{(t)}|\theta'^{(t)}) = \int_{-\infty}^{\infty} p_\Psi^{(t)}(x)\, \mathrm{E}_{F,m,h}[\log p(x,F,m,h|\theta^{(t)}) \mid x, \theta'^{(t)}]\, dx, \quad (19)$$
   where $Q(\theta^{(t)}|\theta'^{(t)})$ is the conditional expectation of the mean log-likelihood for the maximum likelihood estimation. $\mathrm{E}_{F,m,h}[a|b]$ denotes the conditional expectation of a with respect to the hidden variables F, m, and h, with the probability distribution determined by condition b.
2. (M-step)
   Maximize $Q_{\mathrm{MAP}}(\theta^{(t)}|\theta'^{(t)})$ as a function of $\theta^{(t)}$ to obtain the updated (improved) estimate $\overline{\theta^{(t)}}$:
   $$\overline{\theta^{(t)}} = \operatorname*{argmax}_{\theta^{(t)}} Q_{\mathrm{MAP}}(\theta^{(t)}|\theta'^{(t)}). \quad (20)$$

In the E-step, $Q(\theta^{(t)}|\theta'^{(t)})$ is expressed as

$$Q(\theta^{(t)}|\theta'^{(t)}) = \int_{-\infty}^{\infty} \int_{\mathrm{Fl}_i}^{\mathrm{Fh}_i} \sum_{m=1}^{\mathrm{M}_i} \sum_{h=1}^{\mathrm{H}_i} p_\Psi^{(t)}(x)\, p(F,m,h|x,\theta'^{(t)}) \log p(x,F,m,h|\theta^{(t)}) dF dx, \quad (21)$$

where the complete-data log-likelihood is given by

$$\log p(x,F,m,h|\theta^{(t)}) = \log(w^{(t)}(F,m)\, p(x,h|F,m,\mu^{(t)}(F,m))). \quad (22)$$

Regarding the M-step, Equation (20) is a conditional problem of variation, where the conditions are given by Equations (5) and (10). This problem can be solved by using Euler-Lagrange differential equations with Lagrange multipliers [7, 8] and we obtain the following new parameter estimates:

$$\overline{w^{(t)}(F,m)} = \frac{\overline{w_{\mathrm{ML}}^{(t)}(F,m)} + \beta_{wi}^{(t)} w_{0i}^{(t)}(F,m)}{1 + \beta_{wi}^{(t)}}, \quad (23)$$

$$\overline{c^{(t)}(h|F,m)} = \frac{\overline{w_{\mathrm{ML}}^{(t)}(F,m)}\ \overline{c_{\mathrm{ML}}^{(t)}(h|F,m)} + \beta_{\mu i}^{(t)}(F,m) c_{0i}^{(t)}(h|F,m)}{\overline{w_{\mathrm{ML}}^{(t)}(F,m)} + \beta_{\mu i}^{(t)}(F,m)}, \quad (24)$$

where $\overline{w_{\mathrm{ML}}^{(t)}(F,m)}$ and $\overline{c_{\mathrm{ML}}^{(t)}(h|F,m)}$ are, when the noninformative prior distribution ($\beta_{wi}^{(t)} = 0$ and $\beta_{\mu i}^{(t)}(F,m) = 0$) is given, the following maximum likelihood estimates:

$$\overline{w_{\mathrm{ML}}^{(t)}(F,m)} = \int_{-\infty}^{\infty} p_\Psi^{(t)}(x) \frac{w'^{(t)}(F,m)\, p(x|F,m,\mu'^{(t)}(F,m))}{\int_{\mathrm{Fl}_i}^{\mathrm{Fh}_i} \sum_{\nu=1}^{\mathrm{M}_i} w'^{(t)}(\eta,\nu)\, p(x|\eta,\nu,\mu'^{(t)}(F,\nu))\, d\eta} dx, \quad (25)$$

$$\overline{c_{\mathrm{ML}}^{(t)}(h|F,m)} = \frac{1}{\overline{w_{\mathrm{ML}}^{(t)}(F,m)}} \int_{-\infty}^{\infty} p_\Psi^{(t)}(x) \frac{w'^{(t)}(F,m)\, p(x,h|F,m,\mu'^{(t)}(F,m))}{\int_{\mathrm{Fl}_i}^{\mathrm{Fh}_i} \sum_{\nu=1}^{\mathrm{M}_i} w'^{(t)}(\eta,\nu)\, p(x|\eta,\nu,\mu'^{(t)}(F,\nu))\, d\eta} dx. \quad (26)$$

For an intuitive explanation of Equation (25), we call $\frac{w'^{(t)}(F,m)\, p(x|F,m,\mu'^{(t)}(F,m))}{\int_{\mathrm{Fl}_i}^{\mathrm{Fh}_i} \sum_{\nu=1}^{\mathrm{M}_i} w'^{(t)}(\eta,\nu)\, p(x|\eta,\nu,\mu'^{(t)}(F,\nu))\, d\eta}$ the decomposition filter. For the integrand on the right side of Equation (25), we can consider that, because of this filter, the value of $p_\Psi^{(t)}(x)$ at frequency x is decomposed into (is distributed among) all possible tone models $p(x|F,m,\mu'^{(t)}(F,m))$ ($\mathrm{Fl}_i \le F \le \mathrm{Fh}_i$, $1 \le m \le \mathrm{M}_i$) in proportion to the numerator of the decomposition filter at x. The higher the weight $w'^{(t)}(F,m)$, the larger the decomposed value given to the corresponding tone model. Note that the value of $p_\Psi^{(t)}(x)$ at different x is also decomposed according to a different ratio in proportion to the numerator of the decomposition filter at that x. Finally, the updated weight

$\overline{w^{(t)}_{\mathrm{ML}}(F,m)}$ is obtained by integrating all the decomposed values given to the corresponding m-th tone model for the F0 F.

We think that this decomposition behavior is the advantage of PreFEst in comparison to previous comb-filter-based or autocorrelation-based methods [1, 2, 3]. This is because these previous methods cannot easily support the decomposition of an overlapping frequency component (overtone) shared by several simultaneous tones and tend to have difficulty distinguishing sounds with overlapping overtones. In addition, PreFEst can simultaneously estimate all the weights $\overline{w^{(t)}_{\mathrm{ML}}(F,m)}$ (for all the range of F) so that these weights can be optimally balanced: it does not determine the weight at F after determining the weight at another F. We think this simultaneous estimation of all the weights is an advantage of PreFEst compared to previous recursive-subtraction-based methods [1, 4] where components of the most dominant harmonic structure identified are subtracted from a mixture and then this is recursively done again starting from the residue of the previous subtraction. In these methods, once inappropriate identification or subtraction occurs, the following recursions starting from the wrong residue become unreliable.

After the above iterative computation of Equations (23) and (24), the F0's PDF $p^{(t)}_{F0}(F)$ can be obtained from $w^{(t)}(F,m)$ according to Equation (11). We can also obtain the tone-model shape $c^{(t)}(h|F,m)$, which is the relative amplitude of each harmonic component of all types of tone models $p(x|F,m,\mu^{(t)}(F,m))$.

2.3. PreFEst-back-end: Sequential F0 tracking by multiple-agent architecture

A simple way to identify the most predominant F0 is to find the frequency that maximizes the F0's PDF. This result is not always stable, however, because peaks corresponding to the F0s of simultaneous sounds sometimes compete in the F0's PDF for a moment and are transiently selected, one after another, as the maximum.

We therefore consider the global temporal continuity of the F0 by using a multiple-agent architecture [5, 6, 8] in which agents track different temporal trajectories of the F0. The final F0 output is determined on the basis of the most dominant and stable F0 trajectory.

3. Experimental results

The PreFEst has been implemented in a real-time system that takes a musical audio signal as input and outputs the detected melody and bass lines in several forms, such as audio signals and computer graphics. The current implementation uses two adaptive tone models and main parameter values are described in [7, 8]. The system was tested on excerpts from 10 musical pieces in the popular, jazz, and orchestral genres. The 20-s-long input monaural audio signals — each containing a single-tone melody and the sounds of several instruments — were sampled from compact discs. We evaluated the detection rates by comparing the estimated F0s with the correct F0s that were hand-labeled using our F0 editor program [6, 8].

In our experiment the system correctly detected, for most parts of each audio sample, the melody lines provided by a voice or a single-tone mid-range instrument and the bass lines provided by a bass guitar or a contrabass: the average detection rate was 88.4% for the melody line and 79.9% for the bass line.

4. Conclusion

I have introduced the PreFEst method that estimates the most predominant F0 in a monaural sound mixture without assuming the number of sound sources. Although PreFEst has great potential, I have not fully exploited it. In the future, for example, many different tone models could be prepared by analyzing or learning various kinds of harmonic structure that appear in music and multiple peaks in the F0's PDF, each corresponding to a different sound source, could be tracked simultaneously by using a sound source discrimination method. While I dealt with only harmonic-structure tone models in this paper, PreFEst can be applied to any weighted mixture of arbitrary tone models (even if their components are inharmonic) by simply replacing Equation (2) with $p(x,h|F,m,\mu^{(t)}(F,m)) = c^{(t)}(h|F,m)p_{\mathit{arbit}}(x;F,h,m)$, where $p_{\mathit{arbit}}(x;F,h,m)$ is an arbitrary PDF (h is merely the component number in this case). Even with this general tone model, in theory the F0's PDF can be estimated by using the same Equations (23) and (24). Both any harmonic- and inharmonic-structure tone models can also be used together. Moreover, PreFEst can be applied to non-music audio signals. In fact, Masuda-Katsuse [10] has extended it and demonstrated its effectiveness for speech recognition in realistic noisy environments.

Acknowledgments: I thank Shotaro Akaho and Hideki Asoh for their valuable discussions.

5. References

[1] Alain de Cheveigné, "Separation of concurrent harmonic sounds: Fundamental frequency estimation and a time-domain cancellation model of auditory processing," *J. Acoust. Soc. Am.*, 93(6):3271–3290, 1993.

[2] Alain de Cheveigné and Hideki Kawahara, "Multiple period estimation and pitch perception model," *Speech Communication*, 27(3–4):175–185, 1999.

[3] Tero Tolonen and Matti Karjalainen, "A computationally efficient multipitch analysis model," *IEEE Trans. on Speech and Audio Processing*, 8(6):708–716, 2000.

[4] Anssi P. Klapuri, "Multipitch estimation and sound separation by the spectral smoothness principle," In *Proc. ICASSP 2001*, 2001.

[5] Masataka Goto, "A real-time music scene description system: Detecting melody and bass lines in audio signals," In *Working Notes of the IJCAI-99 Workshop on Computational Auditory Scene Analysis*, pp. 31–40, 1999.

[6] Masataka Goto, "A robust predominant-F0 estimation method for real-time detection of melody and bass lines in CD recordings," In *Proc. ICASSP 2000*, pp. II–757–760, 2000.

[7] Masataka Goto, "A predominant-F0 estimation method for CD recordings: MAP estimation using EM algorithm for adaptive tone models," In *Proc. ICASSP 2001*, pp. V–3365–3368, 2001.

[8] Masataka Goto, "A real-time music scene description system: Predominant-F0 estimation for detecting melody and bass lines in real-world audio signals," *Speech Communication*, 2004, (accepted).

[9] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Roy. Stat. Soc. B*, 39(1):1–38, 1977.

[10] Ikuyo Masuda-Katsuse, "A new method for speech recognition in the presence of non-stationary, unpredictable and high-level noise," In *Proc. Eurospeech 2001*, pp. 1119–1122, 2001.