ICASSP 2001 (The 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing) Proceedings, pp. V-3365-3368, May 2001.

A PREDOMINANT-F0 ESTIMATION METHOD FOR CD RECORDINGS: MAP ESTIMATION USING EM ALGORITHM FOR ADAPTIVE TONE MODELS

Masataka Goto

"Information and Human Activity", PRESTO, Japan Science and Technology Corporation (JST). / Electrotechnical Laboratory (National Institute of Advanced Industrial Science and Technology).
1-1-4 Umezono, Tsukuba, Ibaraki 305-8568, JAPAN
goto@etl.go.jp

ABSTRACT

This paper describes a predominant-F0 (fundamental frequency) estimation method called *PreFEst*, which can detect melody and bass lines in monaural audio signals containing sounds of various instruments. While most previous methods premised mixtures of a few sounds and had difficulty dealing with such complex signals, our method can estimate the F0 of the melody and bass lines without assuming the number of sound sources in compact-disc recordings. In this paper we propose the following three extensions to our previous PreFEst to make it more adaptive and flexible: introducing multiple harmonic-structure tone models, estimating the shape of tone models, and introducing a prior distribution of its shape and F0 estimates. These extensions were implemented by the MAP (Maximum *A Posteriori* Probability) estimation by using the *Expectation-Maximization* algorithm. Experimental results with compact-disc recordings showed that our real-time system based on the extended PreFEst achieved performance improvement.

1. INTRODUCTION

Our goal is to build a real-time system that can detect melody and bass lines in monaural complex real-world audio signals, such as those sampled from commercially distributed compact discs. This detection is an important initial step in computer emulation of human music understanding because the melody and bass lines are fundamental to the perception of Western music. In addition, the detected melody and bass lines are useful in various practical applications, such as automatic music indexing for information retrieval (e.g., searching for a song by singing a melody), computer participation in live human performances, and analysis of recordings of outstanding performances.

Although this detection requires the estimation of the fundamental frequency (F0, perceived as pitch) of the melody and bass lines, it has been considered very difficult to estimate the F0 in complex audio signals sampled from compact discs. The main reasons are: in compact-disc recordings, the number of sound sources cannot be assumed, the frequency components of one sound often overlap frequency components of simultaneous sounds, and the F0's frequency component (the frequency component corresponding to the F0) is sometimes missing or very weak (*missing fundamental*). Most previous F0-estimation methods [1, 2, 3, 4], however, premised that the input contained just a single-pitch sound with aperiodic noises. Although several methods for dealing with multiple-pitch mixtures were proposed [5, 6, 7, 8], they dealt with at most three musical instruments or voices and had difficulty dealing with compact-disc recordings.

We therefore developed a method, called *PreFEst* (Predominant-F0 Estimation method), that can detect the melody and bass lines in complex mixtures containing simultaneous sounds of various instruments (even drums) [9]. PreFEst has the advantages that it does not assume the number of sound sources, it does not locally trace frequency components, and it does not rely on the existence of the F0's frequency component. It basically estimates the F0 of the most predominant harmonic structure in the input sound mixture; it simultaneously takes into consideration all the possibilities of F0 and considers that the input mixture contains every possible harmonic structure with different weights (amplitude). It regards the input frequency components as a weighted mixture of harmonic-structure tone models of all possible F0s and then finds the F0 of the maximum-weight model corresponding to the most predominant harmonic structure.

PreFEst reported in our earlier paper [9] had three limitations. First, although various kinds of harmonic structure appear at different F0s and even at the same F0, just a single harmonic-structure tone model was prepared for each F0. Second, the shape of tone model was fixed as if one ideal tone model was always assumed: the relative amplitude of each harmonic component was constant. Because it did not exactly coincide with the harmonic structure contained in the input, there was room for refining the tone-model management in a more adaptive way. Third, even if prior knowledge about very rough F0 estimates of the melody and bass lines is available, we were not able to incorporate it into the estimation. Such prior rough estimates can be given in some practical applications where more precise F0 with less errors is required; for example, the analysis of expression in a recorded performance needs to estimate the actual F0 by using its rough estimate that can be given by a score or by playing a MIDI instrument along to the original recorded performance.

In the following sections we describe how we extended our previous PreFEst so that it can overcome the above three limitations. We first give an overview of PreFEst described in [9] and then describe the details of the three extensions that make it possible to overcome the above limitations. The main idea is to introduce multiple adaptive tone models and prior knowledge, and to estimate the model parameters on the basis of the MAP (Maximum *A Posteriori* Probability) estimation by using the *Expectation-Maximization (EM)* algorithm. Finally, we show experimental results of a real-time system based on the extended method.

2. OVERVIEW OF PREFEST

PreFEst estimates the most predominant F0 in frequency-range-limited sound mixtures. Since the melody line tends to have

0-7803-7041-4/01/$10.00

the most predominant harmonic structure in middle- and high-frequency regions and the bass line tends to have the most predominant harmonic structure in a low-frequency region, we can estimate the F0s of the melody and bass lines by applying PreFEst with appropriate frequency-range limitation [9, 10].

To use statistical methods, we represent the input bandpass-filtered frequency components as a probability density function (PDF), called *observed PDF*, which is estimated by using multirate signal-processing techniques and instantaneous-frequency-related measure [9, 10]. We then consider that the observed PDF has been generated from a weighted-mixture model of tone models of all possible F0s; a tone model is the PDF corresponding to a typical harmonic structure of every possible F0. Because the weights of tone models represent the relative dominance of every possible harmonic structure, we can regard those weights as the PDF of the F0 (*F0's PDF*): the more dominant a tone model in the mixture, the higher the probability of the F0 of its model. As explained in our earlier papers [9, 10], those weights (i.e., the F0's PDF) can be estimated by using the EM algorithm [11].

A simple way of determining the most predominant F0 is to find the frequency that maximizes the F0's PDF. This result is not 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 in which agents track different temporal trajectories of the F0, and the final F0 output is determined on the basis of the most dominant and stable F0 trajectory.

3. THREE EXTENSIONS

To overcome the three limitations described in the Introduction, we propose the following three extensions of the previous PreFEst.

[Extension 1] Introducing multiple tone models
We prepare multiple tone models for each F0 and consider their mixture model.

[Extension 2] Estimating the shape of tone models
We consider, as model parameters, the relative amplitude of each harmonic component of all the tone models (called *the shape of tone model*) as well as their weights, and estimate them by using the EM algorithm.

[Extension 3] Introducing a prior distribution
To estimate the model parameters on the basis of their prior distribution, we use the MAP estimation instead of the maximum likelihood estimation used in [9, 10]. Note that we can also take into consideration a prior distribution of the shape of tone model added to the model parameters by *[Extension 2]*.

All of these extensions are dealt with in the process of estimating the F0's PDF $p^{(t)}_{F0}(F)$ from the observed PDF (the PDF of the bandpass-filtered frequency components), $p^{(t)}_{\Psi}(x)$. Here, t is the time measured in units of frame-shift (10 msec), and x and F are the log-scale frequency denoted in units of *cents* (a musical-interval measurement). Frequency f_{Hz} in hertz is converted to frequency f_{cent} in cents as follows:

$$f_{\mathrm{cent}} = 1200 \log_2 \frac{f_{\mathrm{Hz}}}{440 \times 2^{\frac{3}{12}-5}}. \quad (1)$$

3.1. Weighted-mixture model of extended tone models

For *[Extension 1]* and *[Extension 2]*, we prepare multiple tone models for each F0, F, and introduce the model parameter $\mu^{(t)}(F, m)$ to the m-th tone model whose PDF is denoted as

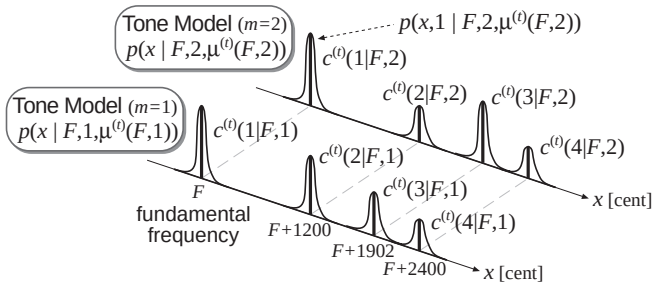

Figure 1: Model parameters of multiple adaptive tone models.

$p(x|F, m, \mu^{(t)}(F, m))$ (Figure 1). 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). The tone model that indicates where the harmonics of the F0, F, tend to occur is defined as

$$p(x|F, m, \mu^{(t)}(F, m)) = \sum_{h=1}^{\mathrm{H}_i} p(x, h|F, m, \mu^{(t)}(F, m)), \quad (2)$$

$$p(x, h|F, m, \mu^{(t)}(F, m)) = c^{(t)}(h|F, m)\ G(x; F + 1200 \log_2 h, \mathrm{W}_i), \quad (3)$$

$$\mu^{(t)}(F, m) = \{c^{(t)}(h|F, m) \mid h = 1, ..., \mathrm{H}_i\}, \quad (4)$$

$$G(x; x_0, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}}\ e^{-\frac{(x-x_0)^2}{2\sigma^2}}, \quad (5)$$

where H_i is the number of harmonics considered, W_i^2 is the variance 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 tone model) and satisfies

$$\sum_{h=1}^{\mathrm{H}_i} c^{(t)}(h|F, m) = 1. \quad (6)$$

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

$$p(x|\theta^{(t)}) = \int_{\mathrm{Fl}_i}^{\mathrm{Fh}_i} \sum_{m=1}^{\mathrm{M}_i} w^{(t)}(F, m)\ p(x|F, m, \mu^{(t)}(F, m))\ dF, \quad (7)$$

$$\theta^{(t)} = \{w^{(t)}, \mu^{(t)}\}, \quad (8)$$

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

$$\mu^{(t)} = \{\mu^{(t)}(F, m) \mid \mathrm{Fl}_i \leq F \leq \mathrm{Fh}_i, m = 1, ..., \mathrm{M}_i\}, \quad (10)$$

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 $p(x|F, m, \mu^{(t)}(F, m))$ which satisfies

$$\int_{\mathrm{Fl}_i}^{\mathrm{Fh}_i} \sum_{m=1}^{\mathrm{M}_i} w^{(t)}(F, m)\ dF = 1. \quad (11)$$

Because we cannot know *a priori* the number of sound sources, it is important that we simultaneously take into consideration all the possibilities of the F0 as expressed in Equation (7). If we can estimate the model parameter $\theta^{(t)}$ such that the observed PDF $p^{(t)}_{\Psi}(x)$ 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 $p^{(t)}_{F0}(F)$ because $w^{(t)}(F, m)$ represents the relative dominance of the harmonic structure:

$$p^{(t)}_{F0}(F) = \sum_{m=1}^{\mathrm{M}_i} w^{(t)}(F, m) \quad (\mathrm{Fl}_i \leq F \leq \mathrm{Fh}_i). \quad (12)$$

3.2. Introducing a prior distribution

For *[Extension 3]*, we define a prior distribution $p_{0i}(\theta^{(t)})$ of the model parameter $\theta^{(t)}$ as follows:

$$p_{0i}(\theta^{(t)}) = p_{0i}(w^{(t)})\, p_{0i}(\mu^{(t)}), \tag{13}$$

$$p_{0i}(w^{(t)}) = \frac{1}{Z_w} e^{-\beta_{wi}^{(t)}\, D_w(w_{0i}^{(t)}; w^{(t)})}, \tag{14}$$

$$p_{0i}(\mu^{(t)}) = \frac{1}{Z_\mu} e^{-\int_{\mathrm{Fl}_i}^{\mathrm{Fh}_i} \sum_{m=1}^{\mathrm{M}_i} \beta_{\mu i}^{(t)}(F,m)\, D_\mu(\mu_{0i}^{(t)}(F,m); \mu^{(t)}(F,m))\, dF}. \tag{15}$$

Here, $p_{0i}(w^{(t)})$ and $p_{0i}(\mu^{(t)})$ are unimodal distributions; $p_{0i}(w^{(t)})$ takes the maximum value at $w_{0i}^{(t)}(F,m)$ and $p_{0i}(\mu^{(t)})$ takes the 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 the normalization factors, and $\beta_{wi}^{(t)}$ and $\beta_{\mu i}^{(t)}(F,m)$ are the parameters determining how much emphasis is put on the maximum value; $\beta_{wi}^{(t)} = 0$ and $\beta_{\mu i}^{(t)}(F,m) = 0$ represent the noninformative prior distribution (uniform distribution). In Equations (14) and (15), $D_w(w_{0i}^{(t)}; w^{(t)})$ and $D_\mu(\mu_{0i}^{(t)}(F,m); \mu^{(t)}(F,m))$ are defined as the following Kullback-Leibler's information:

$$D_w(w_{0i}^{(t)}; w^{(t)}) = \int_{\mathrm{Fl}_i}^{\mathrm{Fh}_i} \sum_{m=1}^{\mathrm{M}_i} w_{0i}^{(t)}(F,m)\, log \frac{w_{0i}^{(t)}(F,m)}{w^{(t)}(F,m)}\, dF, \tag{16}$$

$$D_\mu(\mu_{0i}^{(t)}(F,m); \mu^{(t)}(F,m)) = \sum_{h=1}^{\mathrm{H}_i} c_{0i}^{(t)}(h|F,m)\, log \frac{c_{0i}^{(t)}(h|F,m)}{c^{(t)}(h|F,m)}. \tag{17}$$

3.3. MAP estimation using EM algorithm

The problem to be solved is to estimate the model parameter $\theta^{(t)}$ on the basis of the prior distribution $p_{0i}(\theta^{(t)})$ when we observe $p_\Psi^{(t)}(x)$. The MAP (Maximum *A Posteriori* Probability) estimator of $\theta^{(t)}$ is obtained by maximizing

$$\int_{-\infty}^{\infty} p_\Psi^{(t)}(x)\, (\log p(x|\theta^{(t)}) + \log p_{0i}(\theta^{(t)}))\, dx. \tag{18}$$

Because this maximization problem is too difficult to be solved analytically, we use the EM algorithm to estimate $\theta^{(t)}$. While the EM algorithm is usually used for computing maximum likelihood estimates from incomplete observed data, it can also be used for computing MAP estimates as described in [11]. In the maximum likelihood estimation, the EM algorithm iteratively applies two steps, the *expectation step (E-step)* computing the conditional expectation of the mean log-likelihood and the *maximization step (M-step)* maximizing its expectation. On the other hand, in the MAP estimation, it iteratively applies the E-step computing the sum of the conditional expectation and the log prior distribution and the M-step maximizing it. With respect to $\theta^{(t)}$, each iteration updates the 'old' estimate $\theta'^{(t)} = \{w'^{(t)}, \mu'^{(t)}\}$ to obtain the 'new' improved estimate $\overline{\theta^{(t)}} = \{\overline{w^{(t)}}, \overline{\mu^{(t)}}\}$.

By introducing 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:
   $$Q_{\mathrm{MAP}}(\theta^{(t)}|\theta'^{(t)}) = Q(\theta^{(t)}|\theta'^{(t)}) + \log p_{0i}(\theta^{(t)}), \tag{19}$$
   $$Q(\theta^{(t)}|\theta'^{(t)}) = \int_{-\infty}^{\infty} p_\Psi^{(t)}(x)\, \mathrm{E}_{F,m,h}[\log p(x,F,m,h|\theta^{(t)}) \mid x, \theta'^{(t)}]\, dx, \tag{20}$$
   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 $\theta^{(t)}$ in order to obtain the updated (improved) estimate $\overline{\theta^{(t)}}$:
   $$\overline{\theta^{(t)}} = \operatorname*{argmax}_{\theta^{(t)}} Q_{\mathrm{MAP}}(\theta^{(t)}|\theta'^{(t)}). \tag{21}$$
   In the E-step, $Q(\theta^{(t)}|\theta'^{(t)})$ is expressed as
   $$Q(\theta^{(t)}|\theta'^{(t)}) = \int_{-\infty}^{\infty} \int_{\mathrm{Fl}_i}^{\mathrm{Fh}_i} \sum_{m=1}^{\mathrm{M}_i} \sum_{h=1}^{\mathrm{H}_i} p_\Psi^{(t)}(x)\, p(F,m,h|x,\theta'^{(t)}) \log p(x,F,m,h|\theta^{(t)}) dF dx, \tag{22}$$
   where the complete-data log-likelihood is given by
   $$\log p(x,F,m,h|\theta^{(t)}) = \log(w^{(t)}(F,m)\, p(x,h|F,m,\mu^{(t)}(F,m))). \tag{23}$$
   From Equation (13) the log prior distribution is given by
   $$\log p_{0i}(\theta^{(t)}) = -\log Z_w Z_\mu - \int_{\mathrm{Fl}_i}^{\mathrm{Fh}_i} \sum_{m=1}^{\mathrm{M}_i} \left(\beta_{wi}^{(t)} w_{0i}^{(t)}(F,m)\, \log \frac{w_{0i}^{(t)}(F,m)}{w^{(t)}(F,m)} + \beta_{\mu i}^{(t)}(F,m) \sum_{h=1}^{\mathrm{H}_i} c_{0i}^{(t)}(h|F,m)\, \log \frac{c_{0i}^{(t)}(h|F,m)}{c^{(t)}(h|F,m)} \right) dF. \tag{24}$$

   Regarding the M-step, Equation (21) is a conditional problem of variation, where the conditions are given by Equations (6) and (11). This problem can be solved by using the following Euler-Lagrange differential equations with Lagrange multipliers λ_w and λ_μ:
   $$\frac{\partial}{\partial w^{(t)}} \left(\int_{-\infty}^{\infty} \sum_{h=1}^{\mathrm{H}_i} p_\Psi^{(t)}(x)\, p(F,m,h|x,\theta'^{(t)}) (\log w^{(t)}(F,m) + \log p(x,h|F,m,\mu^{(t)}(F,m)))\, dx - \beta_{wi}^{(t)}\, w_{0i}^{(t)}(F,m)\, \log \frac{w_{0i}^{(t)}(F,m)}{w^{(t)}(F,m)} - \lambda_w (w^{(t)}(F,m) - \tfrac{1}{\mathrm{M}_i(\mathrm{Fh}_i - \mathrm{Fl}_i)}) \right) = 0, \tag{25}$$
   $$\frac{\partial}{\partial c^{(t)}} \left(\int_{-\infty}^{\infty} p_\Psi^{(t)}(x)\, p(F,m,h|x,\theta'^{(t)}) (\log w^{(t)}(F,m) + \log c^{(t)}(h|F,m) + \log G(x; F + 1200 \log_2 h, \mathrm{W}_i))\, dx - \beta_{\mu i}^{(t)}(F,m)\, c_{0i}^{(t)}(h|F,m)\, \log \frac{c_{0i}^{(t)}(h|F,m)}{c^{(t)}(h|F,m)} - \lambda_\mu (c^{(t)}(h|F,m) - \tfrac{1}{\mathrm{H}_i}) \right) = 0. \tag{26}$$
   From these equations we get
   $$w^{(t)}(F,m) = \frac{1}{\lambda_w} \left(\int_{-\infty}^{\infty} p_\Psi^{(t)}(x)\, p(F,m|x,\theta'^{(t)})\, dx + \beta_{wi}^{(t)} w_{0i}^{(t)}(F,m) \right), \tag{27}$$

$$c^{(t)}(h|F,m) = \frac{1}{\lambda_\mu}\left(\int_{-\infty}^{\infty} p_\Psi^{(t)}(x) p(F,m,h|x,\theta'^{(t)})\,dx + \beta_{\mu i}^{(t)}(F,m) c_{0i}^{(t)}(h|F,m)\right). \quad (28)$$

In these equations, λ_w and λ_μ are determined from Equations (6) and (11) as

$$\lambda_w = 1 + \beta_{wi}^{(t)}, \quad (29)$$

$$\lambda_\mu = \int_{-\infty}^{\infty} p_\Psi^{(t)}(x) p(F,m|x,\theta'^{(t)})\,dx + \beta_{\mu i}^{(t)}(F,m). \quad (30)$$

From the Bayes' theorem, $p(F,m,h|x,\theta'^{(t)})$ and $p(F,m|x,\theta'^{(t)})$ are given by

$$p(F,m,h|x,\theta'^{(t)}) = \frac{w'^{(t)}(F,m)\; p(x,h|F,m,\mu'^{(t)}(F,m))}{p(x|\theta'^{(t)})}, \quad (31)$$

$$p(F,m|x,\theta'^{(t)}) = \frac{w'^{(t)}(F,m)\; p(x|F,m,\mu'^{(t)}(F,m))}{p(x|\theta'^{(t)})}. \quad (32)$$

Finally we obtain the 'new' parameter estimates $\overline{w^{(t)}(F,m)}$ and $\overline{c^{(t)}(h|F,m)}$:

$$\overline{w^{(t)}(F,m)} = \frac{\overline{w_{\mathrm{ML}}^{(t)}(F,m)} + \beta_{wi}^{(t)} w_{0i}^{(t)}(F,m)}{1+\beta_{wi}^{(t)}}, \quad (33)$$

$$\overline{c^{(t)}(h|F,m)} = \frac{\overline{w_{\mathrm{ML}}^{(t)}(F,m)}\;\overline{c_{\mathrm{ML}}^{(t)}(h|F,m)} + \beta_{\mu i}^{(t)}(F,m) c_{0i}^{(t)}(h|F,m)}{\overline{w_{\mathrm{ML}}^{(t)}(F,m)} + \beta_{\mu i}^{(t)}(F,m)}, \quad (34)$$

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

$$\overline{w_{\mathrm{ML}}^{(t)}(F,m)} = \int_{-\infty}^{\infty} p_\Psi^{(t)}(x) \frac{w'^{(t)}(F,m)\; p(x|F,m,\mu'^{(t)}(F,m))}{\int_{\mathrm{Fl}_i}^{\mathrm{Fh}_i}\sum_{\nu=1}^{\mathrm{M}_i} w'^{(t)}(\eta,\nu)\; p(x|\eta,\nu,\mu'^{(t)}(F,\nu))\,d\eta}\,dx, \quad (35)$$

$$\overline{c_{\mathrm{ML}}^{(t)}(h|F,m)} = \frac{1}{\overline{w_{\mathrm{ML}}^{(t)}(F,m)}}\int_{-\infty}^{\infty} p_\Psi^{(t)}(x) \frac{w'^{(t)}(F,m)\; p(x,h|F,m,\mu'^{(t)}(F,m))}{\int_{\mathrm{Fl}_i}^{\mathrm{Fh}_i}\sum_{\nu=1}^{\mathrm{M}_i} w'^{(t)}(\eta,\nu)\; p(x|\eta,\nu,\mu'^{(t)}(F,\nu))\,d\eta}\,dx. \quad (36)$$

After the above iterative computation, the F0's PDF $p_{F0}^{(t)}(F)$ estimated by considering the prior distribution can be obtained from $w^{(t)}(F,m)$ according to Equation (12). We can also obtain $c^{(t)}(h|F,m)$, which is the relative amplitude of each harmonic component of all the multiple tone models $p(x|F,m,\mu^{(t)}(F,m))$. All three extensions are thus fulfilled.

4. EXPERIMENTAL RESULTS

The extended 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 for auralization and computer graphics for visualization [9, 10]. The current implementation uses the following parameter values with two adaptive tone models: $\mathrm{Fh_m}$ = 8400 cent, $\mathrm{Fl_m}$ = 3600 cent, $\mathrm{M_m}$ = 2, $\mathrm{H_m}$ = 16, $\mathrm{W_m}$ = 17 cent, $\mathrm{Fh_b}$ = 4800 cent, $\mathrm{Fl_b}$ = 1000 cent, $\mathrm{M_b}$ = 2, $\mathrm{H_b}$ = 6, and $\mathrm{W_b}$ = 17 cent. For the prior distribution of the shape of tone models, we use $c_{0i}^{(t)}(h|F,m) = \alpha_{i,m}\; g_{m,h}\; G(h;1,\mathrm{U}_i)$, where m is 1 or 2, $\alpha_{i,m}$ is a normalization factor, $g_{m,h}$ is 2/3 (m = 2 and h is even) or 1 (otherwise), $\mathrm{U_m}$ = 5.5, and $\mathrm{U_b}$ = 2.7.

The system was tested on excerpts of 10 songs in popular, jazz, and orchestral genres. The 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 by using the F0 editor program we previously developed [9].

In our experiment the system correctly detected the melody and bass lines for most of each audio sample; the average detection rate for the melody line was 88.4% and that for the bass line was 79.9%. The results of comparing the extended PreFEst (with the two adaptive tone models) with the previous PreFEst without any extension showed that the detection rates for three songs were greatly improved (at most 29.2% improvement).

5. CONCLUSION

We have described how to improve a method called PreFEst that detects the melody and bass lines in complex real-world audio signals. The MAP estimation executed by using the EM algorithm enabled three extensions: introducing multiple tone models, estimating the shape of tone models, and introducing a prior distribution of its shape and F0 estimates. Experimental results showed that a real-time system using the extended PreFEst improved in performance and is robust enough to estimate the F0s of the melody and bass lines in compact-disc recordings.

Although the three extensions we made have great potential, we have not fully exploited them. In the future, for example, a lot of tone models could be prepared by analyzing various kinds of harmonic structure appearing in a music (sound) database. We also plan to report experimental results that have shown the effectiveness of using a prior distribution of F0 estimates.

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

6. REFERENCES

[1] L. R. Rabiner *et al.*, "A comparative performance study of several pitch detection algorithms," *IEEE Trans. on ASSP*, vol. ASSP-24, no. 5, pp. 399–418, 1976.

[2] A. Nehorai and B. Porat, "Adaptive comb filtering for harmonic signal enhancement," *IEEE Trans. on ASSP*, vol. ASSP-34, no. 5, pp. 1124–1138, 1986.

[3] T. Abe *et al.*, "Robust pitch estimation with harmonics enhancement in noisy environments based on instantaneous frequency," in *ICSLP 96*, pp. 1277–1280, 1996.

[4] H. Kawahara *et al.*, "Fixed point analysis of frequency to instantaneous frequency mapping for accurate estimation of F0 and periodicity," in *Eurospeech 99*, pp. 2781–2784, 1999.

[5] C. Chafe and D. Jaffe, "Source separation and note identification in polyphonic music," in *ICASSP 86*, pp. 1289–1292, 1986.

[6] H. Katayose and S. Inokuchi, "The kansei music system," *Computer Music Journal*, vol. 13, no. 4, pp. 72–77, 1989.

[7] G. J. Brown and M. Cooke, "Perceptual grouping of musical sounds: A computational model," *JNMR*, vol. 23, pp. 107–132, 1994.

[8] K. Kashino and H. Murase, "Music recognition using note transition context," in *ICASSP 98*, pp. 3593–3596, 1998.

[9] M. Goto, "A robust predominant-f0 estimation method for real-time detection of melody and bass lines in CD recordings," in *ICASSP 2000*, pp. II–757–760, 2000.

[10] M. 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 CASA*, pp. 31–40, 1999.

[11] A. P. Dempster *et al.*, "Maximum likelihood from incomplete data via the EM algorithm," *J. Roy. Stat. Soc. B*, vol. 39, no. 1, pp. 1–38, 1977.