Proceedings of CRAC-2001 (Workshop on Consistent & Reliable Acoustic Cues for Sound Analysis), September 2001.

A Predominant-F0 Estimation Method for Real-world Musical Audio Signals:
MAP Estimation for Incorporating Prior Knowledge about F0s and Tone Models

Masataka Goto

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

Abstract

In this paper we describe a robust method, called *PreFEst*, for estimating the fundamental frequency (F0) of melody and bass lines in monaural audio signals containing sounds of various instruments. Most previous F0-estimation methods have difficulty dealing with such complex audio signals because they are designed for mixtures of only a few sounds. Without assuming the number of sound sources, PreFEst can obtain the most predominant F0 — corresponding to the melody or bass line — supported by harmonics within an intentionally-limited frequency range. It estimates the relative dominance of every possible F0 (represented as a probability density function of the F0) and the shape of harmonic-structure tone models by using the MAP (Maximum *A Posteriori* Probability) estimation considering their prior distribution. Experimental results showed that a real-time system implementing this method is robust enough to detect the melody and bass lines in compact-disc recordings.

1. Introduction

What cues can we rely on when we want to enable a system to understand music? Most people think that reliable and fundamental information in music is provided by the notes in musical scores, and most previous systems attempting to understand music have dealt with automatic music transcription, which transforms audio signals into a symbolic note-level representation. Such a representation, however, is not a good starting point for understanding musical audio signals. Identifying the names (symbols) of the notes corresponding to the sounds of music is a skill mastered only by trained musicians. Even though many trained musicians use scores for composing and interpreting music, untrained listeners understand music without mentally representing audio signals as musical scores.

We therefore proposed a research approach, *music scene description* [1], intended to obtain more fundamental and reliable descriptions of musical audio signals. We have been working on obtaining intuitive and basic descriptions — such as the melody line, bass line [2, 3], and hierarchical beat structure [4, 5, 6] — without identifying musical notes and without segregating sound sources. This paper focuses on detecting melody and bass lines in compact-disc recordings. These lines are fundamental to the perception of Western music and are useful in various practical applications, such as automatic music indexing for information retrieval.

This paper describes a predominant-F0 (fundamental frequency) estimation method, called *PreFEst* [2], that can detect the melody and bass lines in monaural complex mixtures containing simultaneous sounds of various musical instruments. It has been considered difficult to estimate the F0 in such audio signals 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 [7, 8, 9, 10, 11], however, assumed that the input contained just a single-pitch sound with aperiodic noises. Although several methods for dealing with multiple-pitch mixtures were proposed [12, 13, 14, 15], they assumed the number of simultaneous sounds and had difficulty dealing with compact-disc recordings.

Advantages of our method are that it does not assume the number of sound sources, locally trace frequency components, or even 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, simultaneously taking into consideration all the possibilities of the F0 and treating 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 the harmonic-structure tone models (represented by PDFs) of all possible F0s and then finds the F0 of the maximum-weight model corresponding to the most predominant harmonic structure.

The method has recently been made more adaptive and flexible by the following three extensions [3]: introducing multiple types of harmonic-structure tone models, estimating the shape of tone models, and introducing a prior distribution of the model shapes and F0 estimates. For example, when prior knowledge about very rough F0 estimates of the melody and bass lines is available, it can be used in the estimation process. These extensions were made possible by the MAP (Maximum *A Posteriori* Probability) estimation executed by using the EM (Expectation-Maximization) algorithm.

The following sections first describe the extended PreFEst in detail and then present experimental results showing that a real-time system based on the PreFEst can detect the melody and bass lines and that the use of a prior distribution of F0 estimates enables the F0 to be determined more accurately.

2. Predominant-F0 Estimation Method: PreFEst

PreFEst consists of three components, *PreFEst-front-end* for frequency analysis, *PreFEst-core* estimating the most predominant F0, and *PreFEst-back-end* considering temporal continuity of the F0. Since the melody line tends to have 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-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 (Figure 1) in order to obtain adequate time and fre-

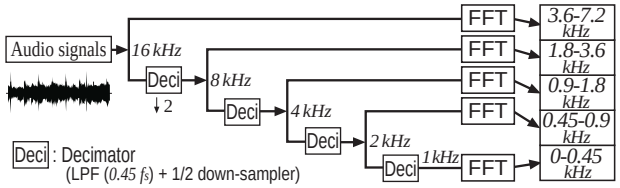

Figure 1: Structure of the multirate filter bank: The cut-off frequency of the anti-aliasing filter (FIR LPF) in each decimator is 0.45 f_s, where f_s is the sampling rate at that branch. The input signal is digitized at 16 bit / 16 kHz and is finally down-sampled to 1 kHz.

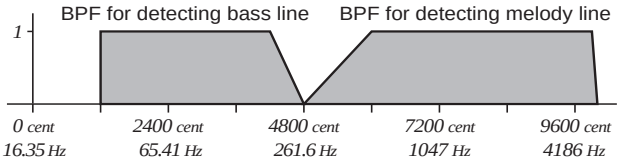

Figure 2: Frequency responses of bandpass filters (BPFs).

quency resolution under the constraint of real-time operation. It then extracts frequency components by using an instantaneous-frequency-related measure [1, 2] and obtains two sets of the input bandpass-filtered frequency components (Figure 2), one for the melody line and the other for the bass line. To use statistical methods, we represent each of the bandpass-filtered frequency components 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 msec), and x is 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:

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

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 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, we can regard those weights as the F0's PDF: the more dominant a tone model 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 harmonic structure, the PreFEst-core can use multiple 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 3) 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

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

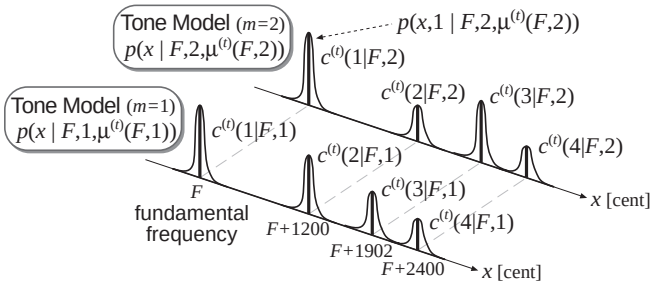

Figure 3: Model parameters of multiple adaptive tone models.

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

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

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

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

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

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

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

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

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 $p(x|F, m, \mu^{(t)}(F, m))$ that 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 $p_\Psi^{(t)}(x)$ is likely to have been generated from the model $p(x|\theta^{(t)})$, $w^{(t)}(F, m)$ can be interpreted as the F0's PDF $p_{F0}^{(t)}(F)$:

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

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:

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

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

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

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 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. 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 (14) and (15), $D_w(w_{0i}^{(t)}; w^{(t)})$ and $D_\mu(\mu_{0i}^{(t)}(F,m); \mu^{(t)}(F,m))$ are the following Kullback-Leibler information:

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

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

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 (Maximum *A Posteriori* Probability) estimator of $\theta^{(t)}$ is obtained by maximizing

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

Because this maximization problem is too difficult to be solved analytically, we use the Expectation-Maximization (EM) algorithm [16], which is an iterative algorithm successively 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 $\overline{\theta^{(t)}} = \{\overline{w^{(t)}}, \overline{\mu^{(t)}}\}$.

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:
   $$Q_{\mathrm{MAP}}(\theta^{(t)}|\theta'^{(t)}) = Q(\theta^{(t)}|\theta'^{(t)}) + \log p_{0i}(\theta^{(t)}), \quad (19)$$
   $$Q(\theta^{(t)}|\theta'^{(t)}) = \int_{-\infty}^{\infty} p_\Psi^{(t)}(x)\, \mathrm{E}_{F,m,h}[\log p(x,F,m,h|\theta^{(t)}) \mid x, \theta'^{(t)}]\, dx, \quad (20)$$
   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 $\overline{\theta^{(t)}}$:
   $$\overline{\theta^{(t)}} = \operatorname*{argmax}_{\theta^{(t)}} Q_{\mathrm{MAP}}(\theta^{(t)}|\theta'^{(t)}). \quad (21)$$

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

$$Q(\theta^{(t)}|\theta'^{(t)}) = \int_{-\infty}^{\infty} \int_{\mathrm{Fl}_i}^{\mathrm{Fh}_i} \sum_{m=1}^{\mathrm{M}_i} \sum_{h=1}^{\mathrm{H}_i} p_\Psi^{(t)}(x)\, p(F,m,h|x,\theta'^{(t)}) \log p(x,F,m,h|\theta^{(t)}) dF dx, \quad (22)$$

where the complete-data log-likelihood is given by

$$\log p(x,F,m,h|\theta^{(t)}) = \log(w^{(t)}(F,m)\, p(x,h|F,m,\mu^{(t)}(F,m))). \quad (23)$$

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:

$$\frac{\partial}{\partial w^{(t)}} \left(\int_{-\infty}^{\infty} \sum_{h=1}^{\mathrm{H}_i} p_\Psi^{(t)}(x)\, p(F,m,h|x,\theta'^{(t)}) (\log w^{(t)}(F,m) + \log p(x,h|F,m,\mu^{(t)}(F,m)))\, dx - \beta_{wi}^{(t)}\, w_{0i}^{(t)}(F,m)\, \log \frac{w_{0i}^{(t)}(F,m)}{w^{(t)}(F,m)} - \lambda_w (w^{(t)}(F,m) - \tfrac{1}{\mathrm{M}_i(\mathrm{Fh}_i - \mathrm{Fl}_i)}) \right) = 0, \quad (24)$$

$$\frac{\partial}{\partial c^{(t)}} \left(\int_{-\infty}^{\infty} p_\Psi^{(t)}(x)\, p(F,m,h|x,\theta'^{(t)})\, (\log w^{(t)}(F,m) + \log c^{(t)}(h|F,m) + \log G(x; F + 1200 \log_2 h, \mathrm{W}_i))\, dx - \beta_{\mu i}^{(t)}(F,m)\, c_{0i}^{(t)}(h|F,m)\, \log \frac{c_{0i}^{(t)}(h|F,m)}{c^{(t)}(h|F,m)} - \lambda_\mu (c^{(t)}(h|F,m) - \tfrac{1}{\mathrm{H}_i}) \right) = 0. \quad (25)$$

From these equations we get

$$w^{(t)}(F,m) = \frac{\int_{-\infty}^{\infty} p_\Psi^{(t)}(x)\, p(F,m|x,\theta'^{(t)})\, dx + \beta_{wi}^{(t)} w_{0i}^{(t)}(F,m)}{1 + \beta_{wi}^{(t)}}, \quad (26)$$

$$c^{(t)}(h|F,m) = \frac{\int_{-\infty}^{\infty} p_\Psi^{(t)}(x) p(F,m,h|x,\theta'^{(t)})\, dx + \beta_{\mu i}^{(t)}(F,m) c_{0i}^{(t)}(h|F,m)}{\int_{-\infty}^{\infty} p_\Psi^{(t)}(x) p(F,m|x,\theta'^{(t)})\, dx + \beta_{\mu i}^{(t)}(F,m)}. \quad (27)$$

According to the Bayes' theorem, $p(F,m,h|x,\theta'^{(t)})$ is given by

$$p(F,m,h|x,\theta'^{(t)}) = \frac{w'^{(t)}(F,m)\, p(x,h|F,m,\mu'^{(t)}(F,m))}{p(x|\theta'^{(t)})}. \quad (28)$$

Finally we obtain the following new parameter estimates:

$$\overline{w^{(t)}(F,m)} = \frac{\overline{w_{\mathrm{ML}}^{(t)}(F,m)} + \beta_{wi}^{(t)} w_{0i}^{(t)}(F,m)}{1 + \beta_{wi}^{(t)}}, \quad (29)$$

$$\overline{c^{(t)}(h|F,m)} = \frac{\overline{w_{\mathrm{ML}}^{(t)}(F,m)}\ \overline{c_{\mathrm{ML}}^{(t)}(h|F,m)} + \beta_{\mu i}^{(t)}(F,m) c_{0i}^{(t)}(h|F,m)}{\overline{w_{\mathrm{ML}}^{(t)}(F,m)} + \beta_{\mu i}^{(t)}(F,m)}, \quad (30)$$

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:

$$\overline{w_{\mathrm{ML}}^{(t)}(F,m)} = \int_{-\infty}^{\infty} p_\Psi^{(t)}(x) \frac{w'^{(t)}(F,m)\, p(x|F,m,\mu'^{(t)}(F,m))}{\int_{\mathrm{Fl}_i}^{\mathrm{Fh}_i} \sum_{\nu=1}^{\mathrm{M}_i} w'^{(t)}(\eta,\nu)\, p(x|\eta,\nu,\mu'^{(t)}(F,\nu))\, d\eta}\, dx, \quad (31)$$

$$\overline{c^{(t)}_{\mathrm{ML}}(h|F,m)} = \frac{1}{\overline{w^{(t)}_{\mathrm{ML}}(F,m)}} \int_{-\infty}^{\infty} p^{(t)}_{\Psi}(x) \frac{w'^{(t)}(F,m)\; p(x,h|F,m,\mu'^{(t)}(F,m))}{\int_{\mathrm{Fl}_i}^{\mathrm{Fh}_i} \sum_{\nu=1}^{\mathrm{M}_i} w'^{(t)}(\eta,\nu)\; p(x|\eta,\nu,\mu'^{(t)}(F,\nu))\; d\eta}\, dx. \quad (32)$$

After the above iterative computation, the F0's PDF $p^{(t)}_{F0}(F)$ estimated by considering the prior distribution can be obtained from $w^{(t)}(F,m)$ according to Equation (12). We can also obtain the tone-model shape $c^{(t)}(h|F,m)$, which is the relative amplitude of each harmonic component of all the 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 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 [1, 2] 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 [1, 2]. The current implementation uses two adaptive tone models with the following parameter values: $\mathrm{Fh_m}$ = 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 $c^{(t)}_{0i}(h|F,m) = \alpha_{i,m}\; g_{m,h}\; G(h;1,\mathrm{U}_i)$, where m is 1 or 2, $\alpha_{i,m}$ is a normalization factor, $g_{m,h}$ is 2/3 (when 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 from a total of 10 songs in the 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 hand-labeled using the F0 editor program we previously developed [2]. In our experiment the system correctly detected the melody and bass lines for most of each audio sample: the average detection rate was 88.4% for the melody line and 79.9% for the bass line.

We also tested the system by providing prior knowledge about rough F0 estimates of the melody line. For the prior F0 distribution, we used $w^{(t)}_{0i}(F,m) = G(F;F^{(t)}_{0i},100\ \mathrm{cent})/\mathrm{M}_i$, where $F^{(t)}_{0i}$ is the F0 estimate given by playing a MIDI keyboard while listening to each excerpt of the songs. Comparing the results obtained with and without using the prior F0 distribution showed that the use of prior knowledge improved the average detection rate for the melody line (from 88.4% to 91.2%). In particular, the detection rate for an orchestral song was greatly improved (11.9% improvement).

4. Conclusion

We have described a method, called PreFEst, that estimates the most predominant F0 in a monaural complex sound mixture without assuming the number of sound sources. Its MAP estimation executed by using the EM algorithm makes it possible to estimate the F0's PDF and the shape of tone models while considering their prior distribution. Experimental results showed that a system implementing this method is robust enough to estimate the F0s of the melody and bass lines in compact-disc recordings in real time.

Although the PreFEst has great potential, we have not fully exploited it. In the future, for example, a lot of tone models could be prepared by analyzing various kinds of harmonic structure appearing 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. The PreFEst can also be applied to non-music audio signals. In fact, Masuda-Katsuse [17, 18] has extended it and shown that it is effective for speech recognition in realistic noisy environments.

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

5. References

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

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

[3] M. Goto, "A predominant-F0 estimation method for CD recordings: MAP estimation using EM algorithm for adaptive tone models," *Proc. ICASSP 2001*, 2001.

[4] M. Goto and Y. Muraoka, "A beat tracking system for acoustic signals of music," *Proc. the Second ACM Intl. Conf. on Multimedia*, pp. 365–372, 1994.

[5] M. Goto and Y. Muraoka, "Real-time beat tracking for drumless audio signals: Chord change detection for musical decisions," *Speech Communication*, vol. 27, nos. 3–4, pp. 311–335, 1999.

[6] M. Goto, "An audio-based real-time beat tracking system for music with or without drum-sounds," *Journal of New Music Research*, vol. 30, 2001. (in press)

[7] 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.

[8] 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.

[9] F. J. Charpentier, "Pitch detection using the short-term phase spectrum," *Proc. ICASSP 86*, pp. 113–116, 1986.

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

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

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

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

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

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

[16] 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.

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

[18] I. Masuda-Katsuse and Y. Sugano, "Speech estimation biased by phonemic expectation in the presence of non-stationary and unpredictable noise," *Proc. CRAC workshop*, 2001.