PreFEst: A Predominant-F0 Estimation Method for Polyphonic Musical Audio Signals

Masataka Goto
National Institute of Advanced Industrial Science and Technology (AIST).
IT, AIST, 1-1-1 Umezono, Tsukuba, Ibaraki 305-8568, Japan
m.goto@aist.go.jp

ABSTRACT

This paper describes a real-time method, called *PreFEst (Predominant-F0 Estimation method)*, for estimating the fundamental frequency (F0) of simultaneous sounds in monaural polyphonic audio signals. 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 (Maximum *A Posteriori* Probability) 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.

1 Introduction

This paper introduces a Predominant-F0 Estimation method, called *PreFEst*, that I developed during 1998 and 2000 [1, 2, 3, 4]. 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.

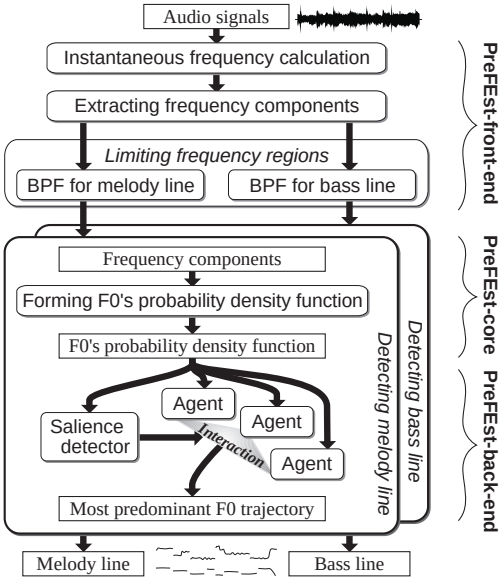

Figure 1: Overview of PreFEst.

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 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 [1, 4] and obtains two sets of bandpass-filtered frequency components, one for the melody line and the other for the bass line. 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. The PreFEst-core considers 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 $p(x|F, m, \mu^{(t)}(F, m))$, where the model parameter $\mu^{(t)}(F, m)$ represents the shape of the tone model. The number of tone models is M_i ($m = 1, \ldots, \mathrm{M}_i$) where i denotes the melody line ($i = \mathrm{m}$) or the bass line ($i = \mathrm{b}$). Each tone model is defined by

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

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

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

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

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.

The PreFEst-core then considers 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 the number of sound sources cannot be known *a priori*, it is important to simultaneously take into consideration all F0 possibilities as expressed in Equation (6). If it is possible to 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, a prior distribution $p_{0i}(\theta^{(t)})$ of $\theta^{(t)}$ is defined 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:

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

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

[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: $f_{\mathrm{cent}} = 1200 \log_2(f_{\mathrm{Hz}} \,/\, (440 \times 2^{\frac{3}{12}-5}))$.

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 $p_\Psi^{(t)}(x)$ is observed. 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, the PreFEst-core uses the Expectation-Maximization (EM) algorithm, 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, the two steps can be specified 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 [3, 4] and the following new parameter estimates are obtained:

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

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

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

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

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

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

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.

The PreFEst-back-end therefore considers the global temporal continuity of the F0 by using a multiple-agent architecture [1] in which agents track different temporal trajectories of the F0. Each agent starts tracking from each salient peak in the F0's PDF, keeps tracking as long as it is temporally continued, and stops tracking when its next peak cannot be found for a while. The final F0 output is determined on the basis of the most dominant and stable F0 trajectory.

3 MIREX 2005 Evaluation Results

I participated in the 2005 MIREX Contest "Audio Melody Extraction". The goal of the contest is to compare the accuracy of state-of-the-art methods of extracting melodic content from polyphonic audio. The data set used for the evaluation contains 25 phrase excerpts of 10-40 sec from the following genres: Rock, R&B, Pop, Jazz, Solo classical piano. The task consists of two parts:

Table 1: Evaluation results for voiced (pitched) time frame only (sorted by *D. Raw Pitch Accuracy*).

Rank	Participant	*D. Raw Pitch Acc.*	*E. Raw Chroma Acc.*
1.	Ryynanen & Klapuri	0.6855	0.7409
2.	Dressler	0.6807	0.7142
3.	Poliner & Ellis	0.6725	0.7339
4.	**Goto (PreFEst)**	**0.6583**	**0.7178**
5.	Paiva 1	0.6273	0.6673
6.	Marolt	0.6006	0.6706
7.	Vincent & Plumbley 1	0.5985	0.6758
8.	Vincent & Plumbley 2	0.5958	0.7114
9.	Paiva 2	0.5854	0.6197
10.	Brossier	0.0393	0.0809

These results are excerpts of the "2005 MIREX Contest Results - Audio Melody Extraction". Note that the original results are sorted by using *F. Overall Accuracy* (proportion of frames with voicing and pitch correct). Goto, Vincent & Plumbley, and Brossier did not perform voiced/unvoiced detection. Scores for Brossier are artificially low due to an unresolved algorithmic issue.

1. voiced/unvoiced detection (deciding whether a particular time frame contains a "melody pitch" or not)

2. pitch (F0) estimation (deciding the most likely melody pitch for each time frame)

In evaluating the F0 estimation, the estimated frequency is scored as a successful estimation within 1/4 tone of the reference frequency for time frames in which the predominant melody is present.

The results of the F0 estimation — *D. Raw Pitch Accuracy* (pitch accuracy on pitched frames ignoring voicing) and *E. Raw Chroma Accuracy* (pitch accuracy on pitched frames ignoring voicing and octave errors) — are shown in Table 1. The methods are sorted by the Raw Pitch Accuracy in this table. Because the PreFEst method does not have a function of discriminating voice and unvoiced portions, the results affected by the voiced/unvoiced detection (*A. Voicing Detection*, *B. Voicing False Alarm*, *C. Voicing d-prime*, and *F. Overall Accuracy*) are not meaningful for the evaluation of PreFEst and omitted in the table.

Although PreFEst can be considered a classic baseline method, the performance differences from the top-ranked scores were 2.72% for the Raw Pitch Accuracy and 2.31% for the Raw Chroma Accuracy. Despite the fact that the PreFEst method was developed during 1998 and 2000 and has not been refined afterwards by the author, it is still competitive in estimating the melody line and can be considered a promising approach with an interesting potential. Further investigation of the performance differences will also be interesting.

The shorter version of this paper was reviewed before the evaluation. The comments of the reviewer (Dan Ellis) regarding this paper were as follows:

> *This approach is the now-classic PreFEst that was used by others last year. The interesting thing about the algorithm is that it defines a stochastic parametric model for the observed set of spectral peaks, then solves for the maximum-likelihood parameterization of this model to match the observations. This allows the system to resolve ambiguous analyses (since although there may be multiple candidate explanations, they will fit with different likelihoods) and provides a nice way to incorporate prior information.*
>
> *Per-frame fundamental estimates are assembled into continuous streams with an agent-based architecture, which in principle could also implement some likelihood-maximizing criterion, although that is not described in the abstract.*
>
> *The theory behind this approach is considerably more sophisticated than most other approaches; it would be interesting to know how much more computationally intensive this makes the algorithm.*

In response to the last part, the implementation of the PreFEst-based melody estimation system was executed in real time even in 1999. The latency to the audio input was 303 ms. It can be executed even faster than real time in processing audio files. It took 6.3 sec for 20 sec excerpt (0.315 $times$ real time) on average on Intel Xeon 3.60 GHz CPU. In addition, it has been ported on multiple operating systems, such as Linux, SGI IRIX, and Microsoft Windows XP.

4 Conclusion

This paper has introduced the PreFEst method that estimates the most predominant F0 in a monaural sound mixture without assuming the number of sound sources. The main contributions of my research on the PreFEst method [1, 2, 3, 4] can be summarized as follows:

- The first predominant-F0 estimation method for estimating melody and bass lines was developed. It was important to show that the melody estimation for CD recordings was actually possible in 1999.

- This research consequently established a new research topic *"predominant-F0 estimation"* by introducing this original term. Note that this is not like a traditional approach of transcribing all musical notes as a score.

- An original framework for multiple F0 estimation by introducing a mixture density model for general sound mixtures was proposed. Model parameters of the powerful harmonic-structure models can be estimated by using MAP estimation solved by the EM algorithm while incorporating their prior information. This is an influential approach followed by other researchers in terms of research theme and probabilistic modeling.

Although PreFEst has great potential, it has not been fully exploited. 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.

References

[1] 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*, pages 31–40, 1999.

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

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

[4] 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*, 43(4):311–329, 2004.