ICASSP 2000 (2000 IEEE International Conference on Acoustics, Speech, and Signal Processing) Proceedings, pp. II-757-760, June 2000.

A ROBUST PREDOMINANT-F0 ESTIMATION METHOD FOR REAL-TIME DETECTION OF MELODY AND BASS LINES IN CD RECORDINGS

Masataka Goto

Electrotechnical Laboratory
1-1-4 Umezono, Tsukuba, Ibaraki 305-8568 JAPAN.
goto@etl.go.jp

ABSTRACT

This paper describes a robust method for estimating the fundamental frequency (F0) of melody and bass lines in monaural real-world musical audio signals containing sounds of various instruments. Most previous F0-estimation methods had great difficulty dealing with such complex audio signals because they were designed to deal with mixtures of only a few sounds. To make it possible to estimate the F0 of the melody and bass lines, we propose a predominant-F0 estimation method called *PreFEst* that does not rely on the F0's unreliable frequency component and obtains the most predominant F0 supported by harmonics within an intentionally limited frequency range. It evaluates the relative dominance of every possible F0 by using the *Expectation-Maximization* algorithm and considers the temporal continuity of F0s by using a multiple-agent architecture. Experimental results show that our real-time system can detect the melody and bass lines in audio signals sampled from commercially distributed compact discs.

1. INTRODUCTION

In order to build a computational model that can understand musical audio signals in a human-like fashion, the detection of melody and bass lines is essential because the melody forms the core of Western music and is very influential in the identity of a musical piece and because the bass is closely related with the tonality. The detected melody and bass lines are also useful in various practical applications, such as automatic music indexing for information retrieval (e.g. submitting a song query by singing a melody), automatic transcription, computer participation in human live performances, analysis of recordings of outstanding performances, and producing accompaniment tracks for *Karaoke* or *Music Minus One* automatically by making use of compact discs.

This detection requires the estimation of the fundamental frequency (F0, perceived as pitch) of the melody and bass lines. It has, however, been considered very difficult to estimate the F0 of a particular instrument or voice in the monaural audio signal of an ensemble performed by more than three musical instruments. Most previous F0-estimation methods [1, 2, 3, 4, 5] premised that the input audio signal contained just a single-pitch sound with aperiodic noises. Although several methods for dealing with multiple-pitch mixtures were proposed [6, 7, 8, 9], they dealt with at most three musical instruments or voices and had great difficulty estimating the F0 in complex audio signals sampled from compact discs. The main reason for this difficulty is that, in the time-frequency domain, the frequency components of one sound often overlap frequency components of simultaneous sounds. In typical popular music, for example, part of the voice's harmonic structure is often overlapped by harmonics of the keyboard instrument or guitar, by higher harmonics of the bass guitar, and by noisy inharmonic frequency components of the snare drum. A simple method locally tracing a frequency component therefore cannot be reliable and stable. Moreover, sophisticated F0 estimation methods relying on the existence of a frequency component corresponding to the F0 not only cannot handle the *missing fundamental* but are also unreliable when the frequency component of the F0 is smeared by the harmonics of simultaneous sounds.

This paper describes a method, called *PreFEst* (Predominant-F0 Estimation Method), that can detect the melody and bass lines in these complex real-world audio signals. Because the PreFEst does not rely on the F0's frequency component that tends to be unreliable and evaluates the relative dominance of every possible F0 by using the *Expectation-Maximization (EM)* algorithm [10] without assuming the number of sound sources, it can estimate the F0 of the most predominant harmonic structure in sound mixtures containing simultaneous sounds of various instruments (even drums). In addition, to obtain a stable estimate in ambiguous situations, it considers the global temporal continuity of the F0 by using a multiple-agent architecture.

The following sections describe the details of the PreFEst and the implementation of a system that can perform the PreFEst calculation in real time. It then shows the results of experiments detecting the melody and bass lines in monaural audio signals of compact disc recordings.

2. PREDOMINANT-F0 ESTIMATION METHOD: PREFEST

The PreFEst obtains the temporal F0 trajectories of the melody and bass lines under the following assumptions that fit a large class of music.

- The melody and bass sounds have the harmonic structure. We do not care about the existence of the F0's frequency component, however.
- The melody line has the most predominant harmonic structure in middle and high frequency regions and the bass line has the most predominant harmonic structure in a low frequency region.
- The melody and bass lines tend to have temporally continuous trajectories.

Figure 1 shows an overview of the PreFEst. It first calculates instantaneous frequencies by using multirate signal processing techniques and extracts candidate frequency components on the basis of an instantaneous-frequency-related measure. The PreFEst basically estimates the F0 which is supported by predominant harmonic frequency components within an intentionally limited frequency range; by using two bandpass filters (BPFs) it limits the frequency range to middle and high regions for the melody line and to a low region for the bass line. It then forms a probabil-

0-7803-6293-4/00/$10.00

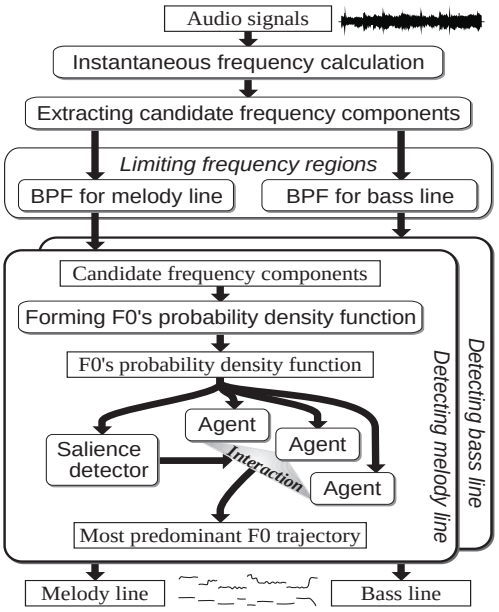

Figure 1: Overview of the *PreFEst*.

ity density function (PDF) of the F0, which represents the relative dominance of every possible harmonic structure. To form this F0's PDF, it regards each set of the filtered frequency components as a weighted mixture of all possible harmonic-structure tone models and estimates their weights that can be interpreted as the F0's PDF: the maximum-weight model corresponds to the most predominant harmonic structure. This estimation is carried out by using the EM algorithm, which is an iterative technique for computing maximum likelihood estimates from incomplete data. Finally, multiple agents track the temporal trajectories of salient promising peaks in the F0's PDF and the output F0 is determined on the basis of the most dominant and stable trajectory.

2.1. Instantaneous Frequency Calculation

The method first calculates the instantaneous frequency of filterbank outputs by using the short-time Fourier transform (STFT) whose output can be interpreted as a collection of uniform-filter outputs. When the STFT of a signal $x(t)$ with a window function $h(t)$ is defined as

$$X(\omega, t) = \int_{-\infty}^{\infty} x(\tau)h(\tau - t)e^{-j\omega\tau}d\tau = a + jb, \quad (1)$$

the instantaneous frequency $\lambda(\omega, t)$ is given by the following equation [11]:

$$\lambda(\omega, t) = \omega + \frac{a\frac{\partial b}{\partial t} - b\frac{\partial a}{\partial t}}{a^2 + b^2}. \quad (2)$$

To obtain an adequate time-frequency resolution under the real-time constraint, we designed an STFT-based multirate filter bank shown in Figure 2. At each level of binary branches, the audio signal is down-sampled by a decimator. 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. In the current implementation, the input signal is digitized at 16 bit / 16 kHz and is finally down-sampled to 1 kHz. Then the STFT whose window size is 512 samples is calculated at each leaf by using the FFT while compensating for time delays of the different multirate layers. Since at 16 kHz the FFT frame is shifted by 160 samples, the discrete time step (1 *frame-time*) is 10 ms. This paper uses time t for the time measured in units of frame-time.

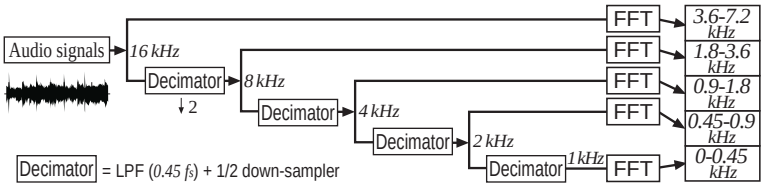

Figure 2: Overview of multirate filter bank.

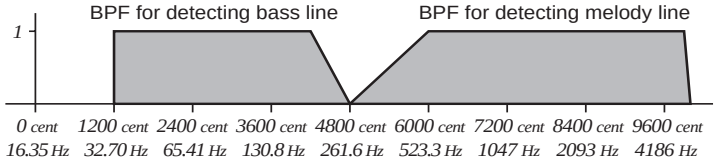

Figure 3: Frequency responses of bandpass filters (BPFs).

2.2. Extracting Candidate Frequency Components

The extraction of candidate frequency components is based on the the mapping from the center frequency ω of an STFT filter to the instantaneous frequency $\lambda(\omega, t)$ of its output [3, 4, 5]. Finding fixed stable points of the mapping, we can extract a set $\Psi_f^{(t)}$ of instantaneous frequencies of the frequency components by using the following equation [12]:

$$\Psi_f^{(t)} = \{\ \psi\ \mid\ \lambda(\psi, t) - \psi = 0,\ \frac{\partial}{\partial\psi}(\lambda(\psi, t) - \psi) < 0\}. \quad (3)$$

By calculating the power of those frequencies, we can define the power distribution function $\Psi_p^{(t)}(\omega)$ as

$$\Psi_p^{(t)}(\omega) = \begin{cases} |\,X(\omega, t)\,| & \text{if } \omega \in \Psi_f^{(t)} \\ 0 & \text{otherwise.} \end{cases} \quad (4)$$

2.3. Limiting Frequency Regions

The frequency range is intentionally limited by using the two BPFs whose frequency responses are shown in Figure 3. The BPF for the melody line is designed so that it covers most dominant harmonics of typical melody lines and deemphasizes a crowded frequency region around the F0: it does not matter if the F0 is not within the passband. The BPF for the bass line is designed so that it covers most dominant harmonics of typical bass lines and deemphasizes a frequency region where other parts tend to become more dominant than the bass line. In this paper the log-scale frequency is denoted in units of *cents* (a musical-interval measurement). A frequency f_{Hz} in hertz is converted to the frequency f_{cent} in cents as follows:

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

There are 100 cents to a tempered semitone and 1200 to an octave.

The filtered frequency components can be represented as $BPF_i(x)\Psi_p'^{(t)}(x)$, where $BPF_i(x)$ (i = m, b) is the BPF's frequency response at frequency x (in cents) for the melody line (i = m) and the bass line (i = b). The power distribution $\Psi_p'^{(t)}(x)$ is the same as $\Psi_p^{(t)}(\omega)$ except that the frequency unit is the cent. The PDF of the filtered frequency components $p_\Psi^{(t)}(x)$ is defined as follows:

$$p_\Psi^{(t)}(x) = \frac{BPF_i(x)\ \Psi_p'^{(t)}(x)}{\int_{-\infty}^{\infty} BPF_i(x)\ \Psi_p'^{(t)}(x)\ dx}. \quad (6)$$

2.4. Forming the F0's Probability Density Function

For each set of the filtered frequency components, the method forms a PDF of the F0. The basic idea is to consider that the observed PDF $p_\Psi^{(t)}(x)$ was generated from a model that is a weighted mixture of harmonic-structure tone models. When the PDF of each tone model whose F0 is frequency F is denoted as $p(x|F)$, the mixture density $p(x;\theta^{(t)})$ is defined as

$$p(x;\theta^{(t)}) = \int_{\mathrm{Fl}_i}^{\mathrm{Fh}_i} w^{(t)}(F)\, p(x|F)\, dF, \tag{7}$$

$$\theta^{(t)} = \{w^{(t)}(F) \mid \mathrm{Fl}_i \le F \le \mathrm{Fh}_i\}, \tag{8}$$

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

$$\int_{\mathrm{Fl}_i}^{\mathrm{Fh}_i} w^{(t)}(F)\, dF = 1. \tag{9}$$

Because we cannot know *a priori* the number of sound sources in real-world audio signals, it is important that we simultaneously take into consideration all the possibilities of the F0 as expressed in the above equations. If we can estimate the model parameter $\theta^{(t)}$ such that $p_\Psi^{(t)}(x)$ is likely to have been generated from $p(x;\theta^{(t)})$, $p_\Psi^{(t)}(x)$ can be considered to be decomposed into harmonic-structure tone models and $w^{(t)}(F)$ can be interpreted as the F0's PDF:

$$p_{F0}^{(t)}(F) = w^{(t)}(F) \quad (\mathrm{Fl}_i \le F \le \mathrm{Fh}_i). \tag{10}$$

The more dominant a tone model $p(x|F)$ in the mixture, the higher the probability of the F0 F of its model.

Therefore the problem to be solved is to estimate the model parameter $\theta^{(t)}$ when we observe $p_\Psi^{(t)}(x)$. The maximum likelihood estimator of $\theta^{(t)}$ is obtained by maximizing the mean log-likelihood defined as $\int_{-\infty}^{\infty} p_\Psi^{(t)}(x)\ \log p(x;\theta^{(t)})\, dx$. Because this maximization problem is too difficult to be solved analytically, the PreFEst uses the *Expectation-Maximization (EM)* algorithm [10], which is an iterative algorithm successively applying two steps — the *expectation step (E-step)* and the *maximization step (M-step)* — to compute maximum likelihood estimates from incomplete observed data (i.e., from $p_\Psi^{(t)}(x)$). With respect to $\theta^{(t)}$, each iteration updates the 'old' estimate $\theta'^{(t)} = \{w'^{(t)}(F)\}$ to obtain the 'new' improved estimate $\overline{\theta^{(t)}} = \{\overline{w^{(t)}(F)}\}$. For the initial estimate of $\theta'^{(t)}$ we simply use the final estimate at $t-1$.

By introducing a hidden (unobservable) variable F describing which tone model was responsible for generating each observed frequency component at x, we can specify the two steps as follows:

1. (E-step)
   Compute the following conditional expectation of the mean log-likelihood:
   $$Q(\theta^{(t)}|\theta'^{(t)}) = \int_{-\infty}^{\infty} p_\Psi^{(t)}(x)\ \mathrm{E}_F[\log p(x, F;\theta^{(t)}) \mid x;\theta'^{(t)}]\, dx, \tag{11}$$
   where $\mathrm{E}_F[a|b]$ denotes the conditional expectation of a with respect to the hidden variable F with the probability distribution determined by the condition b.
2. (M-step)
   Maximize $Q(\theta^{(t)}|\theta'^{(t)})$ as a function of $\theta^{(t)}$ to obtain $\overline{\theta^{(t)}}$:
   $$\overline{\theta^{(t)}} = \mathrm{argmax}_{\theta^{(t)}}\ Q(\theta^{(t)}|\theta'^{(t)}). \tag{12}$$

In the E-step we have

$$Q(\theta^{(t)}|\theta'^{(t)}) = \int_{-\infty}^{\infty}\int_{\mathrm{Fl}_i}^{\mathrm{Fh}_i} p_\Psi^{(t)}(x) p(F|x;\theta'^{(t)}) \log p(x, F;\theta^{(t)}) dF\, dx, \tag{13}$$

where the complete-data log-likelihood is given by

$$\log p(x, F;\theta^{(t)}) = \log(w^{(t)}(F)\ p(x|F)). \tag{14}$$

As for the M-step, Eq. (12) is a conditional problem of variation, where the condition is Eq. (9). This problem can be solved by using the following Euler-Lagrange differential equation:

$$\frac{\partial}{\partial w^{(t)}}\left(\int_{-\infty}^{\infty} p_\Psi^{(t)}(x)\ p(F|x;\theta'^{(t)})\ (\log w^{(t)}(F) + \log p(x|F))\, dx - \lambda(w^{(t)}(F) - \tfrac{1}{\mathrm{Fh}_i - \mathrm{Fl}_i})\right) = 0, \tag{15}$$

where λ is a Lagrange multiplier and is determined from Eq. (9) to be equal to 1. From the Bayes' theorem, $p(F|x;\theta'^{(t)})$ is given by

$$p(F|x;\theta'^{(t)}) = \frac{w'^{(t)}(F)\ p(x|F)}{\int_{\mathrm{Fl}_i}^{\mathrm{Fh}_i} w'^{(t)}(\eta)\ p(x|\eta)\, d\eta}. \tag{16}$$

Finally we obtain the 'new' parameter estimate $\overline{w^{(t)}(F)}$:

$$\overline{w^{(t)}(F)} = \int_{-\infty}^{\infty} p_\Psi^{(t)}(x) \frac{w'^{(t)}(F)\ p(x|F)}{\int_{\mathrm{Fl}_i}^{\mathrm{Fh}_i} w'^{(t)}(\eta)\ p(x|\eta)\, d\eta}\, dx. \tag{17}$$

To compute Eq. (17) we need to assume $p(x|F)$ that indicates where the harmonics of the F0 F tend to occur. The PreFEst assumes the following simple harmonic-structure tone models:

$$p(x|F) = \alpha \sum_{h=1}^{\mathrm{N}_i} c(h)\ G(x;\ F + 1200\log_2 h, \mathrm{W}_i), \tag{18}$$

where α is a normalization factor, N_i is the number of harmonics considered, $c(h)$ determines the amplitude of the h-th harmonic component, and W_i^2 is the variance of the Gaussian distribution $G(x;m,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}}\ e^{-\frac{(x-m)^2}{2\sigma^2}}$. For $c(h)$ it uses $G(h;1,\mathrm{H}_i)$, where H_i is a constant. These models are very simple but work well for the purpose of evaluating the relative dominance of harmonic structure even though they do not coincide exactly with various kinds of harmonic structure contained in real-world audio signals.

A simple way of determining the most predominant F0 is to find the frequency that maximizes the F0's PDF $p_{F0}^{(t)}(F)$, which is the final estimate obtained by the iterative computation of Eq. (17). This result is not stable, however, because peaks corresponding to the F0s of several simultaneous tones sometimes compete in $p_{F0}^{(t)}(F)$ for a moment and are transiently selected, one after another, as the maximum. It is therefore necessary to consider the global temporal continuity of the F0. This is addressed in the next section.

2.5. Sequential F0 Tracking by Multiple-Agent Architecture

To select the F0 trajectory that is most dominant and stable from the viewpoint of global F0 estimation, the method sequentially tracks peak trajectories in the temporal transition of the F0's PDF. To make this possible, we introduce a multiple-agent architecture that enables the tracking process to be controlled dynamically and flexibly. It consists of a salience detector and multiple agents that are dynamically generated and terminated (Figure 4). The salience detector picks up salient promising peaks in the F0's PDF, and agents driven by those peaks track their trajectories. They behave at each frame as follows:

(1) The salience detector picks up salient peaks that are higher than a dynamic threshold adjusted according to the maximum peak. The agents generated interact to allocate the salient peaks among themselves exclusively according to peak closeness. If more than one agent claims the same peak, the peak is allocated to the most reliable agent. If the most salient peak has not been allocated, a new agent for tracking it is generated.

(2) Each agent has an accumulated penalty, and an agent whose penalty exceeds a threshold is terminated. An agent to which a salient peak has not been allocated or which cannot find its next peak in the F0's PDF is penalized. When a peak is allocated to an agent, its penalty is reset.

(3) Each agent evaluates its own reliability by using the reliability at the previous frame and the degree of the peak's salience at the

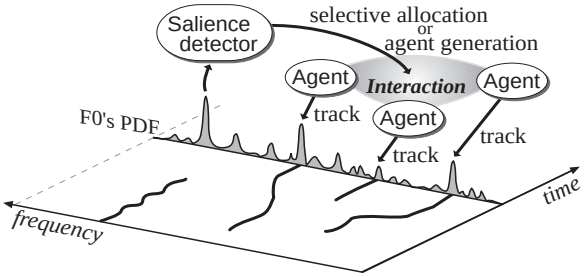

Figure 4: Sequential F0 tracking by multiple-agent architecture.

current frame. The final F0 output is determined on the basis of which agent has the highest reliability and greatest total power along the trajectory of the peak it is tracking.

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: audio signals for auralization, computer graphics for visualization, and continuous values with time stamps for use in applications.[1] The output audio signals are generated by sinusoidal synthesis on the basis of the harmonics tracked along the estimated F0. The audio-synchronized graphics output shows the scrolling candidate frequency components and F0 trajectories on a time-frequency plane.[2] The current implementation uses the following parameter values: Fh_m = 9600 cent, Fl_m = 3600 cent, N_m = 16, W_m = 17 cent, H_m = 5.5, Fh_b = 4800 cent, Fl_b = 1000 cent, N_b = 6, W_b = 17 cent, and H_b = 2.7.

The system was tested on excerpts of 10 songs in popular, jazz, and orchestral genres. The input monaural audio signals were sampled from compact discs and each contained a single-tone melody and the sounds of several instruments. The estimated F0s were compared with the correct F0s that were hand-labeled by using an F0 editor program we developed. This F0 editor program enables a user to determine, at each frame, the correct F0 values of the melody and bass lines while listening to the audio playback of the original and the harmonic structure of the currently-labeled F0 and while watching their frequency components. If the F0 error (frequency difference) of a frame was less than 50 cents, the estimated F0 at that frame was judged to be correct.

The detection rates thus obtained are listed in Table 1. In the absence of the melody or bass line, the system detected the F0 of a dominant accompaniment part because the method simply estimates the predominant F0 trajectory every moment and does not discriminate sound sources. The evaluation was therefore restricted to the periods when the hand-labeled melody or bass line was present. Typical errors were half-F0 or double-F0 errors and errors where a short-term trajectory around the onset was missing.

4. CONCLUSION

We have described a method called PreFEst that detects the melody and bass lines in complex real-world audio signals by estimating the most predominant F0 trajectory. Those lines can be detected separately by using only partial information in intentionally limited frequency ranges. The use of the EM algorithm without assuming the number of sound sources enables the F0 to be estimated without relying on the existence of the F0's frequency component. In addition, the multiple-agent architecture makes it possible to determine the most dominant and stable F0 trajectory from the viewpoint of global temporal continuity of the F0. Experimental results show that a system implementing the PreFEst is robust enough to estimate, in real time, the predominant F0 of the melody and bass lines in audio signals sampled from compact discs.

The F0's PDF estimated by the PreFEst contains the information of every harmonic structure in sound mixtures and has not been fully exploited. We therefore plan to extend the method to track several sound sources simultaneously in the F0's PDF by using a selective-attention mechanism with sound source discrimination.

[1] The main signal processing is performed on a personal computer with two Pentium II 450MHz CPUs, and the audio I/O and visualization processing is performed on an SGI workstation with an R10000 250MHz CPU.

[2] Further information, including screen snapshots, is available at the following URL: http://www.etl.go.jp/~goto/ICASSP2000/

Table 1: Detection rates of the melody and bass lines.

title	genre	detection rates [%]	
		melody	bass
My Heart Will Go On (Celine Dion)	popular	88.7	92.2
Vision of Love (Mariah Carey)	popular	74.5	83.8
Always (Bon Jovi)	popular	92.4	84.5
Time Goes By (Every Little Thing)	popular	89.9	64.7
Spirit of Love (Sing Like Talking)	popular	85.9	80.0
Hoshi no Furu Oka (Misia)	popular	89.1	76.6
Scarborough Fair (Herbie Hancock)	jazz	93.6	53.4
Autumn Leaves ("Cannonball" Adderley)	jazz	81.2	86.2
On Green Dolphin Street (Miles Davis)	jazz	90.8	54.3
Violin Con. in D, Op. 35 (Tchaikovsky)	classical	78.6	77.6

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

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

[5] 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.

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

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

[8] 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.

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

[10] 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.

[11] J. L. Flanagan and R. M. Golden, "Phase vocoder," *The Bell System Technical Journal*, vol. 45, pp. 1493–1509, 1966.

[12] T. Abe *et al.*, "The IF spectrogram: a new spectral representation," *Proc. ASVA 97*, pp. 423–430, 1997.