Working Notes of the IJCAI-99 Workshop on Computational Auditory Scene Analysis, pp.31-40, August 1999.

A Real-time Music Scene Description System: Detecting Melody and Bass Lines in Audio Signals

Masataka Goto **Satoru Hayamizu**
Machine Understanding Division, Electrotechnical Laboratory
1-1-4 Umezono, Tsukuba, Ibaraki 305-8568 JAPAN.
goto@etl.go.jp hayamizu@etl.go.jp

Abstract

This paper describes a predominant-pitch estimation method that enables us to build a real-time system detecting melody and bass lines as a subsystem of our music scene description system. The purpose of this study is to build such a real-time system that is practical from the engineering viewpoint, that gives suggestions to the modeling of music understanding, and that is useful in various applications. Most previous pitch-estimation methods premised either a single-pitch sound with aperiodic noises or a few musical instruments and had great difficulty dealing with complex audio signals sampled from compact discs, especially discs recording jazz or popular music with drum-sounds. Our method can estimate the most predominant fundamental frequency (F0) in such signals containing sounds of various instruments because it does not rely on the F0's frequency component, which is often overlapped by other sounds' components, and instead estimates the F0 by using the *Expectation-Maximization* algorithm on the basis of harmonics' frequency components within an intentionally limited frequency range. It also uses a multiple-agent architecture to stably track the temporal trajectory of the F0. Experimental results show that the system is robust enough to estimate the predominant F0s of the melody and bass lines in real-world audio signals.

1 Introduction

A typical research approach to computational auditory scene analysis is sound source segregation: the extraction, from sound mixtures such as ones people find in the real-world environment, of the audio signal corresponding to each auditory stream. Human listeners can obviously understand various properties of such sound mixtures, and this suggests that the listeners detect the existence of some auditory objects in sound mixtures and obtain a description of them. The understanding, however, is not necessarily evidence that the human auditory system extracts the individual audio signal corresponding to each auditory stream, although the sound source segregation is valuable from the viewpoint of engineering. This is because segregation is in general not a necessary condition for understanding: even if a mixture of two objects cannot be segregated, it can be understood that the two objects are included in the mixture on the basis of salient feature points of them. In developing a computational model of monaural or binaural sound source segregation, there is some possibility that we are dealing with the problem which is not solved by any mechanism in this world, not even by a human brain.

We therefore think that it is important to first build a computational model that can obtain a certain description of auditory scene from sound mixtures. To emphasize this approach, we dare to call it *auditory scene description*. If we consider *perceptual sounds*[1] to be the description of an auditory scene, the term *auditory scene description* has the same meaning as the term *auditory scene analysis* utilized by Kashino [Kashino, 1994; Kashino *et al.*, 1998]. In fact, Kashino [1994] also discussed the auditory scene analysis problem from a standpoint similar to ours by pointing out that the extraction of symbolic representation is more natural and essential than the restoration of a target signal wave from a sound mixture.

In modeling the auditory scene description, it is important that we discuss what an appropriate *description* of audio signals is. An easy way of specifying the description is to borrow the terminology of existing discrete symbol systems, such as musical scores consisting of musical notes and such as speech transcriptions consisting of text characters. Those symbols, however, represent

[1] The term *perceptual sound* was proposed by Kashino [Kashino, 1994; Kashino *et al.*, 1998] and means a cluster of acoustic energy which humans hear as one sound. It is defined as a symbol that corresponds to an acoustic (or auditory) entity.

only limited properties of audio signals. For example, they discard nonsymbolic properties such as the expressiveness of music performances (the deviation of pitch, loudness, and timing) and the prosody of spontaneous speeches. To take such properties into account, we need to introduce a subsymbolic description represented as continuous quantitative values. At the same time, we need to choose an appropriate level of abstraction for the description, because even though descriptions such as raw waveforms and spectra have continuous values they are too concrete. The appropriateness of the abstraction level will of course depend on the purpose of the description and on the use to which it will be put.

In this paper we address the problem of *music scene description*, auditory scene description in music, for monaural complex real-world audio signals such as those sampled from commercially distributed compact discs. We deal with various musical genres, such as popular music, jazz music, and classical works. The audio signals thus contain simultaneous sounds of various instruments (even drums). This real-world oriented approach with realistic assumptions is important to address the scaling-up problem and facilitate the implementation of practical applications [Goto and Muraoka, 1996; 1998a; 1998b].

The main contribution of this paper is to propose a predominant-pitch estimation method that makes it possible to detect the melody and bass lines in such audio signals. On the basis of the method, a real-time system detecting those lines has been implemented as a subsystem of the entire music-scene-description system. In the following sections, we first discuss the description used in our music scene description system and difficulties encountered in detecting the melody and bass lines. We then describe the algorithm of the predominant-pitch estimation method that is a core part of our system. Finally, we show experimental results obtained using our system.

2 Music Scene Description Problem

We first specify the entire music-scene-description problem and present the main difficulties in detecting the melody and bass lines, which is the subproblem that we are dealing with in this paper.

2.1 Problem Specification

Music scene description is defined as a process that obtains a description representing the input musical audio signal. Since various levels of description are possible, it is necessary to decide a description that is appropriate as the first step toward the ultimate description in human brains. We think that the music score is not adequate because, as we have already pointed out [Goto and Muraoka, 1999], an untrained listener understands music to some extent without mentally representing audio signals as musical scores. Music transcription, identifying the names (symbols) of musical notes and chords, is in fact a skill mastered only by trained musicians. We think that the appropriate description should satisfy the following requirements:

1. It is an intuitive description which can easily be obtained by untrained listeners.
2. It is a basic description which trained musicians can use as a basis for higher-level music understanding.
3. It is a useful description facilitating the development of various practical applications.

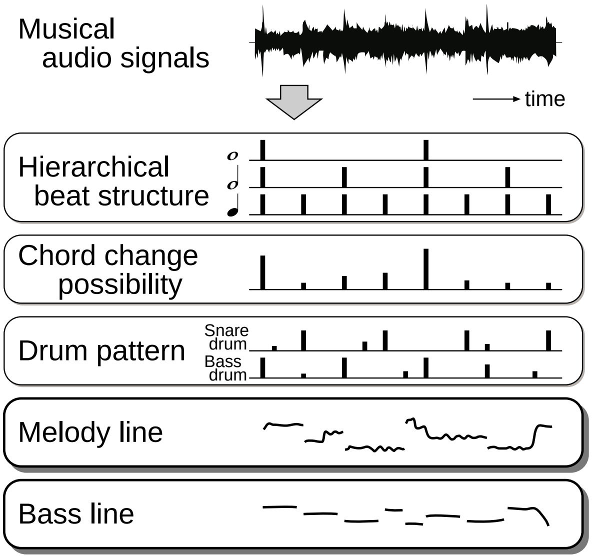

Figure 1: Description in our music scene description system.

According to these requirements, we propose a description consisting of the following five subsymbolic representations shown in Figure 1:

(1) Hierarchical beat structure
This represents the fundamental temporal structure of music and comprises three hierarchical levels.

(2) Chord change possibility
This represents possibilities of chord changes and indicates how much dominant frequency components included in chord tones and their harmonic overtones change.

(3) Drum pattern
This represents temporal patterns of how two principal drums, a bass drum and a snare drum, are played.

(4) *Melody line*
This represents the temporal trajectory of melody, which is a series of single tones and is more distinctly heard than the rest. Note that this is not a series of

musical notes but is a continuous representation of frequency and amplitude transitions.

(5) *Bass line*

This represents the temporal trajectory of bass, which is a series of single tones and is the lowest part in polyphonic music.

The idea of these representations came from the observation of how untrained listeners listen to music. The description consisting of the former three and the methods for obtaining them have already been proposed from the viewpoint of beat-tracking in our previous papers [Goto and Muraoka, 1994; 1996; 1998a; 1998b; 1999; Goto, 1998].

In this paper we introduce the latter two, *melody line* and *bass line*, into the description and address the issues of obtaining them. The detection of the melody and bass lines is important because the melody forms the core of Western music and is very influential in the identity of a musical piece and the bass is closely related with the tonality. For both trained and untrained listeners, the melody and bass lines are fundamental to the perception of music. They are also useful in various applications, such as automatic transcription, automatic music indexing for information retrieval like [Sonoda *et al.*, 1998], computer participation in human live performances, musical performance analysis of formerly-recorded outstanding performances, and automatic production of accompaniment tracks of *Karaoke* or *Music Minus One* by making use of compact discs.

In short, we solve the problem of obtaining the description of the melody line $D_m(t)$ and the bass line $D_b(t)$ given by

$$D_m(t) = \{F_m(t), A_m(t)\}, \quad (1)$$

$$D_b(t) = \{F_b(t), A_b(t)\}, \quad (2)$$

where $F_i(t)$ $(i = m, b)$ denotes the fundamental frequency (F0) at time t and $A_i(t)$ denotes the amplitude at t.

2.2 Issues in Detecting Melody and Bass Lines

It has 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 [Noll, 1967; Schroeder, 1968; Rabiner *et al.*, 1976; Nehorai and Porat, 1986; Charpentier, 1986; Ohmura, 1994; Abe *et al.*, 1996; Kawahara *et al.*, 1998a] premised that the input audio signal contained just a single-pitch sound or a single-pitch sound with a noisy aperiodic sound. Although several methods for dealing with multiple-pitch mixtures were proposed in the context of sound source segregation and automatic music transcription [Parsons, 1976; Chafe and Jaffe, 1986; Katayose and Inokuchi, 1989; Brown and Cooke, 1994; Nakatani *et al.*, 1995; Kashino and Murase, 1997; Kashino *et al.*, 1998], 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 the difficulty of F0 estimation in sound mixtures is that frequency components of one sound overlap frequency components of simultaneous sounds in the time-frequency domain. In the case of typical popular music performed by a voice, a keyboard instrument (like the piano), an electric guitar, a bass guitar, and drums, for example, a part of the voice's harmonic structure — especially its F0's frequency component — is often overlapped by harmonics of the keyboard instrument and 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 that rely on the existence of the F0's frequency component not only cannot handle the *missing fundamental* but are also unreliable when used with complex mixtures where the F0's component is smeared by the harmonics of simultaneous sounds.

3 Predominant-Pitch Estimation Method

We propose a method for estimating the fundamental frequency (F0) of the most predominant harmonic structure in a limited frequency region of sound mixtures. The method makes it possible to detect the melody and bass lines because the melody line usually keeps the most predominant harmonic structure in middle and high frequency regions and the bass line usually keeps the most predominant harmonic structure in a low frequency region. The method has the following features that enable the robust F0 estimation in complex sound mixtures:

- While the method assumes that the target predominant sound has the harmonic structure, it does not rely on the existence of the F0's frequency component and can deal with the *missing fundamental*. In other words, it can estimate the F0 by using a subset of the harmonic structure.
- The method dares to make use of frequency components in a limited frequency range and finds the F0 whose harmonics are predominant in that range. In other words, whether the F0 is within that range or not, the method tries to estimate the F0 which is supported by predominant frequency components as the harmonics.
- The method regards the observed frequency components as a weighted mixture of harmonic-structure

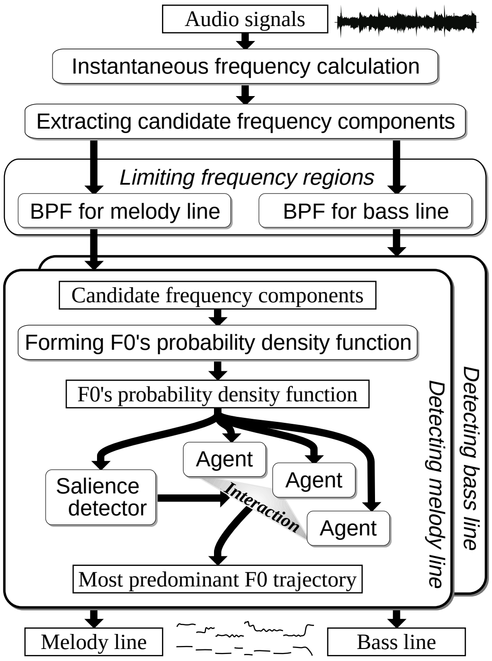
Audio signals
Instantaneous frequency calculation
Extracting candidate frequency components
Limiting frequency regions
BPF for melody line
BPF for bass line
Candidate frequency components
Forming F0's probability density function
F0's probability density function
Agent
Agent
Salience detector
Interaction
Agent
Most predominant F0 trajectory
Detecting melody line
Detecting bass line
Melody line
Bass line

Figure 2: Overview of our predominant-pitch estimation method.

tone models and estimates their weights by using the *Expectation-Maximization (EM)* algorithm [Dempster *et al.*, 1977], which is an iterative technique for computing maximum likelihood estimates from incomplete data. The method then considers the maximum-weight model as the most predominant harmonic structure and obtains its F0.

- Since the local F0 estimation is not reliable, the method supports sequential F0 tracking on the basis of a multiple-agent architecture in which agents track different temporal trajectories of the F0.

In particular, it is important to deemphasize a frequency region around the F0 in estimating the F0 of the melody line, because its frequency region is typically very crowded with other frequency components.

The strategy of this method is related to the *singing formant*, a high spectrum-envelope peak near 2.8 kHz of vowel sounds produced in male opera singing, though the application of the method is not limited to opera singing. Although sounds from an orchestral accompaniment tend to mask the singer's voice around a peak (about 450 Hz) of their long-time-average spectrum, the singing formant enables listeners to hear the voice over the high level of sounds from the orchestra because it has predominant frequency components in the higher limited range [Richards, 1988]. While we do not intend to build a psychoacoustical model of human perception, our strategy also has likely relevance to the following psychoacoustical results: Ritsma [1967] reported that the ear uses a rather limited spectral region in producing a well-defined pitch perception; Plomp [1967] concluded that for fundamental frequencies up to about 1400 Hz, the pitch of a complex tone is determined by the second and higher harmonics rather than by the fundamental. We need to note, however, that those results do not directly support our strategy since they were obtained by using the pitch of a single sound.

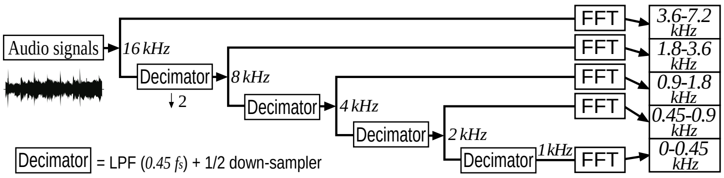

Figure 3: Overview of multirate filter bank.

Figure 2 shows an overview of the method. It first calculates instantaneous frequencies by using multirate signal processing techniques. Candidate frequency components are then extracted on the basis of an instantaneous-frequency-related measure. Those components are filtered with two bandpass filters, one for the melody line and the other for the bass line. Each set of the filtered frequency components is utilized to form a probability density function (PDF) of the F0, and the salient promising peaks in the F0's PDF are sequentially tracked by a multiple-agent architecture, where each agent forms an F0 trajectory and evaluates its reliability. Finally, the most predominant F0 trajectory of the most reliable agent is selected as the output.

3.1 Instantaneous Frequency Calculation

The method first calculates the *instantaneous frequency* [Flanagan and Golden, 1966; Cohen, 1989; Boashash, 1992], the rate of change of the phase of a signal, of filter-bank outputs. Instead of calculating the phase directly, it utilizes an efficient calculation method [Flanagan and Golden, 1966] based on the short-time Fourier transform (STFT) whose output can be interpreted as a collection of uniform-filter outputs. When the STFT of a signal $x(t)$ is defined as

$$X(\omega, t) = \int_{-\infty}^{\infty} x(\tau) h(\tau - t) e^{-j\omega\tau} d\tau \quad (3)$$

$$= a + jb, \quad (4)$$

the instantaneous frequency $\lambda(\omega, t)$ is given by

$$\lambda(\omega, t) = \omega + \frac{a\frac{\partial b}{\partial t} - b\frac{\partial a}{\partial t}}{a^2 + b^2}. \quad (5)$$

We use the STFT window function $h(t)$ obtained by convolving a basis function of the 2nd-order cardinal B-spline with an isometric Gaussian window function [Kawahara *et al.*, 1998b]. Although the importance of the appropriate choice of $h(t)$ is discussed in [Kawahara *et al.*, 1998b], our window function is not optimized to the F0 of a single tone since there may be several different F0s within an STFT window.

Because a single STFT provides bad time-frequency resolution for a certain frequency range, we use a multirate filter bank [Vetterli, 1987]. Since the Wavelet transform providing the minimum uncertainty is hard to be performed in real time, we design an STFT-based filter bank that provides an adequate time-frequency resolution compromise under the real-time constraint.

Figure 3 shows an overview of our binary-tree filter bank. At each level of binary branches, the audio signal is down-sampled by a decimator that consists of an anti-aliasing filter (FIR lowpass filter (LPF)) and a 1/2 down-sampler. The cut-off frequency of the LPF in each decimator is 0.45 f_s where f_s is the sampling rate at that branch. In our current implementation, the input signal is digitized at 16 bit / 16 kHz, and it is finally down-sampled to 1 kHz. Then the STFT whose window size is 512 samples is calculated at each leaf by using the Fast Fourier Transform (FFT) while compensating for the different 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*[2]) is 10 ms.

3.2 Extracting Candidate Frequency Components

The extraction of candidate frequency components is based on the frequency-to-instantaneous-frequency mapping [Charpentier, 1986; Abe *et al.*, 1996; Kawahara *et al.*, 1998b]. We consider the mapping from the center frequency ω of an STFT filter to the instantaneous frequency $\lambda(\omega, t)$ of its output. If there is a frequency component at frequency ψ, ψ is placed at the fixed point of the mapping and the instantaneous frequencies around ψ stay almost constant in the mapping [Kawahara *et al.*, 1998b]. Therefore a set $\Psi_f^{(t)}$ of instantaneous frequencies of the candidate frequency components[3] can be extracted by using the following equation [Abe *et al.*, 1997]:

$$\Psi_f^{(t)} = \{\ \psi \ \mid\ \lambda(\psi, t) - \psi = 0, \\ \frac{\partial}{\partial\psi}(\lambda(\psi, t) - \psi) < 0\}. \qquad (6)$$

By calculating the power of those frequencies which is given by the STFT power spectrum at $\Psi_f^{(t)}$, we can define the power distribution function $\Psi_p^{(t)}(\omega)$ as

$$\Psi_p^{(t)}(\omega) = \begin{cases} \mid X(\omega, t) \mid & \text{if } \omega \in \Psi_f^{(t)} \\ 0 & \text{otherwise.} \end{cases} \qquad (7)$$

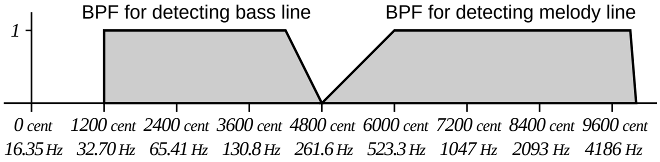

Figure 4: Frequency responses of bandpass filters (BPFs).

3.3 Limiting Frequency Regions

Limiting (i.e., weighting) the frequency region of frequency components is important to estimate a predominant F0. We introduce different kinds of bandpass filters (BPFs) for the melody and bass lines. 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. 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 our current implementation we use BPFs whose frequency responses are shown in Figure 4. In the rest of this paper, to denote the log-scale frequency we use the unit of *cent* (a musical-interval measurement). The frequency f_{Hz} in Hz is converted to the frequency f_{cent} in cent as follows:

$$f_{cent} = 1200 \log_2 \frac{f_{Hz}}{\mathrm{REF}_{Hz}}, \qquad (8)$$

$$\mathrm{REF}_{Hz} = 440 \times 2^{\frac{3}{12}-5}. \qquad (9)$$

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

The bandpass-filtered frequency components can be represented as $BPF_i(x)\Psi_p'^{(t)}(x)$ where $BPF_i(x)$ $(i = m, b)$ denotes the BPF's frequency response at frequency x (in cents) and $\Psi_p'^{(t)}(x)$ is the same as the power distribution $\Psi_p^{(t)}(\omega)$ except that the frequency unit is the cent. In preparation for the next step, we define the probability density function (PDF) of the bandpass-filtered frequency components $p_\Psi^{(t)}(x)$:

$$p_\Psi^{(t)}(x) = \frac{BPF_i(x)\ \Psi_p'^{(t)}(x)}{Pow^{(t)}}, \qquad (10)$$

where $Pow^{(t)}$ is the total power of the bandpass-filtered frequency components:

$$Pow^{(t)} = \int_{-\infty}^{\infty} BPF_i(x)\ \Psi_p'^{(t)}(x)\ dx. \qquad (11)$$

3.4 Forming the F0's Probability Density Function

For each set of candidate frequency components filtered by the BPF, we form a probability density function (PDF) of the F0. The basic idea is to consider that

[2] The frame-time is the unit of time used in our system, and the term *time* in this paper is the time measured in units of frame-time.

[3] Abe *et al.* [1997] called the temporal contours of these components *IF attractors*.

the PDF of the bandpass-filtered frequency components, $p_{\Psi}^{(t)}(x)$, 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

$$p(x;\theta^{(t)}) = \int_{\mathrm{Fl}_i}^{\mathrm{Fh}_i} w^{(t)}(F)\ p(x|F)\ dF, \qquad (12)$$

$$\theta^{(t)} = \{w^{(t)}(F)\ \mid\ \mathrm{Fl}_i \leq F \leq \mathrm{Fh}_i\}, \qquad (13)$$

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

$$\int_{\mathrm{Fl}_i}^{\mathrm{Fh}_i} w^{(t)}(F)\ dF = 1. \qquad (14)$$

Note that we simultaneously take into consideration all the possibilities of the F0 because we cannot assume the number of sound sources in real-world audio signals sampled from compact discs. If we can estimate the model parameter $\theta^{(t)}$ so that the observed PDF $p_{\Psi}^{(t)}(x)$ is likely to have been generated from the model $p(x;\theta^{(t)})$, $p_{\Psi}^{(t)}(x)$ is considered to be decomposed into harmonic-structure tone models, and the weight $w^{(t)}(F)$ can be interpreted as the PDF of the F0:

$$p_{F0}^{(t)}(F) = w^{(t)}(F) \quad (\mathrm{Fl}_i \leq F \leq \mathrm{Fh}_i). \qquad (15)$$

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 parameter $\theta^{(t)}$ of the model $p(x;\theta^{(t)})$ when we observed the PDF $p_{\Psi}^{(t)}(x)$. The maximum likelihood estimator of $\theta^{(t)}$ is obtained by maximizing the mean log-likelihood defined as

$$\int_{-\infty}^{\infty} p_{\Psi}^{(t)}(x)\ \log p(x;\theta^{(t)})\ dx. \qquad (16)$$

Because this maximization problem is too difficult to be solved analytically, we use the *Expectation-Maximization (EM)* algorithm [Dempster *et al.*, 1977], which is an iterative algorithm successively applying two steps — the *expectation step (E-step)* and the *maximization step (M-step)* — for computing maximum likelihood estimates from incomplete observed data (in our case, $p_{\Psi}^{(t)}(x)$). With respect to the parameter $\theta^{(t)}$, each iteration updates the 'old' parameter estimate $\theta'^{(t)}$ to obtain the 'new' parameter estimate $\overline{\theta^{(t)}}$. We simply use the final estimate at the previous frame-time $t-1$ for the initial estimate of $\theta'^{(t)}$.

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

1. (E-step)
   Compute the following conditional expectation of the mean log-likelihood:
   $$Q(\theta^{(t)}|\theta'^{(t)}) = \int_{-\infty}^{\infty} p_{\Psi}^{(t)}(x)\ \mathrm{E}_F[\log p(x,F;\theta^{(t)})\mid x;\theta'^{(t)}]\ dx, \qquad (17)$$
   where $\mathrm{E}_F[a|b]$ denotes the conditional expectation of a with respect to the hidden variable F with probability distribution determined by the condition b.
2. (M-step)
   Maximize $Q(\theta^{(t)}|\theta'^{(t)})$ as a function of $\theta^{(t)}$ in order to obtain the updated (improved) estimate $\overline{\theta^{(t)}}$:
   $$\overline{\theta^{(t)}} = \underset{\theta^{(t)}}{\mathrm{argmax}}\ Q(\theta^{(t)}|\theta'^{(t)}). \qquad (18)$$

In the E-step we have

$$Q(\theta^{(t)}|\theta'^{(t)}) = \int_{-\infty}^{\infty}\int_{\mathrm{Fl}_i}^{\mathrm{Fh}_i} p_{\Psi}^{(t)}(x) p(F|x;\theta'^{(t)}) \log p(x,F;\theta^{(t)}) dF dx, \qquad (19)$$

where the complete-data log-likelihood is given by

$$\log p(x,F;\theta^{(t)}) = \log(w^{(t)}(F)\ p(x|F)) = \log w^{(t)}(F) + \log p(x|F). \qquad (20)$$

As for the M-step, Equation (18) is a conditional problem of variation where the condition is Equation (14). This problem can be solved by using the following Euler-Lagrange differential equation:

$$\frac{\partial}{\partial w^{(t)}}\left(\int_{-\infty}^{\infty} p_{\Psi}^{(t)}(x)\ p(F|x;\theta'^{(t)})\ (\log w^{(t)}(F) + \log p(x|F))\ dx - \lambda(w^{(t)}(F) - \tfrac{1}{\mathrm{Fh}_i - \mathrm{Fl}_i})\right) = 0, \qquad (21)$$

where λ is a Lagrange multiplier. From Equation (21), we have

$$w^{(t)}(F) = \frac{1}{\lambda}\int_{-\infty}^{\infty} p_{\Psi}^{(t)}(x)\ p(F|x;\theta'^{(t)})\ dx. \qquad (22)$$

In this equation, λ is determined from Equation (14) to be $\lambda = 1$ and we know from the Bayes' theorem that $p(F|x;\theta'^{(t)})$ is given

$$p(F|x;\theta'^{(t)}) = \frac{w'^{(t)}(F)\ p(x|F)}{\int_{\mathrm{Fl}_i}^{\mathrm{Fh}_i} w'^{(t)}(\eta)\ p(x|\eta)\ d\eta}, \qquad (23)$$

where $w'^{(t)}(F)$ is the 'old' parameter estimate ($\theta'^{(t)} = \{w'^{(t)}(F)\}$). Finally we obtain the 'new' improved parameter estimate $\overline{w^{(t)}(F)}$:

$$\overline{w^{(t)}(F)} = \int_{-\infty}^{\infty} p_{\Psi}^{(t)}(x) \frac{w'^{(t)}(F)\ p(x|F)}{\int_{\mathrm{Fl}_i}^{\mathrm{Fh}_i} w'^{(t)}(\eta)\ p(x|\eta)\ d\eta}\ dx. \qquad (24)$$

To compute Equation (24), we need to assume the PDF of a tone model $p(x|F)$, which indicates where the harmonics of the F0 F tend to occur. We accordingly assume the following simple harmonic-structure tone models for the melody line ($i = m$) and the bass line ($i = b$):

$$p(x|F) = \alpha \sum_{h=1}^{\mathrm{N}_i} c(h)\ G(x; F + 1200 \log_2 h, \mathrm{W}_i), \quad (25)$$

$$G(x; m, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}}\ e^{-\frac{(x-m)^2}{2\sigma^2}}, \quad (26)$$

where α is a normalization factor, N_i is the number of harmonics considered, W_i^2 is the variance of the Gaussian distribution $G(x; m, \sigma)$, and $c(h)$ determines the amplitude of the h-th harmonic component. For $c(h)$ we use $G(h; 0, \mathrm{H}_i)$ where H_i is a constant. Since these models are very simple, there is a great deal of room for refining them in future implementations. This could be done, for example, by introducing tone memories.

A simple way of determining the frequency $F_i(t)$ of the most predominant F0 is to find the frequency that maximizes the F0's PDF $p_{F0}^{(t)}(F)$ (Equation (15)), which is the final estimate obtained by the iterative computation of Equation (24):

$$F_i(t) = \underset{F}{\operatorname{argmax}}\ p_{F0}^{(t)}(F). \quad (27)$$

This result is not stable, however, because peaks corresponding to the F0s of several simultaneous tones sometimes compete in the F0's PDF for a moment and are transiently selected, one after another, as the maximum of the F0's PDF. It is therefore necessary to consider the global temporal continuity of the F0 peak. This is addressed in the next section.

3.5 Sequential F0 Tracking by Multiple-Agent Architecture

The method sequentially tracks peak trajectories in the temporal transition of the F0's PDF in order to select the most predominant and stable F0 trajectory from the viewpoint of global F0 estimation.[4] To perform this, we introduce a multiple-agent architecture that enables dynamic and flexible control of the tracking process. In the multiple-agent architecture we proposed earlier [Goto and Muraoka, 1996] the number of agents was fixed during the processing. Our new architecture, however, generates and terminates agents dynamically by using a mechanism similar to one in the residue-driven architecture [Nakatani *et al.*, 1995].

The architecture consists of a salience detector and multiple agents. At each frame the salience detector picks up salient promising peaks in the F0's PDF, and agents driven by those peaks track their trajectories (Figure 5). At the first of the processing, no agent has been generated and there is only the salience detector. They then behave at each frame as follows:

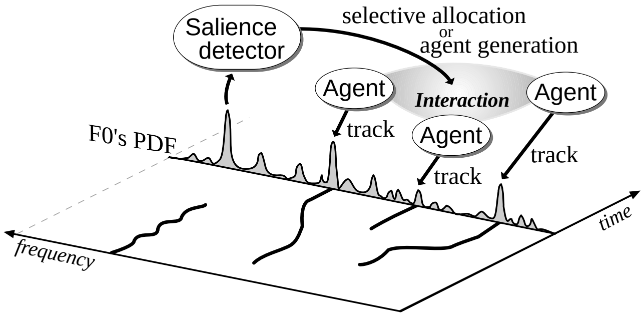

Figure 5: Sequential F0 tracking by multiple-agent architecture.

(1) After forming the F0's PDF at each frame, the salience detector picks up several salient peaks.

(2) If there are agents, they interact to allocate the salient peaks to agents exclusively according to peak-closeness criteria.

(3) If the most salient peak has not been allocated, a new agent for tracking its peak is generated.

(4) Each agent has an accumulated penalty. An agent whose accumulated penalty exceeds a certain threshold is terminated.

(5) An agent to which a salient peak has not been allocated is penalized a certain value and tries to find its next peak in the F0's PDF directly. When the agent cannot find the peak even in the F0's PDF, it is further penalized a certain value. Otherwise, the penalty is reset.

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

(7) The output at t — the F0 $f_0^{(t)}$ and its reliability $r_0^{(t)}$ — 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.

Salience Detector

The salience detector picks up salient peaks $\Phi_f^{(t)}$ of the current F0's PDF. To track peaks in the temporal transition of the F0's PDF, we first define the F0 peak map $m_{F0}^{(t)}(F)$ considering the total power transition:

$$m_{F0}^{(t)}(F) = \begin{cases} Pow^{(t)}\ p_{F0}^{(t)}(F) & \text{if } \frac{\partial}{\partial F} p_{F0}^{(t)}(F) = 0, \\ & \quad \frac{\partial^2}{\partial F^2} p_{F0}^{(t)}(F) < 0 \\ 0 & \text{otherwise.} \end{cases} \quad (28)$$

[4] Because the F0's PDF is obtained without assuming the number of sounds contained, our method can, by using an appropriate sound-source discrimination method, be extended to the problem of tracking multiple simultaneous sounds.

The salient peaks $\Phi_f^{(t)}$ are then given by

$$\Phi_f^{(t)} = \{\, F \mid m_{F0}^{(t)}(F) \ \geq\ \mathrm{Th_s} \max_F \ m_{F0}^{(t)}(F)\}, \qquad (29)$$

where $\mathrm{Th_s}$ is a constant threshold ($0 \leq \mathrm{Th_s} \leq 1$) for the judgement of the salience.

By tracking a provisional peak trajectory of $m_{F0}^{(t)}(F)$ in the near future, the salience detector evaluates the *salience degree* $\Phi_r^{(t)}(F)$ of each peak, which represents how promising its peak is:[5]

$$\Phi_r^{(t)}(F) = \begin{cases} \frac{1}{(\mathrm{Period_s}+1)\ SPow^{(t)}} \Big(m_{F0}^{(t)}(F)\ G(0;0,\mathrm{W_s}) \\ \qquad + \sum_{\tau=1}^{\mathrm{Period_s}} \max_{|f|<2\mathrm{W_s}} \phi_r^{(t+\tau)}(f;F) \Big) \\ \qquad\qquad \text{if } F \in \Phi_f^{(t)} \\ 0 \qquad\qquad \text{otherwise,} \end{cases} \qquad (30)$$

$$SPow^{(t)} = \frac{1}{\mathrm{Period_s}+1} \sum_{\tau=0}^{\mathrm{Period_s}} Pow^{(t+\tau)}, \qquad (31)$$

where $\mathrm{Period_s}$ is the period of tracking a provisional peak trajectory and W_s^2 is the variance of the Gaussian distribution $G(x;m,\sigma)$. The term $\phi_r^{(t+\tau)}(f;F)$ is defined as

$$\phi_r^{(t+\tau)}(f;F) = m_{F0}^{(t+\tau)}(\eta^{(t+\tau-1)}(F) + f)\ G(f;0,\mathrm{W_s}), \qquad (32)$$

where f is the amount of the frequency change from the previous frame, and $\eta^{(t+\tau)}(F)$ denotes the tracked peak frequency at $t+\tau$:

$$\eta^{(t+\tau)}(F) = \begin{cases} \eta^{(t+\tau-1)}(F) + \operatorname*{argmax}_{|f|<2\mathrm{W_s}} \phi_r^{(t+\tau)}(f;F) \\ \qquad\qquad \text{if } \tau > 0 \\ F \qquad\qquad \text{otherwise.} \end{cases} \qquad (33)$$

The term $\phi_r^{(t+\tau)}(f;F)$ denotes a possibility indicating how much the peak frequency is likely to change from the previous peak frequency $\eta^{(t+\tau-1)}(F)$.

Agent

Each agent j has the following set of parameters and updates it every frame: the frequency $Af_j^{(t)}$ of the peak it is tracking, the salience degree $Ad_j^{(t)}$ of the peak, an accumulated peak-not-found penalty $Ap_j^{(t)}$, the sum $As_j^{(t)}$ of $SPow^{(t)}$ since the agent was generated, and the reliability $Ar_j^{(t)}$.

Among the salient peaks $\Phi_f^{(t)}$, each agent finds the most salient peak $Cf_j^{(t)}$ closest to its previous peak frequency $Af_j^{(t-1)}$:

$$Cf_j^{(t)} = \operatorname*{argmax}_F \Phi_r^{(t)}(F)\ G(F; Af_j^{(t-1)}, \mathrm{W_s}). \qquad (34)$$

When $Cf_j^{(t)}$ is close enough to $Af_j^{(t-1)}$ — that is, when $|Cf_j^{(t)} - Af_j^{(t-1)}| < 2\mathrm{W_s}$ — the peak is allocated to agent j and its parameter $Af_j^{(t)}$ is updated to $Cf_j^{(t)}$. If more than one agent claims the same peak, it is exclusively allocated to the most predominant agent.

If the most salient peak, $\mathrm{argmax}_F\, \Phi_r^{(t)}(F)$, is not allocated, a new agent is generated and its parameter $Af_j^{(t)}$ is set to $\mathrm{argmax}_F\, \Phi_r^{(t)}(F)$. On the other hand, an agent whose penalty $Ap_j^{(t)}$ reaches 1.0 is terminated. When an agent cannot find the next peak in $\Phi_f^{(t)}$, its $Ap_j^{(t)}$ is increased by a positive constant $\mathrm{V_a}$, and when it cannot find the next peak even in $m_{F0}^{(t)}(F)$, its $Ap_j^{(t)}$ is further increased by $\mathrm{V_a}$. Otherwise, $Ap_j^{(t)}$ is reset to 0.0.

After each agent updates the salient degree $Ad_j^{(t)} = \Phi_r^{(t)}(Af_j^{(t)})$ and the power sum $As_j^{(t)} = As_j^{(t-1)} + SPow^{(t)}$, it updates its reliability $Ar_j^{(t)}$ as follows:

$$Ar_j^{(t)} = \mathrm{R}_a \frac{As_j^{(t-1)}}{As_j^{(t)}} Ar_j^{(t-1)} + (1 - \mathrm{R}_a) \frac{SPow^{(t)}}{As_j^{(t)}} Ad_j^{(t)}, \qquad (35)$$

where R_a is a weight ($0 \leq \mathrm{R}_a < 1$) determining how much the previous reliability is taken into account. The most predominant F0 trajectory ($f_0^{(t)}$ and $r_0^{(t)}$) is then determined on the basis of the parameters $Af_j^{(t)}$ and $Ar_j^{(t)}$ of the most predominant agent whose number is $\mathrm{argmax}_j\, Ar_j^{(t)} As_j^{(t)}$.

Finally, the F0 $F_i(t)$ and the amplitude $A_i(t)$ of the melody line $D_m(t)$ (Equation (1)) and the bass line $D_b(t)$ (Equation (2)) are determined according to the harmonic structure of those lines. The harmonic structure of each line is estimated by tracking, in the candidate frequency components $\Psi_p^{(t)}(\omega)$, the harmonics along its F0 trajectory. The $F_i(t)$ is finely adjusted according to the frequencies of the harmonics and $A_i(t)$ is obtained as the total power of the harmonics.

4 System Implementation and Experimental Results

Using the proposed method (the values of the parameters are listed in Table 1), we have built a real-time system that takes a musical audio signal as input and

Table 1: Values of parameters.

Fh_m = 9600 cent	Fh_b = 4800 cent
Fl_m = 3600 cent	Fl_b = 1000 cent
N_m = 16	N_b = 6
W_m = 17 cent	W_b = 17 cent
H_m = 5.5	H_b = 2.7
$\mathrm{Th_s}$ = 0.5	$\mathrm{W_s}$ = 50 cent
$\mathrm{Period_s}$ = 5 frame-time	
$\mathrm{V_a}$ = 0.125	R_a = 0.91

[5] In a real-time system, the $m_{F0}^{(t+\tau)}(F)$ ($0 \leq \tau \leq \mathrm{Period_s}$) can be accessed by regarding the actual current time as the future time $t + \mathrm{Period_s}$; the system output is consequently delayed for a short period, at least $\mathrm{Period_s}$.

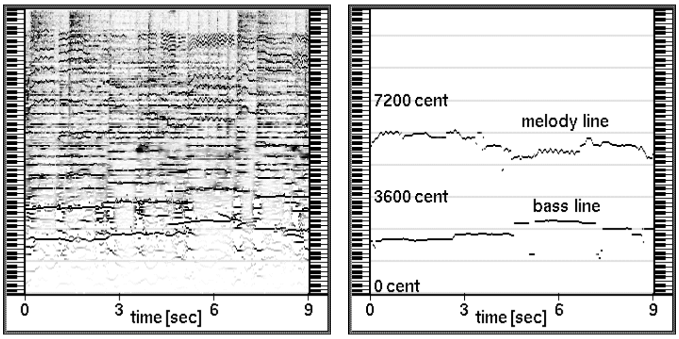

Figure 6: Scrolling-window snapshots of candidate frequency components (left) and the corresponding detected melody and bass lines (right) for a popular-music excerpt with drum-sounds.

Table 2: Test songs sampled from compact discs.

title	genre
My Heart Will Go On (Celine Dion)	popular
Vision of Love (Mariah Carey)	popular
Always (Bon Jovi)	popular
Time Goes By (Every Little Thing)	popular
Spirit of Love (Sing Like Talking)	popular
Hoshi no Furu Oka (Misia)	popular
Scarborough Fair (Herbie Hancock)	jazz
Autumn Leaves (Julian “Cannonball” Adderley)	jazz
On Green Dolphin Street (Miles Davis)	jazz
Violin Concerto in D, Op. 35 (Tchaikovsky)	classical

outputs the detected melody and bass lines in several forms: computer graphics for visualization, audio signals for auralization, and continuous quantitative values (with time stamps) for use in applications. The graphics output shows a window representing the scrolling F0 trajectories on a time-frequency plane and an adjacent interlocking window representing the candidate frequency components (Figure 6). The output audio signals are generated by sinusoidal synthesis on the basis of the harmonics that are tracked by a 2nd-order autoregressive tracking model [Aikawa *et al.*, 1996] guided by the detected $D_i(t)$.

The system has been implemented using a distributed-processing technique so that different system functions — such as audio input and output (I/O), main calculation, and intermediate-state and output visualization — are performed by different processes distributed over a LAN (Ethernet). To facilitate system expansion and application development, those processes are implemented on the basis of a network protocol called *RACP (Remote Audio Control Protocol)*, which is an extension of the *RMCP (Remote Music Control Protocol)* [Goto, ICMC97]. The main signal processing is performed on a personal computer with two Pentium II 450 MHz CPUs (Linux 2.2), and the audio I/O and visualization processing is performed on a workstation, the SGI Octane with R10000 250 MHz CPU (Irix 6.4).

We tested the system on excerpts of 10 songs in popular, jazz, and orchestral genres (Table 2). The input monaural audio signals were sampled from commercially distributed compact discs and each contained a single-tone melody with sounds of several instruments.

In our experiment the system correctly detected, for the most part of each audio sample, melody lines provided by a voice or a single-tone mid-range instrument and bass lines provided by a bass guitar or a contrabass. It tended to perform best on jazz music in which a wind instrument such as a trumpet and a saxophone provided the melody line because the tones of such instruments tended to be more dominant and salient in a jazz ensemble than in other genres. In the absence of the main vocal part or the solo part, the system detected the F0 trajectory of a dominant accompaniment part, because our method simply estimates the most predominant F0 trajectory every moment and does not discriminate sound sources.

The detected line, however, sometimes switched from the main vocal part to another obbligato part for a while even when the previously tracked main vocal part continued. Furthermore, a short-term trajectory around the onset of the main vocal part was sometimes missing because of the delay in switching from another part to the vocal part. These errors are due to the absence of a mechanism for selecting just the target part from several simultaneous streams; this issue should be addressed in our future implementation. Other typical errors were half-pitch or double-pitch errors in which the F0 was estimated as half or twice the actual F0.

5 Conclusion

We have described the problem of music scene description for complex real-world audio signals and have addressed the problem of detecting the melody and bass lines. Our method for estimating the most predominant F0 trajectory in monaural audio signals does not presuppose the existence of the F0's frequency component and uses partial information in an intentionally limited frequency range. Using the EM algorithm without assuming the number of sound sources, the method evaluates the probability density function of the F0 which represents the relative dominance of every possible harmonic structure. It also uses a multiple-agent architecture to determine the most predominant and stable F0 trajectory from the viewpoint of global temporal continuity of the F0. Experimental results show that our system implementing the method can estimate, in real time, the predominant F0s of the melody and bass lines in audio signals sampled from compact discs.

We plan to extend the method to track several streams simultaneously and form more complete melody and bass lines from them by using a selective-attention mechanism. That extension will also address the issues of sound source discrimination. Other future work will include integration of the proposed pitch-estimation subsystem with other subsystems detecting the hierarchical beat structure, chord-change possibilities, and drum patterns in order to build the entire music-scene-description system.

Acknowledgments

We thank Shotaro Akaho for his valuable discussions and for his helpful comments on earlier drafts of this paper.

References

[Abe *et al.*, 1996] Toshihiko Abe, Takao Kobayashi, and Satoshi Imai. Robust pitch estimation with harmonics enhancement in noisy environments based on instantaneous frequency. In *Proc. of ICSLP 96*, pages 1277–1280, 1996.

[Abe *et al.*, 1997] Toshihiko Abe, Takao Kobayashi, and Satoshi Imai. The IF spectrogram: a new spectral representation. In *Proc. of ASVA 97*, pages 423–430, 1997.

[Aikawa *et al.*, 1996] Kiyoaki Aikawa, Hideki Kawahara, and Minoru Tsuzaki. A neural matrix model for active tracking of frequency-modulated tones. In *Proc. of ICSLP 96*, pages 578–581, 1996.

[Boashash, 1992] Boualem Boashash. Estimating and interpreting the instantaneous frequency of a signal. *Proc. of the IEEE*, 80(4):520–568, 1992.

[Brown and Cooke, 1994] Guy J. Brown and Martin Cooke. Perceptual grouping of musical sounds: A computational model. *Journal of New Music Research*, 23:107–132, 1994.

[Chafe and Jaffe, 1986] Chris Chafe and David Jaffe. Source separation and note identification in polyphonic music. In *Proc. of ICASSP 86*, pages 1289–1292, 1986.

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

[Cohen, 1989] Leon Cohen. Time-frequency distributions — a review. *Proc. of the IEEE*, 77(7):941–981, 1989.

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

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

[Goto and Muraoka, 1994] Masataka Goto and Yoichi Muraoka. A beat tracking system for acoustic signals of music. In *Proc. of the Second ACM Intl. Conf. on Multimedia*, pages 365–372, 1994.

[Goto and Muraoka, 1996] Masataka Goto and Yoichi Muraoka. Beat tracking based on multiple-agent architecture — a real-time beat tracking system for audio signals —. In *Proc. of the Second Intl. Conf. on Multiagent Systems*, pages 103–110, 1996.

[Goto and Muraoka, 1998a] Masataka Goto and Yoichi Muraoka. An audio-based real-time beat tracking system and its applications. In *Proc. of Intl. Computer Music Conf.*, pages 17–20, 1998.

[Goto and Muraoka, 1998b] Masataka Goto and Yoichi Muraoka. Music understanding at the beat level — real-time beat tracking for audio signals —. In *Computational Auditory Scene Analysis*, pages 157–176. Lawrence Erlbaum Associates, Publishers, 1998.

[Goto and Muraoka, 1999] Masataka Goto and Yoichi Muraoka. Real-time beat tracking for drumless audio signals: Chord change detection for musical decisions. *Speech Communication*, 27(3–4):311–335, 1999.

[Goto, 1998] Masataka Goto. *A Study of Real-time Beat Tracking for Musical Audio Signals* (in Japanese). PhD thesis, Waseda University, 1998.

[Kashino and Murase, 1997] Kunio Kashino and Hiroshi Murase. A music stream segregation system based on adaptive multi-agents. In *IJCAI-97*, pages 1126–1131, 1997.

[Kashino *et al.*, 1998] Kunio Kashino, Kazuhiro Nakadai, Tomoyoshi Kinoshita, and Hidehiko Tanaka. Application of the bayesian probability network to music scene analysis. In *Computational Auditory Scene Analysis*, pages 115–137. Lawrence Erlbaum Associates, Publishers, 1998.

[Kashino, 1994] Kunio Kashino. *Computational Auditory Scene Analysis for Music Signals* (in Japanese). PhD thesis, University of Tokyo, 1994.

[Katayose and Inokuchi, 1989] Haruhiro Katayose and Seiji Inokuchi. The kansei music system. *Computer Music Journal*, 13(4):72–77, 1989.

[Kawahara *et al.*, 1998a] Hideki Kawahara, Alain de Cheveigné, and Roy D. Patterson. An instantaneous-frequency-based pitch extraction method for high-quality speech transformation: Revised TEMPO in the STRAIGHT suite. In *Proc. of ICSLP 98*, 1998.

[Kawahara *et al.*, 1998b] Hideki Kawahara, Haruhiro Katayose, Roy D. Patterson, and Alain de Cheveigné. Highly accurate F0 extraction using instantaneous frequencies *(in Japanese). Tech. Com. Psycho. Physio., Acoust. Soc. of Japan, H-98-116*, pages 31–38, 1998.

[Nakatani *et al.*, 1995] Tomohiro Nakatani, Hiroshi G. Okuno, and Takeshi Kawabata. Residue-driven architecture for computational auditory scene analysis. In *IJCAI-95*, pages 165–172, 1995.

[Nehorai and Porat, 1986] Arye Nehorai and Boaz Porat. Adaptive comb filtering for harmonic signal enhancement. *IEEE Trans. on ASSP*, ASSP-34(5):1124–1138, 1986.

[Noll, 1967] A. Michael Noll. Cepstrum pitch determination. *J. Acoust. Soc. Am.*, 41(2):293–309, 1967.

[Ohmura, 1994] Hiroshi Ohmura. Fine pitch contour extraction by voice fundamental wave filtering method. In *Proc. of ICASSP 94*, pages II–189–192, 1994.

[Parsons, 1976] Thomas W. Parsons. Separation of speech from interfering speech by means of harmonic selection. *J. Acoust. Soc. Am.*, 60(4):911–918, 1976.

[Plomp, 1967] R. Plomp. Pitch of complex tones. *J. Acoust. Soc. Am.*, 41(6):1526–1533, 1967.

[Rabiner *et al.*, 1976] Lawrence R. Rabiner, Michael J. Cheng, Aaron E. Rosenberg, and Carol A. McGonegal. A comparative performance study of several pitch detection algorithms. *IEEE Trans. on ASSP*, ASSP-24(5):399–418, 1976.

[Richards, 1988] Whitman Richards, editor. *Natural Computation*. The MIT Press, 1988.

[Ritsma, 1967] Roelof J. Ritsma. Frequencies dominant in the perception of the pitch of complex sounds. *J. Acoust. Soc. Am.*, 42(1):191–198, 1967.

[Schroeder, 1968] M. R. Schroeder. Period histogram and product spectrum: New methods for fundamental-frequency measurement. *J. Acoust. Soc. Am.*, 43(4):829–834, 1968.

[Sonoda *et al.*, 1998] Tomonari Sonoda, Masataka Goto, and Yoichi Muraoka. A WWW-based melody retrieval system. In *Proc. of Intl. Computer Music Conf.*, pages 349–352, 1998.

[Vetterli, 1987] Martin Vetterli. A theory of multirate filter banks. *IEEE Trans. on ASSP*, ASSP-35(3):356–372, 1987.