Available online at www.sciencedirect.com
SCIENCE DIRECT®
Speech Communication 43 (2004) 311–329
SPEECH COMMUNICATION
www.elsevier.com/locate/specom

A real-time music-scene-description system: predominant-F0 estimation for detecting melody and bass lines in real-world audio signals ☆

Masataka Goto *

National Institute of Advanced Industrial Science and Technology (AIST), 1-1-1 Umezono, Tsukuba, Ibaraki 305-8568, Japan

Received 22 May 2002; received in revised form 9 May 2003; accepted 13 March 2004

Abstract

In this paper, we describe the concept of *music scene description* and address the problem of detecting melody and bass lines in real-world audio signals containing the sounds of various instruments. Most previous pitch-estimation methods have had difficulty dealing with such complex music signals because these methods were designed to deal with mixtures of only a few sounds. To enable estimation of the fundamental frequency (F0) of the melody and bass lines, we propose a predominant-F0 estimation method called *PreFEst* that does not rely on the unreliable fundamental component and obtains the most predominant F0 supported by harmonics within an intentionally limited frequency range. This method estimates the relative dominance of every possible F0 (represented as a probability density function of the F0) by using MAP (maximum a posteriori probability) estimation and considers the F0's temporal continuity by using a multiple-agent architecture. Experimental results with a set of ten music excerpts from compact-disc recordings showed that a real-time system implementing this method was able to detect melody and bass lines about 80% of the time these existed.

Keywords: F0 estimation; MAP estimation; EM algorithm; Music understanding; Computational auditory scene analysis; Music information retrieval

1. Introduction

A typical research approach to computational auditory scene analysis (CASA) (Bregman, 1990; Brown, 1992; Cooke and Brown, 1993; Rosenthal and Okuno, 1995; Okuno and Cooke, 1997;

☆ Supported by "Information and Human Activity", Precursory Research for Embryonic Science and Technology (PRESTO), Japan Science and Technology Corporation (JST).
* Corresponding author. Tel.: +81 29 861 5898; fax: +81 29 861 3313.
E-mail address: m.goto@aist.go.jp

0167-6393/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2004.07.001

Rosenthal and Okuno, 1998) is sound source segregation: the extraction of the audio signal corresponding to each auditory stream in a sound mixture. Human listeners can obviously understand various properties of the sound mixtures they hear in a real-world environment and this suggests that they detect the existence of certain auditory objects in sound mixtures and obtain a description of them. This understanding, however, is not necessarily evidence that the human auditory system extracts the individual audio signal corresponding to each auditory stream. This is because segregation is not a necessary condition for understanding: even if a mixture of two objects cannot be segregated, that the mixture includes the two objects can be understood from their salient features. In developing a computational model of monaural or binaural sound source segregation, we might be dealing with a problem which is not solved by any mechanism in this world, not even by the human brain, although sound source segregation is valuable from the viewpoint of engineering.

In the context of CASA, we therefore consider it essential to build a computational model that can obtain a certain description of the auditory scene from sound mixtures. To emphasize its difference from sound source segregation and restoration, we call this approach *auditory scene description*. Kashino (1994) 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; he did not, however, address the issue of subsymbolic description which we deal with below.

In modeling the auditory scene description, it is important that we discuss what constitutes an appropriate *description* of audio signals. 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 or speech transcriptions consisting of text characters. Those symbols, however, fail to express non-symbolic properties such as the expressiveness of a musical performance and the prosody of spontaneous speech. 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 depend, of course, on the purpose of the description and on the use to which it will be put.

The focus of this paper is on the problem of *music scene description*—that is, auditory scene description in music—for monaural complex real-world audio signals such as those recorded on commercially distributed compact discs. The audio signals are thus assumed to contain simultaneous sounds of various instruments. This real-world-oriented approach with realistic assumptions is important to address the scaling-up problem [1] and facilitate the implementation of practical applications (Goto and Muraoka, 1996, 1998; Goto, 2001).

The main contribution of this paper is to propose a predominant-F0 estimation method that makes it possible to detect the melody and bass lines in such audio signals. On the basis of this method, a real-time system estimating the fundamental frequencies (F0s) of these lines has been implemented as a subsystem of our music-scene-description system. In the following sections, we discuss the description used in the music-scene-description system and the difficulties encountered in detecting the melody and bass lines. We then describe the algorithm of the predominant-F0 estimation method that is a core part of our system. Finally, we show experimental results obtained using our system.

2. Music-scene-description problem

Here, we explain the entire music-scene-description problem. We also explain the main difficulties

[1] As known from the scaling-up problem (Kitano, 1993) in the domain of artificial intelligence, it is hard to scale-up a system whose preliminary implementation works only in laboratory (toy-world) environments with unrealistic assumptions.

in detecting the melody and bass lines—the subproblem that we deal 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, we must decide which level is an appropriate first step toward the ultimate description in human brains. We think that the music score is inadequate for this because, as 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 an appropriate description should be:

- An intuitive description that can be easily obtained by untrained listeners.
- A basic description that trained musicians can use as a basis for higher-level music understanding.
- A useful description facilitating the development of various practical applications.

According to these requirements, we propose a description (Fig. 1) consisting of five subsymbolic representations:

(1) *Hierarchical beat structure*
Represents the fundamental temporal structure of music and comprises the quarter-note and measure levels—i.e., the positions of quarter-note beats and bar-lines.
(2) *Chord change possibility*
Represents the possibilities of chord changes and indicates how much change there is in the dominant frequency components included in chord tones and their harmonic overtones.
(3) *Drum pattern*
Represents temporal patterns of how two principal drums, a bass drum and a snare drum, are played. This representation is not used for music without drums.

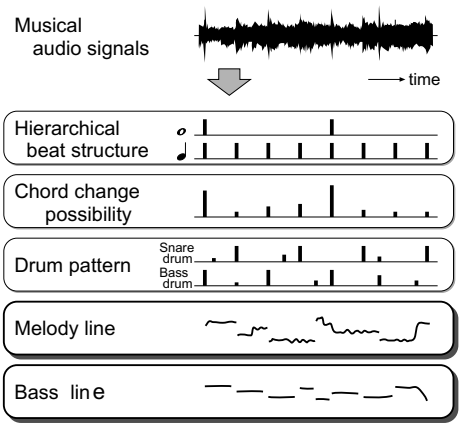

Fig. 1. Description in our music-scene-description system.

(4) *Melody line*
Represents the temporal trajectory of the melody, which is a series of single tones and is heard more distinctly than the rest. Note that this is not a series of musical notes; it is a continuous representation of frequency and power transitions.
(5) *Bass line*
Represents the temporal trajectory of the bass, which is a series of single tones and is the lowest part in polyphonic music. This is a continuous representation of frequency and power transitions.

The idea behind these representations came from introspective observation of how untrained listeners listen to music. A description consisting of the first three representations and the methods for obtaining them were described previously (Goto and Muraoka, 1994, 1996, 1998, 1999; Goto, 1998, 2001) from the viewpoint of beat tracking. [2]

In this paper we deal with the last two representations, the *melody line* and *bass line*. The detection

[2] Although the basic concept of music scene description is independent of music genres, some subsymbolic representations in Fig. 1 depend on genres in our current implementation. The hierarchical beat structure, chord change possibility, and drum pattern are obtained under the assumption that an input song is popular music and its time-signature is 4/4. On the other hand, the melody and bass lines are obtained for music, in any genre, having single-tone melody and bass lines.

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 to the tonality. These lines are fundamental to the perception of music by both trained and untrained listeners. They are also useful in various applications such as automatic transcription, automatic music indexing for information retrieval (e.g., searching for a song by singing a melody), computer participation in live human performances, musical performance analysis of outstanding recorded performances, and automatic production of accompaniment tracks for *Karaoke* or *Music Minus One* using compact discs.

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

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

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

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

2.2. Problems in detecting the melody and bass lines

It has been considered 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., 1999) have been premised upon the input audio signal containing just a single-pitch sound with aperiodic noise. Although several methods for dealing with multiple-pitch mixtures have been proposed (Parsons, 1976; Chafe and Jaffe, 1986; Katayose and Inokuchi, 1989; de Cheveigné, 1993; Brown and Cooke, 1994; Nakatani et al., 1995; Kashino and Murase, 1997; Kashino et al., 1998; de Cheveigné and Kawahara, 1999; Tolonen and Karjalainen, 2000; Klapuri, 2001), these required that the number of simultaneous sounds be assumed and had difficulty estimating the F0 in complex audio signals sampled from compact discs.

The main reason F0 estimation in sound mixtures is difficult is that in the time–frequency domain the frequency components of one sound often overlap the frequency components of simultaneous sounds. In 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 of locally tracing a frequency component is therefore neither reliable nor stable. Moreover, sophisticated F0 estimation methods relying on the existence of the F0's frequency component (the frequency component corresponding to the F0) not only cannot handle the *missing fundamental*, but are also unreliable when the F0's frequency component is smeared by the harmonics of simultaneous sounds.

Taking the above into account, the main problems in detecting the melody and bass lines can be summarized as:

(i) How to decide which F0 belongs to the melody and bass lines in polyphonic music.
(ii) How to estimate the F0 in complex sound mixtures where the number of sound sources is unknown.
(iii) How to select the appropriate F0 when several ambiguous F0 candidates are found.

3. Predominant-F0 estimation method: PreFEst

We propose a method called *PreFEst* (predominant-F0 estimation method) which makes it possible to detect the melody and bass lines in real-world sound mixtures. In solving the above problems, we make three assumptions:

- The melody and bass sounds have a harmonic structure. However, we do not care about the existence of the F0's frequency component.
- The melody line has the most predominant (strongest) 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: the F0 is likely to continue at close to the previous F0 for its duration (i.e., during a musical note).

These assumptions fit a large class of music with single-tone melody and bass lines.

PreFEst basically estimates the F0 of the most predominant harmonic structure within a limited frequency range of a sound mixture. Our solutions to the three problems mentioned above are outlined as follows: [3]

(i) The method intentionally limits the frequency range to middle- and high-frequency regions for the melody line and to a low frequency region for the bass line, and finds the F0 whose harmonics are most predominant in those ranges. In other words, whether the F0 is within the limited range or not, PreFEst tries to estimate the F0 which is supported by predominant harmonic frequency components within that range.

(ii) The method regards the observed frequency components as a weighted mixture of all possible harmonic-structure tone models without assuming the number of sound sources. It estimates their weights by using the *Expectation-Maximization (EM)* algorithm (Dempster et al., 1977), which is an iterative technique for computing maximum likelihood estimates and MAP (maximum a posteriori probability) estimates from incomplete data. The method then considers the maximum-weight model as the most predominant harmonic structure and obtains its F0. Since the above processing does not rely on the existence of the F0's frequency component, it can deal with the *missing fundamental*.

(iii) Because multiple F0 candidates are found in an ambiguous situation, the method considers their temporal continuity and selects the most dominant and stable trajectory of the F0 as the output. For this sequential F0 tracking, we introduce a multiple-agent architecture in which agents track different temporal trajectories of the F0.

While we do not intend to build a psychoacoustical model of human perception, certain psychoacoustical results may have some relevance concerning our strategy: Ritsma (1967) reported that the ear uses a rather limited spectral region in achieving 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. Note, however, that those results do not directly support our strategy since they were obtained by using the pitch of a single sound.

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. Fig. 2 shows an overview of PreFEst. The PreFEst-front-end first calculates

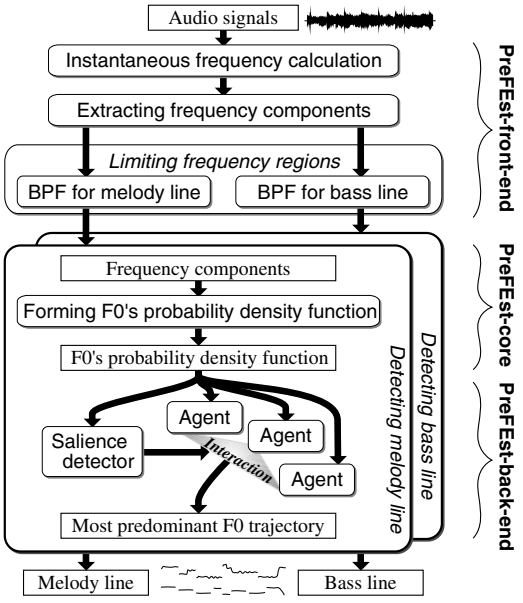

Fig. 2. Overview of PreFEst (predominant-F0 estimation method).

[3] In this paper, we do not deal with the problem of detecting the absence (activity) of melody and bass lines. PreFEst simply estimates the predominant F0 without discriminating between the sound sources.

instantaneous frequencies by using multirate signal processing techniques and extracts frequency components on the basis of an instantaneous-frequency-related measure. By using two bandpass filters (BPFs), it limits the frequency range of these components to middle and high regions for the melody line and to a low region for the bass line. The PreFEst-core then forms a probability density function (PDF) for the F0 which represents the relative dominance of every possible harmonic structure. To form the F0's PDF, it regards each set of filtered frequency components as a weighted mixture of all possible harmonic-structure tone models and then estimates their weights which can be interpreted as the F0's PDF; the maximum-weight model corresponds to the most predominant harmonic structure. This estimation is carried out using MAP estimation and the EM algorithm. Finally, in the PreFEst-back-end, multiple agents track the temporal trajectories of promising salient peaks in the F0's PDF and the output F0 is determined on the basis of the most dominant and stable trajectory.

3.1. PreFEst-front-end: forming the observed probability density functions

The PreFEst-front-end uses a multirate filter bank to obtain adequate time and frequency resolution and extracts frequency components by using an instantaneous-frequency-related measure. It obtains two sets of bandpass-filtered frequency components, one for the melody line and the other for the bass line.

3.1.1. Instantaneous frequency calculation

The PreFEst-front-end first calculates the *instantaneous frequency* (Flanagan and Golden, 1966; Boashash, 1992), the rate of change of the signal phase, of filter-bank outputs. It uses 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 $x(t)$ with a window function $h(t)$ is defined as

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

$$=a+jb, \tag{4}$$

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

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

To obtain adequate time–frequency resolution under the real-time constraint, we designed an STFT-based multirate filter bank (Fig. 3). At each level of the binary branches, the audio signal is down-sampled by a decimator that consists of an anti-aliasing filter (an 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 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 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.

3.1.2. Extracting frequency components

The extraction of frequency components is based on the mapping from the center frequency ω of an STFT filter to the instantaneous frequency

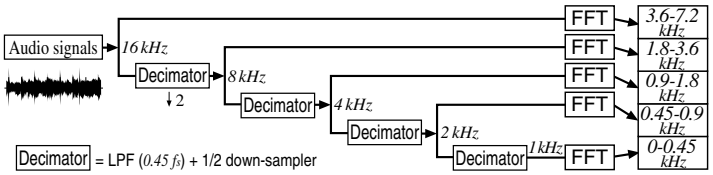

Fig. 3. Structure of the multirate filter bank.

$\lambda(\omega, t)$ of its output (Charpentier, 1986; Abe et al., 1996; Kawahara et al., 1999). If there is a frequency component at frequency ψ, that frequency is placed at the fixed point of the mapping and the instantaneous frequencies around ψ stay almost constant in the mapping (Kawahara et al., 1999). Therefore, a set $\Psi_f^{(t)}$ of instantaneous frequencies of the frequency components can be extracted by using the equation (Abe et al., 1997)

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

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

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

3.1.3. Limiting frequency regions

The frequency range is intentionally limited by using the two BPFs whose frequency responses are shown in Fig. 4. The BPF for the melody line is designed so that it covers most of the dominant harmonics of typical melody lines and deemphasizes the 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 of the dominant harmonics of typical bass lines and deemphasizes a frequency region where other parts tend to become more dominant than the bass line.

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 for the melody line (i = m) and the bass line (i = b), and x is the log-scale frequency denoted in units of *cents* (a musical-interval measurement). Frequency f_{Hz} in Hertz is converted to frequency f_{cent} in cents as follows:

$$f_{\mathrm{cent}} = 1200 \log_2 \frac{f_{\mathrm{Hz}}}{440 \times 2^{\frac{3}{12}-5}}. \tag{8}$$

There are 100 cents to a tempered semitone and 1200 to an octave. The power distribution $\Psi_p'^{(t)}(x)$ is the same as $\Psi_p^{(t)}(\omega)$ except that the frequency unit is the cent.

To enable the application of statistical methods, we represent each of the bandpass-filtered frequency components as a probability density function (PDF), called an *observed PDF*, $p_\Psi^{(t)}(x)$:

$$p_\Psi^{(t)}(x) = \frac{BPF_i(x)\Psi_p'^{(t)}(x)}{\int_{-\infty}^{\infty} BPF_i(x)\Psi_p'^{(t)}(x)\mathrm{d}x}. \tag{9}$$

3.2. 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*, $p_{\mathrm{F0}}^{(t)}(F)$, where F is the log-scale frequency in cents. We consider each observed PDF to have been generated from a weighted-mixture model of 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, we can regard these weights 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.

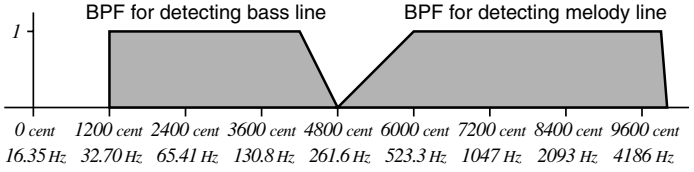

Fig. 4. Frequency responses of bandpass filters (BPFs).

Table 1
List of symbols

Symbol	Description
t	Time
x	Log-scale frequency in cents
i	The melody line (i = m) or the bass line (i = b)
$p_{\Psi}^{(t)}(x)$	Observed PDF (bandpass-filtered frequency components) (Eq. (9))
F	Fundamental frequency (F0) in cents
$p(x\|F,m,\mu^{(t)}(F,m))$	PDF of the mth tone model for each F0 F (Eq. (10))
$\mu^{(t)}(F,m)$	Shape of tone model ($\mu^{(t)}(F,m) = \{c^{(t)}(h\|F,m)\}$) (Eq. (12))
$c^{(t)}(h\|F,m)$	Relative amplitude of the hth harmonic component (shape of tone model)
M_i	The number of tone models
H_i	The number of harmonic components for tone model
W_i	Standard deviation of the Gaussian distribution for harmonic components
$p(x\|\theta^{(t)})$	Weighted-mixture model (weighted mixture of tone models) (Eq. (15))
$\theta^{(t)}$	Model parameter of $p(x\|\theta^{(t)})$ (Eq. (16))
$w^{(t)}(F,m)$	Weight of tone model $p(x\|F,m,\mu^{(t)}(F,m))$
Fl_i, Fh_i	Lower and upper limits of the possible (allowable) F0 range
$p_{\mathrm{F0}}^{(t)}(F)$	F0's PDF (Eq. (20))
$p_{0i}(\theta^{(t)})$	Prior distribution of the model parameter $\theta^{(t)}$ (Eq. (21))
$p_{0i}(w^{(t)})$	Prior distribution of the weight of tone model (Eq. (22))
$p_{0i}(\mu^{(t)})$	Prior distribution of the tone-model shapes (Eq. (23))
$w_{0i}^{(t)}(F,m)$	Most probable parameter of $w^{(t)}(F,m)$ (for $p_{0i}(w^{(t)})$)
$\mu_{0i}^{(t)}(F,m)$ $(c_{0i}^{(t)}(h \mid F,m))$	Most probable parameter of $\mu^{(t)}(F,m)$ (for $p_{0i}(\mu^{(t)})$)
$\beta_{wi}^{(t)}$	Parameter determining how much emphasis is put on $w_{0i}^{(t)}$
$\beta_{\mu i}^{(t)}(F,m)$	Parameter determining how much emphasis is put on $\mu_{0i}^{(t)}(F,m)$
$\theta'^{(t)} = \{w'^{(t)}, \mu'^{(t)}\}$	Old parameter estimate for each iteration of the EM algorithm
$\theta^{(t)} = \{w^{(t)}, \mu^{(t)}\}$	New parameter estimate for each iteration of the EM algorithm

The main symbols used in this section are listed in Table 1.

3.2.1. Weighted mixture 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 mth tone model for each F0 F is denoted by $p(x\,|\,F, m, \mu^{(t)}(F,m))$ (Fig. 5), where the model parameter $\mu^{(t)}(F,m)$ represents the shape of the tone model. The number of tone models is M_i $(1 \leqslant m \leqslant \mathrm{M}_i)$ where i denotes the melody line (i = m) or the bass line (i = b). Each tone model is defined by

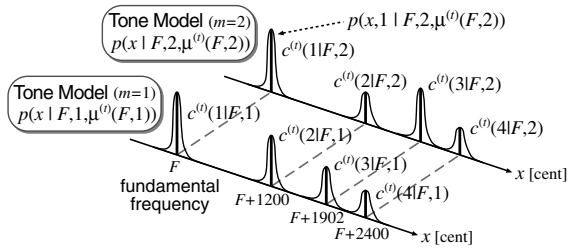

Fig. 5. Model parameters of multiple adaptive tone models.

$$p(x \mid F, m, \mu^{(t)}(F,m)) = \sum_{h=1}^{\mathrm{H}_i} p(x, h \mid F, m, \mu^{(t)}(F,m)), \tag{10}$$

$$p(x, h \mid F, m, \mu^{(t)}(F,m)) = c^{(t)}(h \mid F, m) G(x; F + 1200\log_2 h, \mathrm{W}_i), \tag{11}$$

$$\mu^{(t)}(F,m) = \{c^{(t)}(h \mid F, m) \mid h = 1, \ldots, \mathrm{H}_i\}, \tag{12}$$

$$G(x; x_0, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-x_0)^2}{2\sigma^2}\right], \tag{13}$$

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 hth harmonic component (the shape of the tone model) and satisfies

$$\sum_{h=1}^{\mathrm{H}_i} c^{(t)}(h \mid F, m) = 1. \tag{14}$$

In short, this tone model places a weighted Gaussian distribution at the position of each harmonic component. [4]

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

$$p(x \mid \theta^{(t)}) = \int_{\mathrm{Fl}_i}^{\mathrm{Fh}_i} \sum_{m=1}^{\mathrm{M}_i} w^{(t)}(F, m) \times p(x \mid F, m, \mu^{(t)}(F, m))\mathrm{d}F, \tag{15}$$

$$\theta^{(t)} = \{w^{(t)}, \mu^{(t)}\}, \tag{16}$$

$$w^{(t)} = \{w^{(t)}(F, m) \mid \mathrm{Fl}_i \leqslant F \leqslant \mathrm{Fh}_i, \; m = 1, \ldots, \mathrm{M}_i\}, \tag{17}$$

$$\mu^{(t)} = \{\mu^{(t)}(F, m) \mid \mathrm{Fl}_i \leqslant F \leqslant \mathrm{Fh}_i, \; m = 1, \ldots, \mathrm{M}_i\}, \tag{18}$$

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

$$\int_{\mathrm{Fl}_i}^{\mathrm{Fh}_i} \sum_{m=1}^{\mathrm{M}_i} w^{(t)}(F, m)\mathrm{d}F = 1. \tag{19}$$

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 F0 possibilities as expressed in Eq. (15). If we can estimate the model parameter $\theta^{(t)}$ such that the observed PDF $p_\Psi^{(t)}(x)$ is likely to have been generated from the model $p(x|\theta^{(t)})$, the weight $w^{(t)}(F, m)$ can be interpreted as the F0's PDF $p_{\mathrm{F0}}^{(t)}(F)$:

$$p_{\mathrm{F0}}^{(t)}(F) = \sum_{m=1}^{\mathrm{M}_i} w^{(t)}(F, m) \quad (\mathrm{Fl}_i \leqslant F \leqslant \mathrm{Fh}_i). \tag{20}$$

[4] Although we deal with only harmonic-structure tone models in this paper, we can also support inharmonic-structure tone models as discussed later.

3.2.2. Introducing a prior distribution

To use prior knowledge about F0 estimates and the tone-model shapes, we define a prior distribution $p_{0i}(\theta^{(t)})$ of $\theta^{(t)}$ as follows:

$$p_{0i}(\theta^{(t)}) = p_{0i}(w^{(t)}) p_{0i}(\mu^{(t)}), \tag{21}$$

$$p_{0i}(w^{(t)}) = \frac{1}{Z_w} \exp\left[-\beta_{wi}^{(t)} D_w(w_{0i}^{(t)}; w^{(t)})\right], \tag{22}$$

$$p_{0i}(\mu^{(t)}) = \frac{1}{Z_\mu} \exp\left[-\int_{\mathrm{Fl}_i}^{\mathrm{Fh}_i} \sum_{m=1}^{\mathrm{M}_i} \beta_{\mu i}^{(t)}(F, m) \times D_\mu(\mu_{0i}^{(t)}(F, m); \mu^{(t)}(F, m))\mathrm{d}F\right]. \tag{23}$$

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 $\mu_{0i}^{(t)}(F, m)$ ($c_{0i}^{(t)}(h \mid F, m)$) 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 Eqs. (22) and (23), $D_w(w_{0i}^{(t)}; w^{(t)})$ and $D_\mu(\mu_{0i}^{(t)}(F, m); \mu^{(t)}(F, m))$ are the following Kullback–Leibler information: [5]

$$D_w(w_{0i}^{(t)}; w^{(t)}) = \int_{\mathrm{Fl}_i}^{\mathrm{Fh}_i} \sum_{m=1}^{\mathrm{M}_i} w_{0i}^{(t)}(F, m) \times \log \frac{w_{0i}^{(t)}(F, m)}{w^{(t)}(F, m)}\mathrm{d}F, \tag{24}$$

$$D_\mu(\mu_{0i}^{(t)}(F, m); \mu^{(t)}(F, m)) = \sum_{h=1}^{\mathrm{H}_i} c_{0i}^{(t)}(h \mid F, m) \log \frac{c_{0i}^{(t)}(h \mid F, m)}{c^{(t)}(h \mid F, m)}. \tag{25}$$

$D_w(w_{0i}^{(t)}; w^{(t)})$ represents the closeness between $w_{0i}^{(t)}$ and $w^{(t)}$ and $D_\mu(\mu_{0i}^{(t)}(F, m); \mu^{(t)}(F, m))$ represents the closeness between $\mu_{0i}^{(t)}(F, m)$ and $\mu^{(t)}(F, m)$.

[5] We use Eqs. (24) and (25) because they are intuitive and are also convenient in the derivations that follow.

3.2.3. MAP estimation using the EM algorithm

The problem to be solved is to estimate the model parameter $\theta^{(t)}$, taking into account the prior distribution $p_{0i}(\theta^{(t)})$, when we observe $p_{\Psi}^{(t)}(x)$. The MAP (maximum a posteriori probability) estimator of $\theta^{(t)}$ is obtained by maximizing

$$\int_{-\infty}^{\infty} p_{\Psi}^{(t)}(x)(\log p(x \mid \theta^{(t)}) + \log p_{0i}(\theta^{(t)}))\mathrm{d}x. \tag{26}$$

Because this maximization problem is too difficult to solve analytically, we use the EM (Expectation-Maximization) algorithm to estimate $\theta^{(t)}$. While the EM algorithm is usually used for computing maximum likelihood estimates from incomplete observed data, it can also be used for computing MAP estimates as described in (Dempster et al., 1977). In the maximum likelihood estimation, the EM algorithm iteratively applies two steps, the *expectation step* (*E-step*) to compute the conditional expectation of the mean log-likelihood and the *maximization step* (*M-step*) to maximize its expectation. On the other hand, in the MAP estimation, the algorithm iteratively applies the E-step to compute the sum of the conditional expectation and the log prior distribution and the M-step to maximize it. With respect to $\theta^{(t)}$, each iteration updates the old estimate $\theta'^{(t)} = \{w'^{(t)}, \mu'^{(t)}\}$ to obtain the new (improved) estimate $\overline{\theta^{(t)}} = \{\overline{w^{(t)}}, \overline{\mu^{(t)}}\}$. 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 our 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, we can specify the two steps as follows:

(1) (E-step)

Compute the following $Q_{\mathrm{MAP}}(\theta^{(t)}|\theta'^{(t)})$ for the MAP estimation:

$$Q_{\mathrm{MAP}}(\theta^{(t)} \mid \theta'^{(t)}) = Q(\theta^{(t)} \mid \theta'^{(t)}) + \log p_{0i}(\theta^{(t)}), \tag{27}$$

$$Q(\theta^{(t)} \mid \theta'^{(t)}) = \int_{-\infty}^{\infty} p_{\Psi}^{(t)}(x)\mathrm{E}_{F,m,h}[\log p(x,F,m,h \mid \theta^{(t)}) \mid x, \theta'^{(t)}]\mathrm{d}x, \tag{28}$$

where $Q(\theta^{(t)}|\theta'^{(t)})$ is the conditional expectation of the mean log-likelihood for the maximum likelihood estimation. $E_{F,m,h}[a|b]$ 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 $\overline{\theta^{(t)}}$:

$$\overline{\theta^{(t)}} = \underset{\theta^{(t)}}{\operatorname{argmax}}\, Q_{\mathrm{MAP}}(\theta^{(t)} \mid \theta'^{(t)}). \tag{29}$$

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

$$Q(\theta^{(t)} \mid \theta'^{(t)}) = \int_{-\infty}^{\infty} \int_{\mathrm{Fl}_i}^{\mathrm{Fh}_i} \sum_{m=1}^{\mathrm{M}_i} \sum_{h=1}^{\mathrm{H}_i} p_{\Psi}^{(t)}(x) p(F,m,h \mid x, \theta'^{(t)}) \times \log p(x,F,m,h \mid \theta^{(t)})\mathrm{d}F\,\mathrm{d}x, \tag{30}$$

where the complete-data log-likelihood is given by

$$\log p(x,F,m,h \mid \theta^{(t)}) = \log(w^{(t)}(F,m)p(x,h \mid F,m,\mu^{(t)}(F,m))). \tag{31}$$

From Eq. (21), the log prior distribution is given by

$$\log p_{0i}(\theta^{(t)}) = -\log Z_w Z_\mu - \int_{\mathrm{Fl}_i}^{\mathrm{Fh}_i} \sum_{m=1}^{\mathrm{M}_i} \left(\beta_{wi}^{(t)} w_{0i}^{(t)}(F,m) \log \frac{w_{0i}^{(t)}(F,m)}{w^{(t)}(F,m)} + \beta_{\mu i}^{(t)}(F,m) \sum_{h=1}^{\mathrm{H}_i} c_{0i}^{(t)}(h \mid F,m) \log \frac{c_{0i}^{(t)}(h \mid F,m)}{c^{(t)}(h \mid F,m)} \right) \mathrm{d}F. \tag{32}$$

Regarding the M-step, Eq. (29) is a conditional problem of variation, where the conditions are given by Eqs. (14) and (19). This problem can be solved by using the following Euler-Lagrange differential equations with Lagrange multipliers λ_w and λ_μ:

$$\frac{\partial}{\partial w^{(t)}}\left(\int_{-\infty}^{\infty}\sum_{h=1}^{\mathrm{H}_i} p_{\Psi}^{(t)}(x)p(F,m,h\,|\,x,\theta'^{(t)})(\log w^{(t)}(F,m) + \log p(x,h\,|\,F,m,\mu^{(t)}(F,m)))\mathrm{d}x - \beta_{wi}^{(t)} w_{0i}^{(t)}(F,m)\log\frac{w_{0i}^{(t)}(F,m)}{w^{(t)}(F,m)} - \lambda_w\left(w^{(t)}(F,m) - \frac{1}{\mathrm{M}_i(\mathrm{Fh}_i - \mathrm{Fl}_i)}\right)\right) = 0, \tag{33}$$

$$\frac{\partial}{\partial c^{(t)}}\left(\int_{-\infty}^{\infty} p_{\Psi}^{(t)}(x)p(F,m,h\mid x,\theta'^{(t)})(\log w^{(t)}(F,m) + \log c^{(t)}(h\mid F,m) + \log G(x;F + 1200\log_2 h,\mathrm{W}_i))\mathrm{d}x - \beta_{\mu i}^{(t)}(F,m)c_{0i}^{(t)}(h\mid F,m)\log\frac{c_{0i}^{(t)}(h\mid F,m)}{c^{(t)}(h\mid F,m)} - \lambda_{\mu}\left(c^{(t)}(h\mid F,m) - \frac{1}{\mathrm{H}_i}\right)\right) = 0. \tag{34}$$

From these equations we get

$$w^{(t)}(F,m) = \frac{1}{\lambda_w}\left(\int_{-\infty}^{\infty} p_{\Psi}^{(t)}(x)p(F,m\mid x,\theta'^{(t)})\mathrm{d}x + \beta_{wi}^{(t)} w_{0i}^{(t)}(F,m)\right), \tag{35}$$

$$c^{(t)}(h\mid F,m) = \frac{1}{\lambda_{\mu}}\left(\int_{-\infty}^{\infty} p_{\Psi}^{(t)}(x)p(F,m,h\mid x,\theta'^{(t)})\mathrm{d}x + \beta_{\mu i}^{(t)}(F,m)c_{0i}^{(t)}(h\mid F,m)\right). \tag{36}$$

In these equations, λ_w and λ_{μ} are determined from Eqs. (14) and (19) as

$$\lambda_w = 1 + \beta_{wi}^{(t)}, \tag{37}$$

$$\lambda_{\mu} = \int_{-\infty}^{\infty} p_{\Psi}^{(t)}(x)p(F,m\mid x,\theta'^{(t)})\mathrm{d}x + \beta_{\mu i}^{(t)}(F,m). \tag{38}$$

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

$$p(F,m,h\mid x,\theta'^{(t)}) = \frac{w'^{(t)}(F,m)p(x,h\mid F,m,\mu'^{(t)}(F,m))}{p(x\mid\theta'^{(t)})}, \tag{39}$$

$$p(F,m\mid x,\theta'^{(t)}) = \frac{w'^{(t)}(F,m)p(x\mid F,m,\mu'^{(t)}(F,m))}{p(x\mid\theta'^{(t)})}. \tag{40}$$

Finally, we obtain the new parameter estimates $\overline{w^{(t)}(F,m)}$ and $\overline{c^{(t)}(h|F,m)}$:

$$\overline{w^{(t)}(F,m)} = \frac{\overline{w_{\mathrm{ML}}^{(t)}(F,m)} + \beta_{wi}^{(t)} w_{0i}^{(t)}(F,m)}{1+\beta_{wi}^{(t)}}, \tag{41}$$

$$\overline{c^{(t)}(h\mid F,m)} = \frac{\overline{w_{\mathrm{ML}}^{(t)}(F,m)}\,\overline{c_{\mathrm{ML}}^{(t)}(h\mid F,m)} + \beta_{\mu i}^{(t)}(F,m)c_{0i}^{(t)}(h\mid F,m)}{\overline{w_{\mathrm{ML}}^{(t)}(F,m)} + \beta_{\mu i}^{(t)}(F,m)}, \tag{42}$$

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

$$\overline{w_{\mathrm{ML}}^{(t)}(F,m)} = \int_{-\infty}^{\infty} p_{\Psi}^{(t)}(x) \times \frac{w'^{(t)}(F,m)p(x\mid F,m,\mu'^{(t)}(F,m))}{\int_{\mathrm{Fl}_i}^{\mathrm{Fh}_i}\sum_{v=1}^{\mathrm{M}_i} w'^{(t)}(\eta,v)p(x\mid\eta,v,\mu'^{(t)}(F,v))\mathrm{d}\eta}\mathrm{d}x, \tag{43}$$

$$\overline{c_{\mathrm{ML}}^{(t)}(h\mid F,m)} = \frac{1}{\overline{w_{\mathrm{ML}}^{(t)}(F,m)}}\int_{-\infty}^{\infty} p_{\Psi}^{(t)}(x) \times \frac{w'^{(t)}(F,m)p(x,h\mid F,m,\mu'^{(t)}(F,m))}{\int_{\mathrm{Fl}_i}^{\mathrm{Fh}_i}\sum_{v=1}^{\mathrm{M}_i} w'^{(t)}(\eta,v)p(x\mid\eta,v,\mu'^{(t)}(F,v))\mathrm{d}\eta}\mathrm{d}x. \tag{44}$$

For an intuitive explanation of Eq. (43), we call

$$\frac{w'^{(t)}(F,m)p(x\mid F,m,\mu'^{(t)}(F,m))}{\int_{\mathrm{Fl}_i}^{\mathrm{Fh}_i}\sum_{v=1}^{\mathrm{M}_i} w'^{(t)}(\eta,v)p(x\mid\eta,v,\mu'^{(t)}(F,v))\mathrm{d}\eta}$$

the decomposition filter. For the integrand on the right-hand side of Eq. (43), we can consider that, by this filter, the value of $p_{\Psi}^{(t)}(x)$ at frequency x is decomposed into (is distributed among) all possible tone models $p(x|F,m,\mu'^{(t)}(F,m))$ ($\mathrm{Fl}_i \leqslant F \leqslant \mathrm{Fh}_i$, $1 \leqslant m \leqslant \mathrm{M}_i$) in proportion to the numerator of the decomposition filter at x. The higher the weight

$w'^{(t)}(F,m)$, the larger the decomposed value given to the corresponding tone model. Note that the value of $p_{\Psi}^{(t)}(x)$ at different x is also decomposed according to a different ratio in proportion to the numerator of the decomposition filter at that x. Finally, the updated weight $w_{\mathrm{ML}}^{(t)}(F,m)$ is obtained by integrating all the decomposed values given to the corresponding mth tone model for the F0 F.

We think that this decomposition behavior is the advantage of PreFEst in comparison to previous comb-filter-based or autocorrelation-based methods (de Cheveigné, 1993; de Cheveigné and Kawahara, 1999; Tolonen and Karjalainen, 2000). This is because those previous methods cannot easily support the decomposition of an overlapping frequency component (overtone) shared by several simultaneous tones and tend to have difficulty distinguishing sounds with overlapping overtones. In addition, PreFEst can simultaneously estimate all the weights $w_{\mathrm{ML}}^{(t)}(F, m)$ (for all the range of F) so that these weights can be optimally balanced: it does not determine the weight at F after determining the weight at another F. We think this simultaneous estimation of all the weights is an advantage of PreFEst compared to previous recursive-subtraction-based methods (de Cheveigné, 1993; Klapuri, 2001) where components of the most dominant harmonic structure identified are subtracted from a mixture and then this is recursively done again starting from the residue of the previous subtraction. In those methods, once inappropriate identification or subtraction occurs, the following recursions starting from the wrong residue become unreliable.

After the above iterative computation of Eqs. (41)–(44), the F0's PDF $p_{\mathrm{F0}}^{(t)}(F)$ estimated by considering the prior distribution can be obtained from $w^{(t)}(F,m)$ according to Eq. (20). We can also obtain the tone-model shape $c^{(t)}(h|F,m)$, which is the relative amplitude of each harmonic component of all types of tone models $p(x|F,m,\mu^{(t)}(F,m))$.

A simple way to determine the frequency $F_i(t)$ of the most predominant F0 is to find the frequency that maximizes the F0's PDF $p_{\mathrm{F0}}^{(t)}(F)$:

$$F_i(t) = \operatorname*{argmax}_F p_{\mathrm{F0}}^{(t)}(F). \tag{45}$$

This result is not always stable, however, because peaks corresponding to the F0s of 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. Therefore, we have to consider the global temporal continuity of the F0 peak. This is addressed in the next section.

3.3. PreFEst-back-end: sequential F0 tracking with a multiple-agent architecture

The PreFEst-back-end sequentially tracks peak trajectories in the temporal transition of the F0's PDF to select the most dominant and stable F0 trajectory from the viewpoint of global F0 estimation. [6] To make this possible, we introduced a multiple-agent architecture that enables dynamic and flexible control of the tracking process. In an earlier multiple-agent architecture (Goto and Muraoka, 1996) the number of agents was fixed during the processing. In contrast, our new architecture generates and terminates agents dynamically by using a mechanism similar to one in the residue-driven architecture (Nakatani et al., 1995).

Our architecture consists of a salience detector and multiple agents (Fig. 6). The salience detector picks up promising salient peaks in the F0's PDF, and the agents driven by those peaks track their trajectories. They behave at each frame as follows (the first three steps correspond to the numbers in Fig. 6):

(1) After forming the F0's PDF at each frame, the salience detector picks out several salient peaks that are higher than a dynamic threshold that is adjusted according to the maximum peak. The detector then evaluates how promising each salient peak is by tracking its trajectory in the near future (at most 50 ms) taking into consideration the total power transition. For the real-time implementation, this can be done by regarding the present time as the near-future time.

[6] Because the F0's PDF is obtained without needing to assume 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.

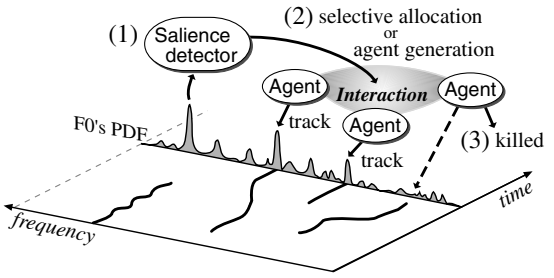

Fig. 6. Sequential F0 tracking with the multiple-agent architecture.

(2) If there are generated agents, they interact to exclusively allocate the salient peaks to agents according to the criterion of peak closeness between the peak frequency and the agent-tracking frequency. 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 its peak is generated.
(3) Each agent has an accumulated penalty. The penalty of an agent to which a salient peak has been allocated is reset. An agent to which a salient peak has not been allocated is penalized a certain value and the agent 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. An agent whose accumulated penalty exceeds a certain threshold is terminated.
(4) 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.
(5) The output F0 $F_i(t)$ 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. The power $A_i(t)$ is obtained as the total power of the harmonics of the F0 $F_i(t)$.

4. System implementation

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: computer graphics for visualization, audio signals for auralization, and continuous quantitative values (with time stamps) for use in applications. The audio-synchronized graphics output (Fig. 7) shows a window representing the scrolling F0 trajectories on a time–frequency plane (Fig. 7(b)), and adjacent interlocking windows representing the frequency components (Fig. 7(a)) and the F0's PDF for the melody and bass lines (Fig. 7(c) and (d)). The output audio signals are generated by sinusoidal synthesis on the basis of the harmonics tracked along the estimated F0.

Our current implementation for experiments uses two adaptive tone models with the parameter values listed in Table 2. Since we cannot assume perfect harmonicity in real-world audio signals, the standard deviation of the Gaussian distribution, W_m and W_b, is effective to take care of any inharmonicity of the harmonic components and its value was set according to psychoacoustical experiments (Kashino and Tanaka, 1994) on the auditory segregation of sounds with a mistuned harmonic. For the prior distribution of the tone-model shapes $\mu^{(t)}$, we use

$$c_{0i}^{(t)}(h \mid F, m) = \alpha_{i,m} g_{m,h} G(h; 1, U_i), \tag{46}$$

where m is 1 or 2, $\alpha_{i,m}$ is a normalization factor, $g_{m,h}$ is 2/3 (when $m = 2$ and h is even) or 1 (otherwise), $U_m = 5.5$, and $U_b = 2.7$. Fig. 8 shows these tone-model shapes which are invariable for all F0 ranges and for all the time. We did not use the prior distribution of $w^{(t)}$ (prior knowledge regarding rough F0 estimates). For the parameters $\beta_{wi}^{(t)}$ and $\beta_{\mu i}^{(t)}(F, m)$, we use

$$\beta_{wi}^{(t)} = 0, \tag{47}$$

$$\beta_{\mu i}^{(t)}(F, m) = B_i \exp\left[-\left(\frac{F - Fl_i}{Fh_i - Fl_i} \right)^2 \Big/ 0.2 \right], \tag{48}$$

where $B_m = 15$ and $B_b = 10$.

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

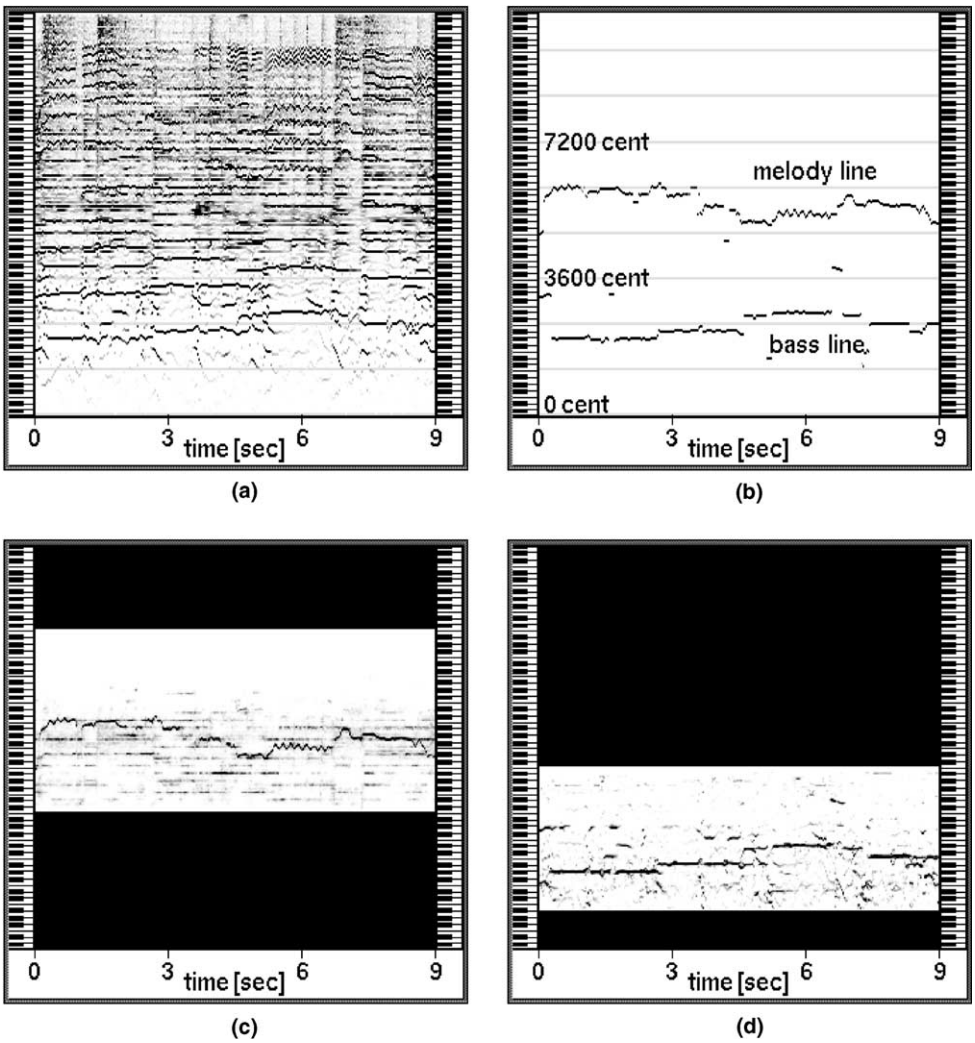

Fig. 7. Scrolling-window snapshots for a popular-music excerpt with drum sounds: (a) frequency components, (b) the corresponding melody and bass lines detected (final output), (c) the corresponding F0's PDF estimated when detecting the melody line, and (d) the corresponding F0's PDF estimated when detecting the bass line. These interlocking windows have the same vertical axis of log-scale frequency.

Table 2
Parameter values

Fh_m = 8400 cent (2093 Hz)	Fh_b = 4800 cent (261.6 Hz)
Fl_m = 3600 cent (130.8 Hz)	Fl_b = 1000 cent (29.14 Hz)
M_m = 2	M_b = 2
H_m = 16	H_b = 6
W_m = 17 cent	W_b = 17 cent

processes distributed over a LAN (Ethernet). To facilitate system expansion and application development, those processes were 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 et al., 1997). The main signal processing was done on a workstation with two Alpha21264 750-MHz CPUs (Linux 2.2), and the audio I/O and visualization processing was done on a workstation, the SGI Octane, with two R12000 300-MHz CPUs (Irix 6.5).

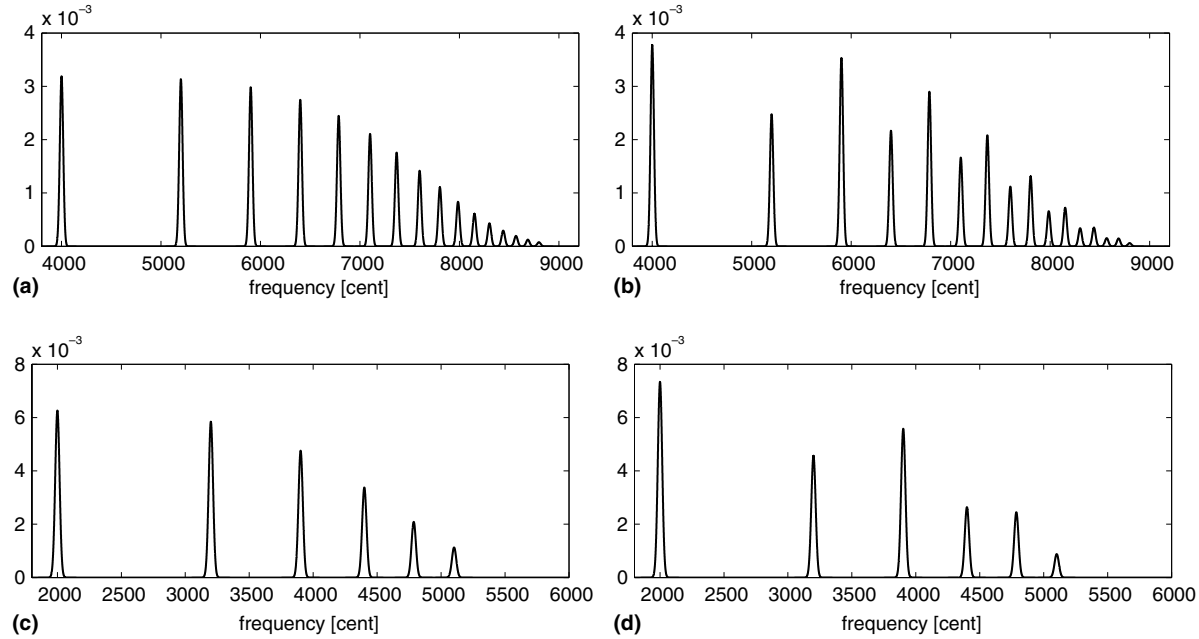

Fig. 8. Prior distribution of the tone-model shapes $p(x \mid F, m, \mu_{0i}^{(t)}(F, m))$ in our experiments: (a) for melody line (i = m, m = 1, F = 4000 cent), (b) for melody line (i = m, m = 2, F = 4000 cent), (c) for base line (i = b, m = 1, F = 2000 cent), (d) for base line (i = b, m = 2, F = 2000 cent).

5. Experimental results

The system was tested on excerpts from 10 musical pieces in the popular, jazz, and orchestral genres (Table 3). The 20-s-long input monaural audio signals—each containing a single-tone melody and the sounds of several instruments—were sampled from compact discs. We evaluated the detection rates by comparing the estimated F0s with the correct F0s that were hand-labeled using an F0 editor program we developed. This F0 editor program enables a user to determine, at each

Table 3
Detection rates for the melody and bass lines

Title	Genre	Detection rates [%]	
		Melody	Bass
My Heart Will Go On (Celine Dion)	Popular	90.8	91.2
Vision of Love (Mariah Carey)	Popular	76.6	87.3
Always (Bon Jovi)	Popular	94.2	85.4
Time Goes By (Every Little Thing)	Popular	91.6	73.1
Spirit of Love (Sing Like Talking)	Popular	90.1	76.4
Hoshi no Furu Oka (Misia)	Popular	91.7	72.1
Scarborough Fair (Herbie Hancock)	Jazz	95.3	65.8
Autumn Leaves (Julian "Cannonball" Adderley)	Jazz	82.2	82.0
On Green Dolphin Street (Miles Davis)	Jazz	92.9	80.8
Violin Con. in D, Op. 35 (Tchaikovsky)	Classical	78.7	84.8
Average		88.4	79.9

frame, the correct F0 values of the melody and bass lines while listening to the audio playback of the original as well as the harmonic structure of the currently labeled F0 while also 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 3. The system correctly detected, for most parts of each audio sample, the melody lines provided by a voice or a single-tone mid-range instrument and the bass lines provided by a bass guitar or a contrabass: the average detection rate was 88.4% for the melody line and 79.9% for the bass line. In the absence of a 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 distinguish between the sound sources. The evaluation was therefore made during periods when a hand-labeled melody or bass line was present.

Typical errors were half-F0 or double-F0 errors, errors where the detected line switched from the target part to another obbligato part for a while even as the previously tracked target part continued, and errors where a short-term trajectory near the onset of the target part was missing because of switching delay from another part to the target part. These errors were essentially due to the absence of a source-discrimination mechanism for selecting just the target part from among several simultaneous streams on the basis of sound source consistency; we plan to address this issue in a future implementation.

6. Discussion

PreFEst has great potential that we have yet to fully exploit. We discuss its future prospects with respect to the following points.

- Incorporating prior knowledge about the tone-model shapes
  While simple tone models (Fig. 8) were used for the prior distribution and were effective enough as shown by the experimental results, PreFEst allows the use of richer tone models. Many different tone models, for example, could be prepared by analyzing various kinds of harmonic structure that appear in music. Future work will also include the use of machine learning techniques to learn these tone models.
- Using more general (inharmonic structure) tone models
  Although we deal with only harmonic-structure tone models in this paper, PreFEst can be applied to any weighted mixture of arbitrary tone models (even if their components are inharmonic) by simply replacing Eq. (11) with

$$p(x, h \mid F, m, \mu^{(t)}(F, m)) = c^{(t)}(h \mid F, m) p_{\text{arbitrary}}(x; F, h, m), \tag{49}$$

  where $p_{\text{arbitrary}}(x; F, h, m)$ is an arbitrary PDF (h is merely the component number in this case). Even with this general tone model, in theory we can estimate the F0's PDF by using the same Eqs. (41)–(44). Both the harmonic-structure tone models (Eq. (11)) and any inharmonic-structure tone models (Eq. (49)) can also be used together.
- Modeling attacks of sounds, consonants, and drum sounds
  By introducing F0-independent tone models having arbitrary PDFs in addition to Eq. (49), we can extend PreFEst to deal with various F0-independent inharmonic-structure sounds—such as attacks of musical-instrument sounds, consonants in the singing voice, and drum sounds—that real-world sound mixtures usually contain. In our current implementation with only harmonic-structure tone models, the estimated F0's PDF is sometimes smeared at the frames with such short sounds. Future work will include the modeling and detection of these inharmonic-structure sounds to more precisely estimate the F0's PDF and obtain a richer description using the detection results.
- Incorporating prior knowledge regarding rough F0 estimates of the melody and bass lines
  Although prior knowledge about rough F0 estimates can be incorporated into the estimation, it was not used in our experiments. This will

be useful for some practical applications, such as the analysis of expression in a recorded performance, where a more precise F0 with fewer errors is required.

- Tracking multiple sound sources with sound-source identification
  Although multiple peaks in the F0's PDF, each corresponding to a different sound source, are tracked by multiple agents in the PreFEst-back-end, we did not fully exploit them. If they could be tracked while considering their sound source consistency by using a sound source identification method, we can investigate other simultaneous sound sources as well as the melody and bass lines. A study of integrating a sound source identification method with tone-model shapes in PreFEst will be necessary.

7. Conclusion

We have described the problem of music scene description—auditory scene description in music—and have addressed the problems regarding the detection of the melody and bass lines in complex real-world audio signals. The predominant-F0 estimation method *PreFEst* makes it possible to detect these lines by estimating the most predominant F0 trajectory. Experimental results showed that our system implementing PreFEst can estimate, in real time, the predominant F0s of the melody and bass lines in audio signals sampled from compact discs.

Our research shows that the pitch-related properties of real-world musical audio signals—like the melody and bass lines—as well as temporal properties like the hierarchical beat structure, can be described without segregating sound sources. Taking a hint from the observation that human listeners can easily listen to the melody and bass lines, we developed PreFEst to detect these lines separately by using only partial information within intentionally limited frequency ranges. Because it is generally impossible to know a priori the number of sound sources in a real-world environment, PreFEst considers all possibilities of the F0 at the same time and estimates, by using the EM algorithm, a probability density function of the F0 which represents the relative dominance of every possible harmonic structure. This approach naturally does not require the existence of the F0's frequency component and can handle the missing fundamental that often occurs in musical audio signals. 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.

In the future, we plan to work on the various extensions discussed in Section 6. While PreFEst was developed for music audio signals, it—especially the PreFEst-core—can also be applied to non-music audio signals. In fact, Masuda-Katsuse (Masuda-Katsuse, 2001; Masuda-Katsuse and Sugano, 2001) has extended it and demonstrated its effectiveness for speech recognition in realistic noisy environments.

Acknowledgments

I thank Shotaro Akaho and Hideki Asoh (National Institute of Advanced Industrial Science and Technology) for their valuable discussions. I also thank the anonymous reviewers for their helpful comments and suggestions.

References

Abe, T., Kobayashi, T., Imai, S., 1996. Robust pitch estimation with harmonics enhancement in noisy environments based on instantaneous frequency. In: Proc. Internat. Conf. on Spoken Language Processing (ICSLP 96), pp. 1277–1280.

Abe, T., Kobayashi, T., Imai, S., 1997. The IF spectrogram: a new spectral representation. In: Proc. Internat. Sympos. on Simulation, Visualization and Auralization for Acoustic Research and Education (ASVA 97), pp. 423–430.

Boashash, B., 1992. Estimating and interpreting the instantaneous frequency of a signal. Proc. IEEE 80 (4), 520–568.

Bregman, A.S., 1990. Auditory Scene Analysis: The Perceptual Organization of Sound. The MIT Press.

Brown, G.J., 1992. Computational auditory scene analysis: a representational approach. Ph.D. thesis, University of Sheffield.

Brown, G.J., Cooke, M., 1994. Perceptual grouping of musical sounds: a computational model. J. New Music Res. 23, 107–132.

Chafe, C., Jaffe, D., 1986. Source separation and note identification in polyphonic music. In: Proc. IEEE Internat. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 86), pp. 1289–1292.

Charpentier, F.J., 1986. Pitch detection using the short-term phase spectrum. In: Proc. IEEE Internat. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 86), pp. 113–116.

Cooke, M., Brown, G., 1993. Computational auditory scene analysis: exploiting principles of perceived continuity. Speech Comm. 13, 391–399.

de Cheveigné, A., 1993. Separation of concurrent harmonic sounds: fundamental frequency estimation and a time-domain cancellation model of auditory processing. J. Acoust. Soc. Amer. 93 (6), 3271–3290.

de Cheveigné, A., Kawahara, H., 1999. Multiple period estimation and pitch perception model. Speech Comm. 27 (3–4), 175–185.

Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood from incomplete data via the EM algorithm. J. Roy. Statist. Soc. B 39 (1), 1–38.

Flanagan, J.L., Golden, R.M., 1966. Phase vocoder. Bell System Tech. J. 45, 1493–1509.

Goto, M., 1998. A study of real-time beat tracking for musical audio signals. Ph.D. thesis, Waseda University (in Japanese).

Goto, M., 2001. An audio-based real-time beat tracking system for music with or without drum-sounds. J. New Music Res. 30 (2), 159–171.

Goto, M., Muraoka, Y., 1994. A beat tracking system for acoustic signals of music. In: Proc. 2nd ACM Internat. Conf. on Multimedia (ACM Multimedia 94), pp. 365–372.

Goto, M., Muraoka, Y., 1996. Beat tracking based on multiple-agent architecture—a real-time beat tracking system for audio signals. In: Proc. 2nd Internat. Conf. on Multiagent Systems (ICMAS-96), pp. 103–110.

Goto, M., Muraoka, Y., 1998. Music understanding at the beat level—real-time beat tracking for audio signals. In: Computational Auditory Scene Analysis. Lawrence Erlbaum Associates, pp. 157–176.

Goto, M., Muraoka, Y., 1999. Real-time beat tracking for drumless audio signals: chord change detection for musical decisions. Speech Comm. 27 (3–4), 311–335.

Goto, M., Neyama, R., Muraoka, Y., 1997. RMCP: remote music control protocol—design and applications. In: Proc. 1997 Internat. Computer Music Conference (ICMC 97), pp. 446–449.

Kashino, K., 1994. Computational auditory scene analysis for music signals. Ph.D. thesis, University of Tokyo (in Japanese).

Kashino, K., Murase, H., 1997. A music stream segregation system based on adaptive multi-agents. In: Proc. Internat. Joint Conf. on Artificial Intelligence (IJCAI-97), pp. 1126–1131.

Kashino, K., Tanaka, H., 1994. A computational model of segregation of two frequency components—evaluation and integration of multiple cues. Electron. Comm. Jpn. (Part III) 77 (7), 35–47.

Kashino, K., Nakadai, K., Kinoshita, T., Tanaka, H., 1998. Application of the Bayesian probability network to music scene analysis. In: Computational Auditory Scene Analysis. Lawrence Erlbaum Associates, pp. 115–137.

Katayose, H., Inokuchi, S., 1989. The kansei music system. Comput. Music J. 13 (4), 72–77.

Kawahara, H., Katayose, H., de Cheveigné, A., Patterson, R.D., 1999. Fixed point analysis of frequency to instantaneous frequency mapping for accurate estimation of F0 and periodicity. In: Proc. European Conf. on Speech Communication and Technology (Eurospeech 99), pp. 2781–2784.

Kitano, H., 1993. Challenges of massive parallelism. In: Proc. Internat. Joint Conf. on Artificial Intelligence (IJCAI-93), pp. 813–834.

Klapuri, A.P., 2001. Multipitch estimation and sound separation by the spectral smoothness principle. In: Proc. IEEE Internat. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 2001).

Masuda-Katsuse, I., 2001. A new method for speech recognition in the presence of non-stationary, unpredictable and high-level noise. In: Proc. European Conf. on Speech Communication and Technology (Eurospeech 2001), pp. 1119–1122.

Masuda-Katsuse, I., Sugano, Y., 2001. Speech estimation biased by phonemic expectation in the presence of non-stationary and unpredictable noise. In: Proc. Workshop on Consistent & Reliable Acoustic Cues for Sound Analysis (CRAC workshop).

Nakatani, T., Okuno, H.G., Kawabata, T., 1995. Residue-driven architecture for computational auditory scene analysis. In: Proc. Internat. Joint Conf. on Artificial Intelligence (IJCAI-95), pp. 165–172.

Nehorai, A., Porat, B., 1986. Adaptive comb filtering for harmonic signal enhancement. IEEE Trans. ASSP ASSP-34 (5), 1124–1138.

Noll, A.M., 1967. Cepstrum pitch determination. J. Acoust. Soc. Amer. 41 (2), 293–309.

Ohmura, H., 1994. Fine pitch contour extraction by voice fundamental wave filtering method. In: Proc. IEEE Internat. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 94), pp. II–189–192.

Okuno, H.G., Cooke, M.P. (Eds.), 1997. Working Notes of the IJCAI-97 Workshop on Computational Auditory Scene Analysis.

Parsons, T.W., 1976. Separation of speech from interfering speech by means of harmonic selection. J. Acoust. Soc. Amer. 60 (4), 911–918.

Plomp, R., 1967. Pitch of complex tones. J. Acoust. Soc. Amer. 41 (6), 1526–1533.

Rabiner, L.R., Cheng, M.J., Rosenberg, A.E., McGonegal, C.A., 1976. A comparative performance study of several pitch detection algorithms. IEEE Trans. ASSP ASSP-24 (5), 399–418.

Ritsma, R.J., 1967. Frequencies dominant in the perception of the pitch of complex sounds. J. Acoust. Soc. Amer. 42 (1), 191–198.

Rosenthal, D., Okuno, H.G. (Eds.), 1995. Working Notes of the IJCAI-95 Workshop on Computational Auditory Scene Analysis.

Rosenthal, D.F., Okuno, H.G. (Eds.), 1998. Computational Auditory Scene Analysis. Lawrence Erlbaum Associates.

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

Tolonen, T., Karjalainen, M., 2000. A computationally efficient multipitch analysis model. IEEE Trans. Speech Audio Process. 8 (6), 708–716.