Pitch Tracking Based on Statistical Anticipation

Mingyang Wu and DeLiang Wang
Department of Computer and Information Science
and Center for Cognitive Science
The Ohio State University
Columbus, OH 43210-1277, USA
Email: {mwu, dwang}@cis.ohio-state.edu

Guy J. Brown
Department of Computer Science
University of Sheffield
Regent Court, 211 Portobello Street,
Sheffield S1 4DP, UK
Email: g.brown@dcs.shef.ac.uk

Abstract

An effective multi-pitch tracking algorithm for noisy speech is critical for auditory processing. However, the performance of existing algorithms is not satisfactory. We have developed a robust algorithm for multi-pitch tracking of noisy speech based on statistical anticipation. By combining an improved channel and peak selection method, a new integration method for extracting periodicity information across the different channels, and a hidden Markov model (HMM) for forming continuous pitch tracks, our algorithm can reliably track single and double pitch tracks in a noisy environment.

1. Introduction

Determination of pitch is a fundamental problem in auditory processing. A reliable algorithm for multi-pitch contour tracking is critical for many tasks such as computational auditory scene analysis (CASA), prosody analysis, speech enhancement and recognition. However, due to the difficulty of dealing with the interference from noise intrusions and mutual interference among multiple harmonic structures, the design of such an algorithm is very challenging and most existing pitch determination algorithms (PDA) are limited to clean speech or a single pitch track in modest noise.

Among the numerous PDAs proposed, some have been specifically designed for detecting a single pitch track with voiced/unvoiced decisions in noisy speech. The majority of these algorithms (for example, see [7]) were tested on clean speech and speech mixed with different levels of white noise. Some systems also have been tested in other speech and noise conditions. For example, the system designed by Rouat et al. [11] was tested on telephone speech, vehicle speech, and speech mixed with white noise. Takagi et al. [12] also tested their single pitch track PDA on speech mixed with pink noise, music, and a male voice. In their study, however, the multi-pitch nature of the signals is ignored and a single pitch decision is given.

An ideal PDA for engineering applications should perform robustly in a variety of acoustic environments. However, the restriction to a single pitch track puts limitations on the background noise in which the PDAs are able to perform well. For example, if the noise background contains harmonic structures such as background music or voiced speech, a multi-pitch tracker is required for providing meaningful pitch tracks.

The tracking of multiple pitches also has been investigated. For examples, Gu and van Bokhoven [3] proposed an algorithm for detecting up to two pitch periods for co-channel speech separation. A model by Tolonen and Karjalainen [14] was tested on musical chords and a mixture of two vowels. Kwon et al. [8] tested their system on mixtures of two single pitch signals. Pernández-Cid and Casajús-Quirós [10] tested their system on polyphonic musical signals. However, these multi-pitch trackers were designed for and tested on clean music signals or mixtures of single-pitch signals with little or no background noise interference. Their performance on tracking speech mixed with broadband interference such as white noise is not clear.

In this paper, we propose a robust algorithm for multi-pitch tracking of noisy speech based on statistical anticipation. By using a statistical approach, the algorithm can maintain multiple hypotheses with different probabilities, making the model more robust in the presence of acoustic noise. Moreover, the modeling process incorporates the statistics extracted from a corpus of natural sound sources. Finally, a hidden Markov model (HMM) is incorporated for detecting continuous pitch tracks.

2. Model description

In this section, we first give an overview of the algorithm and stages of processing. The proposed algorithm consists of four stages. In the first stage - the front-end - the signals are filtered into channels and the envelopes in high-frequency channels are extracted. Then, the normalized correlograms [1] are computed for every channel at every 10-ms interval. Section 2.1 gives the detail of this stage.

Channel and peak selection comprises the second stage. In noisy speech, some channels are significantly corrupted by the noise. By only selecting the less corrupted channels, the robustness of the system is improved. Hunt and Lefèbvre [5] first suggested this idea, and it was implemented on mid- and high-frequency channels

(channels with center frequencies greater then 1400 Hz) by Rouat et al. [11]. We extend the channel selection idea to low-frequency channels and propose an improved method applying to all channels. Furthermore, we broaden the idea to peak selection. Generally speaking, peaks in normalized correlograms suggest periodicity of the signals. However, some peaks give misleading periodicity information and should be removed. Section 2.2 gives the detail of this stage.

The third stage is a statistical integration method for periodicity information across all channels. In the multi-band autocorrelation method, the conventional approach for integrating the periodicity information in a time frame is to summarize the autocorrelations or normalized auto-correlations across all channels. Though simple, the periodicity information contained in each channel is under-utilized. By studying the statistical relationship between the ideal pitch periods and the time lags of selected peaks obtained from the last stage, we propose a statistical integration method for producing the conditional probability of observing the signal in a time frame given a hypothesized pitch period. The relationship between ideal pitch periods and time lags of selected peaks is obtained in Section 2.3 and the integration method is described in Section 2.4.

The last stage of the algorithm is to form continuous pitch tracks using an HMM. In several studies, HMMs have been employed to model pitch track continuity. Weintraub [16] utilized a Markov model to determine whether zero, one or two pitches were present. Gu and van Bokhoven [3] used an HMM to group pitch candidates proposed by a bottom-up PDA and form continuous pitch tracks. Tokuda et al. [13] modeled pitch patterns using an HMM based on a multi-space probability distribution. In both of the studies, pitch is treated as the observation and the HMM has to be trained. In our formulation, the pitch is explicitly modeled as the hidden states and there is no training needed. Finally, the optimal pitch tracks are obtained by using the Viterbi algorithm. This stage is described in Section 2.5.

2.1. Multi-channel front-end

The input signals are sampled at 16 kHz and then passed through a bank of fourth-order "gammatone" filters [9] modeling cochlear filtering. The bandwidth of each filter is set according to its equivalent rectangular bandwidth (ERB) and we use a bank of 128 gammatone filters with center frequencies equally distributed on the ERB scale between 80 Hz to 5 kHz. After the filtering, the signals are re-aligned according to the delay of each filter.

The rest of the front-end is similar to that described by Rouat et al. [11]. The channels are classified into two categories. Channels with center frequencies lower then 800 Hz (channels 1-55) are called low-frequency channels. Others are called high-frequency channels (channels 56-128). The Teager energy operator [6] and a low-pass filter are used to extract the envelopes in high-frequency channels. The Teager energy operator is defined as $E_n = s_n^2 - s_{n+1}s_{n-1}$ for a digital signal s_n. Then, the signals are low-pass filtered at 800 Hz using the 3rd order Butterworth filter.

In order to remove the distortion due to very low frequencies, the outputs of all channels were further high-pass filtered to 64 Hz (FIR, window length of 16 ms). Then, at a given time step j, the normalized correlogram $S(c, j, \tau)$ for channel c with a time lag τ is computed by running the following normalized autocorrelation:

$$S(c, j, \tau) = \frac{\sum_{n=-N/2}^{N/2} x(c, j+n)x(c, j+n+\tau)}{\sqrt{\sum_{n=-N/2}^{N/2} x^2(c, j+n)}\sqrt{\sum_{n=-N/2}^{N/2} x^2(c, j+n+\tau)}}, \quad (1)$$

where x is the filter output.

Here, the window size is 16 ms ($N = 256$) and the normalized correlograms are computed for $\tau = 1, \ldots, 200$.

2.2. Channel and peak selection

Different methods are employed for channel and peak selection in low- and high-frequency channels since the envelopes are used in high-frequency channels.

Low frequency channels

Normalized correlograms are range limited ($-1 \le S(c, j, \tau) \le 1$) and set to 1 at zero time lag. A value of 1 at a non-zero time lag implies a perfect repetition of the signal with a certain scale factor. For a quasi-periodic signal with period T, the greater the normalized correlogram is at time lag T, the stronger the periodicity of the signal. Therefore, the maximum value of all peaks at non-zero lags measures the noise level of this channel. If the maximum value is greater than the threshold $\theta_1 = 0.945$, the channel is relatively "clean" and thus selected. Only the time lags of peaks in selected channels are included in the set of selected peaks denoted as Φ.

High frequency channels

As suggested by Rouat et al. [11], if a channel is less corrupted by noise, the original normalized correlogram computed using a window size of 16 ms and the normalized correlogram $S'(c, j, \tau)$ using a longer window size of 30 ms should have similar shapes. For every local peak of $S(c, j, \tau)$, we search for the closest local peak in $S'(c, j, \tau)$. If the difference between the two time lags is greater then 125 μs (or 2 delay steps), the channel is removed.

Two methods are employed to select peaks in a selected channel. First, for a peak suggesting true periodicity in the signal, a peak around double the time lag of the first one should be found. The second peak will be checked and if it is outside ±5 lag steps around the predicted double time lag of the first peak, the first peak is removed.

A high-frequency channel responds to multiple harmonics, and the nature of beats and combination tones dictates that the response envelope fluctuates at the fundamental frequency [4]. Therefore, the occurrence of strong peaks at time lag *T* and its multiples in a high-frequency channel suggests a fundamental period of *T*. For the second method of peak selection, if the value of the peak at the least non-zero time lag is greater than $\theta_2 = 0.6$, all the multiple peaks are removed.

This second method for peak selection is critical for eliminating multiple and sub-multiple pitch extraction error. In this type of errors, the multiples of period *d* are detected instead of detecting the real pitch period *d*. It is particularly a problem for autocorrelation based PDAs.

The selected peaks in all high frequency channels are added to Φ.

2.3. Pitch period and time lags of selected peaks

The alignment of peaks in the normalized correlograms across the channels signals a pitch period. By studying the difference between the ideal pitch period and the time lag from the closest selected peaks, we can derive the evidence of the normalized correlogram in a particular channel supporting a hypothesis of a pitch delay.

More specifically, consider channel *c*. We denote the ideal pitch period *d* and the relative time lag Δ is defined as

$$\Delta = l - d \ , \tag{2}$$

where l denotes the time lag of the closest peak.

The statistics of the relative time lag Δ are extracted from a corpus of 13 utterances of male and female speech, which is part of the sound database previously used by Cooke [2]. An "ideal" pitch track is obtained by running a correlogram-based PDA on clean speech before mixing and correction by hand. The speech signals are passed through the front-end and the channel/peak selection method described in Section 2.1 and 2.2 respectively. The statistics are collected from the selected channels across all voiced frame for every channel separately.

As an example, the histogram of relative time lags for channel 22 is shown in Fig. 1. As can be seen, the distribution is centered at zero. A mixture of a Laplacian and a uniform distribution is employed for modeling the distribution. The Laplacian represents the majority of channels "supporting" the pitch period and the uniform distribution represents the "background noise" channels,

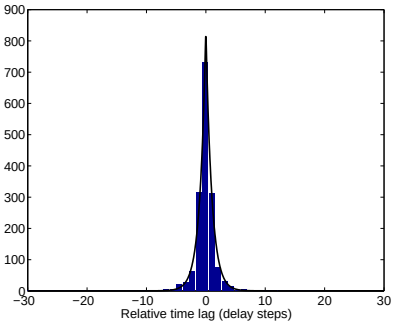

Figure 1: Histogram and estimated distribution of relative time lags for single pitch in channel 22. The bar graph represents the histogram and the solid line represents the estimated distribution.

whose peaks distribute uniformly in the background. The distribution in channel *c* is defined as

$$p_c(\Delta) = (1-q)L(\Delta;\lambda_c) + qU(\Delta;\eta_c) \ , \tag{3}$$

where $0 < q < 1$ is a partition coefficient of the mixture. The Laplacian distribution with parameter λ_c has the formula

$$L(x;\lambda_c) = \frac{1}{2\lambda_c}\exp(-\frac{|x|}{\lambda_c}) \ .$$

The uniform distribution $U(\Delta;\eta_c)$ with range η_c is fixed in a channel according to the possible range of the peak. In a low frequency channel, we set the length of the range as the wavelength of the center frequency, therefore $\eta_c = \left(-F_s/(2F_c), F_s/(2F_c)\right)$, where F_s is the sampling frequency and F_c is the center frequency of channel *c*. In high-frequency channels, however, $U(\Delta;\eta_c)$ is the uniform distribution over all possible pitch periods (between 2 ms to 12.5 ms, that is, 32 to 200 lag steps, in our system).

We also assume a linear relationship between the frequency channel index and the Laplacian distribution parameter λ_c,

$$\lambda_c = a_0 + a_1 c \ . \tag{4}$$

The maximum likelihood method was utilized to estimate the three parameters a_0, a_1, and q. Due to the different properties for low- and high-frequency channels, the parameters were estimated on each set of channels separately and the resulting parameters are shown in the upper half of Table 1. The estimated distribution of channel 22 is illustrated in Fig. 1. As can be seen, the distribution fits the histogram very well.

Likewise, similar statistics are extracted for time frames with two pitch periods. For a selected channel with signals coming from two different harmonic sources, we assumed that the energy from one of the sources is dominant. This assumption holds because otherwise, the channel is likely to be "noisy" and rejected by the selection

Table 1: Four sets of estimated model parameters. LF = low-frequency channels, HF = high-frequency channels.

	Model parameters		
	a_0	a_1	q
One pitch (LF)	1.13	–0.011	0.01
One pitch (HF)	3.17	–0.017	0.10
Two pitches (LF)	1.35	–0.013	0.03
Two pitches (HF)	4.17	–0.026	0.06

method in Section 2.2. We redefine the relative time lags as relative to the pitch period of the dominant source. The statistics are extracted from the mixtures of the 13 utterances of speech mentioned earlier. For a particular time frame and channel, the dominant source is decided by comparing the two energy values of the corresponding time frame and channel from the two speech utterances before mixing. The probability distribution of relative time lags with two pitch periods is denoted as $p'_c(\Delta)$ and has the same formulation as in Equations 3-4. Likewise, the parameters are estimated for low- and high-frequency channels separately and shown in the lower half of Table 1.

2.4. Integration of periodicity information

As noted in Tokuda et al. [13], the state space of pitch is not a discrete or continuous state space in a conventional sense. Rather, it is a union-space Ω consisting of three spaces:

$$\Omega = \Omega_0 \cup \Omega_1 \cup \Omega_2, \tag{5}$$

where Ω_0, Ω_1, Ω_2 are zero, one, and two dimensional spaces representing zero, one, and two pitches, respectively. A state in the union-space is represented as a pair $x = (y, Y)$, where $y \in R^Y$ and Y is the space index. This section derives the conditional probability $p(\Phi \mid x)$ given a pitch state x observing the set of selected peaks.

The hypothesis of a single pitch period d is considered first. For a selected channel, the closest peak relative to the period d was identified and the relative time lag denoted as $\Delta(\Phi_c, d)$, where Φ_c is the set of selected peaks in channel c.

The channel conditional probability is derived as

$$p(\Phi_c \mid x_1) = \begin{cases} p_c(\Delta(\Phi_c, d)), & \textit{if channel c selected} \\ q_1(c)U(0;\eta_c), & \textit{otherwise} \end{cases}, \tag{6}$$

where $x_1 = (d,1) \in \Omega_1$ and $q_1(c)$ is the parameter q of channel c estimated from one-pitch frames as shown in Table 1. Note here that, if a channel has not been selected, the probability of background noise would be assigned.

The channel conditional probability can be easily combined into the frame conditional probability if the mutual independence of the signals of all channels is assumed. However, the signals are usually correlated due to the wide band nature of speech signals and the assumption of independence produces very "spiky" distributions. Hence, the following formula is proposed to combine the information across the channels:

$$p(\Phi \mid x_1) \propto \sqrt[r]{\prod_{c=1}^{C} p(\Phi_c \mid x_1)}, \tag{7}$$

where $C = 128$ is the number of all channels and the parameter $r = 12$ is the smoothing factor.

Then we consider the hypothesis of two pitch periods, d_1 and d_2, corresponding to two different harmonic sources. We further assume that d_1 corresponds to the stronger source. The channels are labeled as the source of d_1 if the relative time lags are small. More specifically, channel c belongs to d_1 if $|\Delta(\Phi_c, d_1)| < \beta\lambda_c$, where $\beta = 5.0$ and λ_c denotes the Laplacian parameter for channel c calculated from Equation 4. The combined probability is defined as

$$p_2(\Phi, d_1, d_2) = \sqrt[r]{\prod_{c=1}^{C} p'_2(\Phi_c, d_1, d_2)}, \tag{8}$$

where

$$p'_2(\Phi_c, d_1, d_2) = \begin{cases} q_2(c)U(0;\eta_c) & \textit{if channel c not selected} \\ p'_c(\Delta(\Phi_c, d_1)), & \textit{if channel c belongs to } d_1 \\ \max(p'_c(\Delta(\Phi_c, d_1)), p'_c(\Delta(\Phi_c, d_2))), & \textit{otherwise} \end{cases}, \tag{9}$$

with $q_2(c)$ denotes the parameter q of channel c estimated from two-pitch frames.

The conditional probability for the time frame is the larger of assuming either d_1 or d_2 to be the stronger source:

$$p(\Phi \mid x_2) \propto \alpha_2 \max[p_2(\Phi, d_1, d_2), p_2(\Phi, d_2, d_1)], \tag{10}$$

where $x_2 = ((d_1, d_2), 2) \in \Omega_2$ and $\alpha_2 = 0.29$.

Finally, we fix the probability of zero pitch,

$$p(\Phi \mid x_0) \propto \alpha_0, \tag{11}$$

where $x_0 \in \Omega_0$ and $\alpha_0 = 2.3 \times 10^{-35}$.

2.5. Pitch tracking using an HMM

Our approach utilizes a hidden Markov model for approximating the generation process of harmonic structure in natural environments. The hidden nodes represent possible pitch states in every time frame. The observation nodes represent the set of selected peaks in each time

Table 2: Transition probabilities between state spaces of pitch.

	$\rightarrow \Omega_0$	$\rightarrow \Omega_1$	$\rightarrow \Omega_2$
Ω_0	0.8	0.2	0.0
Ω_1	0.05	0.75	0.2
Ω_2	0.0	0.2	0.8

frame. The temporal links in the Markov model represent the probabilistic pitch dynamics. The links between a hidden node and an observation node are called observation probabilities, which have been formulated in the last section representing bottom-up pitch estimation.

There are two parts of probabilistic pitch dynamics. The first part is the dynamics of a continuous pitch track. The pitch period of time frame $t+1$ was modeled as a normal distribution centered at the pitch period of time frame t with the standard deviation of $\sigma = 7.0$.

The second part is the probabilities of jumping between the state spaces of zero pitch, one pitch, and two pitch. The values of these probabilities are given in Table 2.

Finally, the state spaces of one and two pitch are discretized and the Viterbi algorithm is employed for finding the optimal sequence of states. Note here, the sequence can be a mixture of zero, one, and two pitch states.

3. Results

The algorithm has been evaluated using a corpus of 100 mixtures of speech and noise [2] commonly used for CASA research. The mixtures are obtained by mixing 10 voiced speech samples with 10 noise samples (1kHz tone, white noise, noise burst, "cocktail party" noise, rock music, siren, trill telephone, two utterances of female speech, and one utterance of male speech).

Our results show that the proposed algorithm reliably tracks pitch points in various situations, such as one speaker, speech mixed with other acoustic sources, and simultaneous multiple speakers. As examples, Fig. 2 shows our result of tracking two simultaneous utterances of a male speaker and a female speaker (signal-to-signal ratio = 9 dB). As a comparison, Fig. 3 shows the result from an existing bottom-up pitch estimation method [15]. By analyzing peak patterns of the summary autocorrelation, the bottom-up PDA detects the first pitch if there are repeating peaks of pitch period multiples on the summary autocorrelation. The second pitch is detected by using the same analysis on the residue summary autocorrelation obtained by subtracting the peaks responsible for the first pitch. Fig. 4 shows our result of tracking a mixture of a male utterance and white noise (signal-to-noise ratio = −2 dB). Note here that the white noise is very strong. As a comparison, Fig. 5 shows the result of the bottom-up pitch estimation method described above. As can be seen, the tracking of the pitch tracks in both examples has been significantly improved.

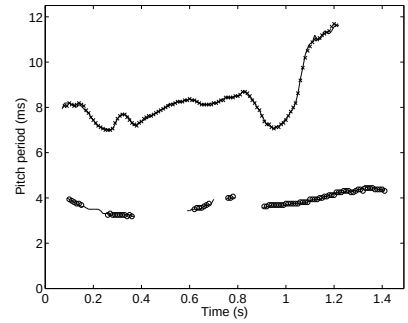

Figure 2: Result of tracking two simultaneous utterances of a male and a female speaker. The solid lines represent the hand-labeled pitch tracks estimated using one utterance before it is mixed with the other one. The '×' and 'o' tracks represent the pitch tracks estimated by our algorithm.

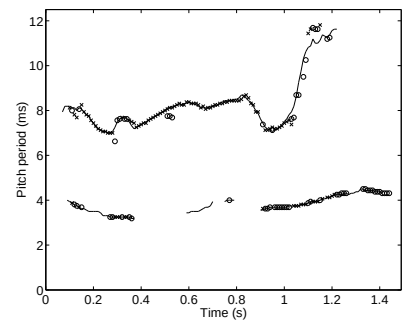

Figure 3: Result of tracking the same signal as in Fig. 2 using a bottom-up PDA. The solid lines represent the hand-labeled pitch tracks. The '×' and 'o' tracks represent the pitch tracks estimated by the bottom-up algorithm.

The results in a more systematic evaluation demonstrate that our algorithm recovers pitch tracks that match closely ideal pitch tracks.

We have also divided the 13 utterances of speech in the corpus into two separate sets, estimated the model parameters from the first set and tested our algorithm on the second set. The results show that the performance during the testing phase is very similar.

4. Discussion

Our algorithm has been shown to perform reliably for tracking single and double pitch tracks in a noisy acoustic environment. A combination of several novel ideas enables our algorithm to perform robustly. First, an improved channel and peak selection method effectively removes the corrupted channels and invalid peaks. Second, a statistical integration method utilizes the periodicity information across different channels. Finally, an HMM is employed for realizing the pitch continuity constraint.

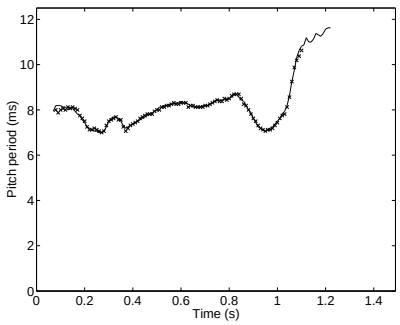

Figure 4: Result of tracking the mixture of a male utterance and white noise. The solid lines represent the hand-labeled pitch tracks estimated using the clean utterance. The '×' tracks represent the pitch tracks estimated by our algorithm.

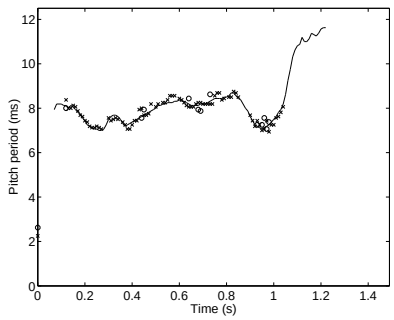

Figure 5: Result of tracking the same signal as in Fig. 4 using a bottom-up PDA. The solid lines represent the hand-labeled pitch tracks. The '×' and 'o' tracks represent the pitch tracks estimated by the bottom-up algorithm.

The probabilistic pitch dynamics defined in Section 2.5 are currently specified in our model explicitly. However, the dynamics can also be learned from natural pitch tracks and the total number of parameters hence can be significantly reduced. This will be addressed in future research. Moreover, our model can be extended to tracking more than two pitch tracks by augmenting the union-space of pitch and formulating the conditional probability of the multi-pitch states.

In CASA research, a reliable algorithm for multi-pitch tracking is critical for segregating harmonic structures, such as speech, from noise intrusions. Our system can thus provide a much needed front-end to general CASA systems, including the multistage neural model of Wang and Brown [15].

A neural network model for multi-pitch tracking could also be formed using our model as a foundation. Part of the model can already be implemented in a biologically plausible way. For examples, normalized correlograms, the rest of the front-end and the channel/peak selection method, are biologically plausible. Also, our model for bottom-up pitch estimation can be implemented as a neural network.

Acknowledgements This research was supported in part by an ONR Young Investigator Award, an NSF grant (IIS-0081058), and an AFOSR grant (F49620-01-1-0027) to DLW.

References

[1] G.J. Brown and M.P. Cooke, "Computational auditory scene analysis," *Computer Speech and Language*, vol. 8, pp. 297-336, 1994.

[2] M.P. Cooke, *Modeling Auditory Processing and Organization*, Cambridge, U.K.: Cambridge Univeristy Press, 1993.

[3] Y.H. Gu and W.M.G. van Bokhoven, "Co-channel speech separation using frequency bin non-linear adaptive filter," in *Proc. IEEE ICASSP*, 1991, pp. 949-952.

[4] H. Helmholtz, *On the Sensations of Tone as a Physiological Basis for the Theory of Music*, 1863 (Translation by A.J. Ellis, Dover Publications, 1954).

[5] M.J. Hunt and C. Lefèbvre, "Speaker dependent and independent speech recognition experiments with an auditory model," in *Proc. IEEE ICASSP*, 1988, pp. 215-218.

[6] J.F. Kaiser, "On a simple algorithm to calculate the 'energy' of a signal," in *Proc. IEEE ICASSP*, 1990, PP. 381-384.

[7] D.A. Krubsack and R.J. Niederjohn, "An autocorrelation pitch detector and voicing decision with confidence measures developed for noise-corrupted speech," *IEEE trans. Signal Process.*, vol. 39, no. 2, 1991.

[8] Y.-H. Kwon, D.-J. Park and B.-C. Ihm, "Simplified pitch detection algorithm of mixed speech signals," in *Proc. IEEE ISCAS*, 2000, pp. III-722-III-725.

[9] R.D. Patterson, I. Nimmo-Smith, J. Holdsworth, and P. Price, *APU Report 2341: An Efficient Auditory Filterbank Based on the Gammatone Function*, Cambridge: Applied Psychology Unit, 1988.

[10] P. Pernández-Cid and F.J. Casajús-Quirós, "Multi-pitch estimation for polyphonic musical signals," in *Proc. IEEE ICASSP*, 1998, pp. 3565-3568.

[11] J. Rouat, Y.C. Liu, and D. Morissette, "A pitch determination and voiced/unvoiced decision algorithm for noisy speech," *Speech Communication*, vol. 21, 191-207, 1997.

[12] T. Takagi, N. Seiyama, and E. Miyasaka, "A method for pitch extraction of speech signals using autocorrelation functions through multiple window lengths," *Electronics and Communications in Japan*, Part 3, vol. 83, no. 2, pp. 67-79, 2000.

[13] K. Tokuda, T. Masuko, N. Miyazaki, and T. Kobayashi, "Hidden Markov models based on multi-space probability distribution for pitch pattern modeling," in *Proc. IEEE ICASSP*, 1999, vol. 1, pp. 229-232.

[14] T. Tolonen and M. Karjalainen, "A computationally efficient multipitch analysis model," *IEEE trans. Speech and Audio Process.*, vol. 8, no. 6, pp. 708-716, 2000.

[15] D.L. Wang and G.J. Brown, "Separation of speech from interfering sounds based on oscillatory correlation," *IEEE trans. Neural Networks*, vol. 10, no. 3, 1999.

[16] M. Weintraub, "A computational model for separating two simultaneous talkers," in *Proc. IEEE ICASSP*, 1986, pp. 81-84.