Technical Report OSU-CISRC-12/01-TR25
Department of Computer and Information Science
The Ohio State University
Columbus, OH 43210-1277

ftp site: **ftp.cis.ohio-state.edu**
Login: **anonymous**
Directory: **pub/tech-report/2001**
File in pdf format: **TR25.pdf**
Web site: **http://www.cis.ohio-state.edu/research/tech-report.html**

A Multipitch Tracking Algorithm for Noisy Speech

Mingyang Wu, DeLiang Wang, and Guy J. Brown

***Abstract*—An effective multipitch tracking algorithm for noisy speech is critical for acoustic signal processing. However, the performance of existing algorithms is not satisfactory. In this paper, we present a robust algorithm for multipitch tracking of noisy speech. Our approach integrates an improved channel and peak selection method, a new method for extracting periodicity information across different channels, and a hidden Markov model (HMM) for forming continuous pitch tracks. The resulting algorithm can reliably track single and double pitch tracks in a noisy environment. We suggest a pitch error measure for the multipitch situation. The proposed algorithm is evaluated on a database of speech utterances mixed with various types of interference. Quantitative comparisons show that our algorithm significantly outperforms existing ones.**

***Index Terms*— Channel selection, correlogram, hidden Markov model (HMM), multipitch tracking, noisy speech, and pitch detection.**

I. Introduction

DETERMINATION of pitch is a fundamental problem in acoustic signal processing. A reliable algorithm for multipitch tracking is critical for many applications, including computational auditory scene analysis (CASA), prosody analysis, speech enhancement, speech recognition, and speaker identification (for example, see [9] [26] [38] [40]). However, due to the difficulty of dealing with noise intrusions and mutual interference among multiple harmonic structures, the design of such an algorithm has proven to be very challenging and most existing pitch determination algorithms (PDAs) are limited to clean speech or a single pitch track in modest noise.

Numerous PDAs have been proposed [13] and are generally classified into three categories: time-domain, frequency-domain and time-frequency domain algorithms. Time-domain PDAs directly examine the temporal structure of a signal waveform. Typically, peak and valley positions, zero-crossings, autocorrelations and residues of comb-filtered signals (for example, see [6]) are analyzed for detecting the pitch period. Frequency-domain PDAs distinguish the fundamental frequency by utilizing the harmonic structure in the short-term spectrum. Time-frequency domain algorithms perform time-domain analysis on band-filtered signals obtained via a multi-channel front-end.

Many PDAs have been specifically designed for detecting a single pitch track with voiced/unvoiced decisions in noisy speech. The majority of these algorithms were tested on clean speech and speech mixed with different levels of white noise (for example, see [1] [3] [18] [19] [23] [24] [33]). Some systems also have been tested in other speech and noise conditions. For example, Wang and Seneff [39] showed that their algorithm is particularly robust for telephone speech without a voiced/unvoiced decision. The system by Rouat et al. [31] was tested on telephone speech, vehicle speech, and speech mixed with white noise. Takagi et al. [34] tested their single pitch track PDA on speech mixed with pink noise, music, and a male voice. In their study, multiple pitches in the mixtures are ignored and a single pitch decision is given.

This research was supported in part by an NSF grant (IIS-0081058) and an AFOSR grant (F49620-01-1-0027).

M. Wu and D.L. Wang are with the 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).

G. J. Brown is with the Department of Computer Science, University of Sheffield, Regent Court, 211 Portobello Street, Sheffield S1 4DP, UK (email: g.brown@dcs.shef.ac.uk).

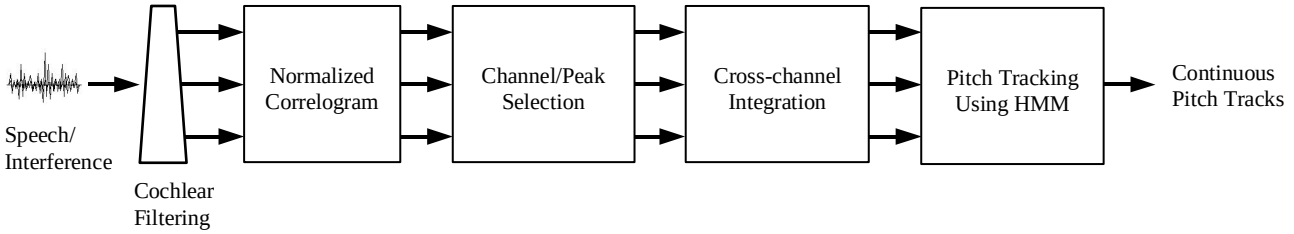

Fig. 1. Schematic diagram of the proposed model. A mixture of speech and interference is processed in four main stages. In the first stage, the normalized correlogram is obtained within each channel after the mixture is decomposed into a multi-channel representation by cochlear filtering. Channel/peak selection is performed in the second stage. In the third stage, the periodicity information is integrated across different channels using a statistical method. Finally, an HMM is utilized to form continuous pitch tracks.

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

The tracking of multiple pitches has also been investigated. For example, Gu and van Bokhoven [11] and Chazan et al. [4] proposed algorithms for detecting up to two pitch periods for co-channel speech separation. A recent model by Tolonen and Karjalainen [36] was tested on musical chords and a mixture of two vowels. Kwon et al. [20] proposed a system to segregate mixtures of two single pitch signals. Pernández-Cid and Casajús-Quirós [29] presented an algorithm to deal with polyphonic musical signals. However, these multipitch trackers were designed for and tested on clean music signals or mixtures of single-pitch signals with little or no background noise. Their performance on tracking speech mixed with broadband interference (e.g. white noise) is not clear.

In this paper, we propose a robust algorithm for multipitch tracking of noisy speech. 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. A database consisting of mixtures of speech and a variety of interfering sounds is used to evaluate the proposed algorithm, and very good performance is obtained. In addition, we have carried out quantitative comparisons with related algorithms and the results show that our model performs significantly better.

The article is organized as follows. In the next section, we give an overview of our model. Detailed explanations of the model are presented in Section III-VII. Section VIII and IX discuss model parameters and computationally efficient implementation of the proposed model. In Section X, we present the evaluation experiments and show the results. Finally, we discuss related issues in Section XI.

II. Model Overview

In this section, we first give an overview of the algorithm and stages of processing. As shown in Fig. 1, the proposed algorithm consists of four stages. In the first stage, the front-end, the signals are filtered into channels by an auditory peripheral model and the envelopes in high-frequency channels are extracted. Then, normalized correlograms [2] [38] are computed. Section III gives the details of this stage.

Channel and peak selection comprises the second stage. In noisy speech, some channels are significantly corrupted by noise. By selecting the less corrupted channels, the robustness of the system is improved. Hunt and Lefèbvre [14] first suggested this idea, and it was implemented on mid- and high-frequency channels (with center frequencies greater than 1400 Hz) by Rouat et al. [31]. We extend the channel selection idea to low-frequency channels and propose an improved method that applies to all channels. Furthermore, we employ the idea for peak selection as well. Generally speaking, peaks in normalized correlograms indicate periodicity of the signals. However, some peaks give misleading information and should be removed. Section IV gives the detail of this stage.

The third stage integrates periodicity information across all channels. Most time-frequency domain PDAs stem from Licklider's duplex model for pitch perception [22], which extracts periodicity in two steps. First, the contribution of each frequency channel to a pitch hypothesis is calculated. Then, the contributions from all channels are combined into a single score. In the multi-band autocorrelation method, the conventional approach for integrating the periodicity information in a time frame is to summate the (normalized) autocorrelations across all channels. Though simple, the periodicity information contained in each channel is under-utilized in the summary. By studying the statistical relationship between the true pitch periods and the time lags of selected peaks obtained in the previous stage, we first formulate the probability of a channel supporting a pitch hypothesis and then employ a statistical integration method for producing the conditional probability of observing the signal in a time frame given the hypothesized pitch. The

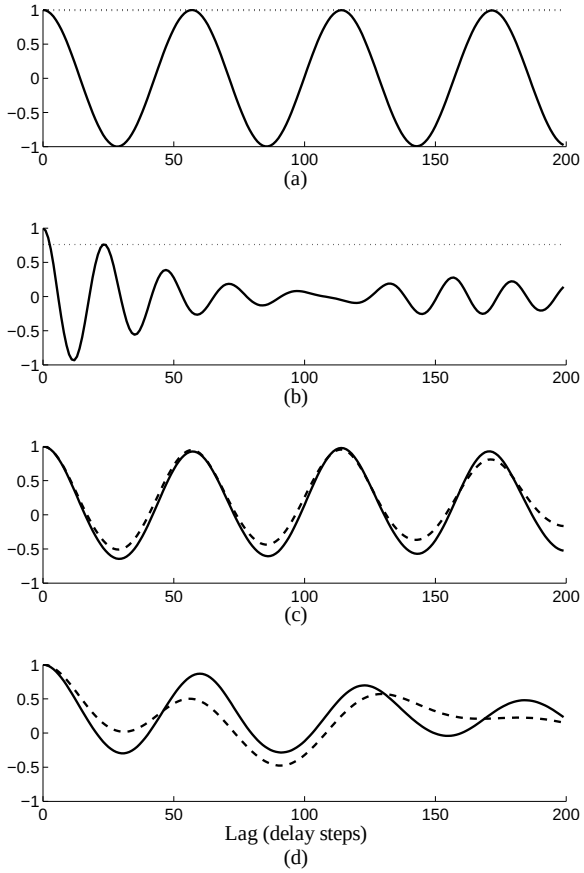

Fig. 2. Examples of normalized correlograms: (a) normalized correlogram of a clean low-frequency channel, (b) that of a noisy low-frequency channel, (c) that of a clean high-frequency channel, and (d) that of a noisy high-frequency channel. Solid lines represent the correlogram using the original time window of 16 ms and dashed lines represent the correlogram using a longer time window of 30 ms. Dotted lines indicate the maximum height of non-zero peaks. All correlograms are computed from the mixture of two simultaneous utterances of a male and a female speaker. The utterances are "Why are you all weary" and "Don't ask me to carry an oily rag like that."

relationship between true pitch periods and time lags of selected peaks is obtained in Section V and the integration method is described in Section VI.

The last stage of the algorithm is to form continuous pitch tracks using an HMM. In several previous studies, HMMs have been employed to model pitch track continuity. Weintraub [40] utilized a Markov model to determine whether zero, one or two pitches were present. Gu and van Bokhoven [11] used an HMM to group pitch candidates proposed by a bottom-up PDA and form continuous pitch tracks. Tokuda et al. [35] modeled pitch patterns using an HMM based on a multi-space probability distribution. In these studies, pitch is treated as an observation and both transition and observation probabilities of the HMM must be trained. In our formulation, pitch is explicitly modeled as hidden states and hence only transition probabilities need to be specified by extracting pitch statistics from natural speech. Finally, optimal pitch tracks are obtained by using the Viterbi algorithm. This stage is described in Section VII.

III. Multi-channel Front-end

The input signal is sampled at a rate of 16 kHz and then passed through a bank of fourth-order gammatone filters [28], which is a standard model for 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 and 5 kHz [5] [38]. 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. [31]. The channels are classified into two categories. Channels with center frequencies lower than 800 Hz (channels 1-55) are called low-frequency channels. Others are called high-frequency channels (channels 56-128). The Teager energy operator [16] 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 are further high-pass filtered to 64 Hz (FIR, window length of 16 ms). Then, at a given time step *j*, which indicates the center step of a 16 ms long time frame, the normalized correlogram $A(c, j, \tau)$ for channel *c* with a time lag τ is computed by running the following normalized autocorrelation in every 10-ms interval:

$$A(c, j, \tau) = \frac{\sum_{n=-N/2}^{N/2} r(c, j+n) r(c, j+n+\tau)}{\sqrt{\sum_{n=-N/2}^{N/2} r^2(c, j+n)} \sqrt{\sum_{n=-N/2}^{N/2} r^2(c, j+n+\tau)}}, \qquad (1)$$

where r is the filter output. Here, $N = 256$ corresponds to the 16 ms window size (one frame) and the normalized correlograms are computed for $\tau = 0, \ldots, 200$.

IV. Channel and Peak Selection

In low-frequency channels, the normalized correlograms are computed directly from filter outputs, while in high-frequency channels, they are computed from envelopes. Due to their distinct properties, separate methods are employed for channel and peak selection in the two categories of frequency channels.

A. Low-frequency Channels

Fig. 2a and 2b show the normalized correlograms in the low-frequency range for a clean and noisy channel respectively. As can be seen, normalized correlograms are range limited ($-1 \le A(c, j, \tau) \le 1$) and set to 1 at the zero time lag. A value of 1 at a non-zero time lag implies a perfect

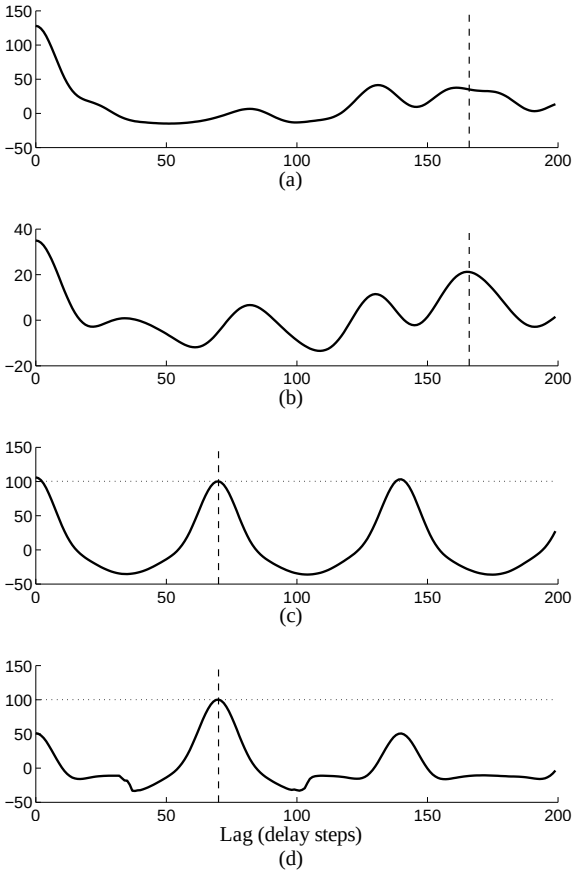

Fig. 3. (a) Summary normalized correlogram of all channels in a time frame from a speech utterance mixed with white noise. The utterance is "Why are you all weary." (b) Summary normalized correlogram of only selected channels in the same time frame as shown in (a). (c) Summary normalized correlogram of selected channels in a time frame from the speech utterance "Don't ask me to carry an oily rag like that." (d) Summary normalized correlogram of selected channels where the removed peaks are excluded in the same time frame as shown in (c). To exclude a removed peak means that the segment of correlogram between the two adjacent minima surrounding the peak is not considered. Dashed lines represent the delay corresponding to the true pitch periods. Dotted lines indicate the peak heights at pitch periods.

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 indicates the noise level of this channel. If the maximum value is greater than a threshold $\theta_1 = 0.945$, the channel is considered clean and thus selected. Only the time lags of peaks in selected channels are included in the set of selected peaks, which is denoted as Φ.

B. High-frequency Channels

As suggested by Rouat et al. [31], if a channel is not severely corrupted by noise, the original normalized correlogram computed using a window size of 16 ms and the normalized correlogram $A'(c, j, \tau)$ using a longer window size of 30 ms should have similar shapes. This is illustrated in Fig. 2c and 2d which show the normalized correlograms of a clean and a noisy channel in the high-frequency range respectively. For every local peak of $A(c, j, \tau)$, we search for the closest local peak in $A'(c, j, \tau)$. If the difference between the two corresponding time lags is greater than $\theta_2 = 2$ lag steps, the channel is removed.

Two methods are employed to select peaks in a selected channel. The first method is motivated by the observation that, for a peak suggesting true periodicity in the signal, a peak that is around the double of the time lag of the first one should be found. This second peak is thus checked and if it is outside $\theta_3 = \pm 5$ lag steps around the predicted double time lag of the first peak, the first peak is removed.

It is well known that a high-frequency channel responds to multiple harmonics, and the nature of beats and combinational tones dictates that the response envelope fluctuates at the fundamental frequency [12]. Therefore, the occurrence of strong peaks at time lag *T* and its multiples in a high-frequency channel suggests a fundamental period of *T*. In the second method of peak selection, if the value of the peak at the smallest non-zero time lag is greater than $\theta_4 = 0.6$, all of its multiple peaks are removed. The second method is critical for reducing the errors caused by multiple and sub-multiple pitch peaks in autocorrelation functions.

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

To demonstrate the effects of channel selection, Fig. 3a shows the summary normalized correlograms of a speech utterance mixed with white noise from all channels, and Fig. 3b from only selected channels. As can be seen, selected channels are much less noisy and their summary correlogram reveals the most prominent peak near the true pitch period whereas the summary correlogram of all channels fails to indicate the true pitch period. To further demonstrate the effects of peak selection, Fig. 3c shows the summary normalized correlogram of a speech utterance from selected channels, and Fig. 3d that from selected channels where removed peaks are excluded. To exclude a removed peak means that the segment of the correlogram between the two adjacent minima surrounding the peak is not considered. As can be seen, without peak selection, the height of the peak that is around double the time lag of the true pitch period is comparable or even slightly greater than the height of the peak that is around the true pitch period. With peak selection, the height of the peak at the double of the true pitch period has been significantly reduced.

V. Pitch Period and Time Lags of Selected Peaks

The alignment of peaks in the normalized correlograms across different channels signals a pitch period. By studying the difference between the true pitch period and the time lag from the closest selected peaks, we can derive the evidence of

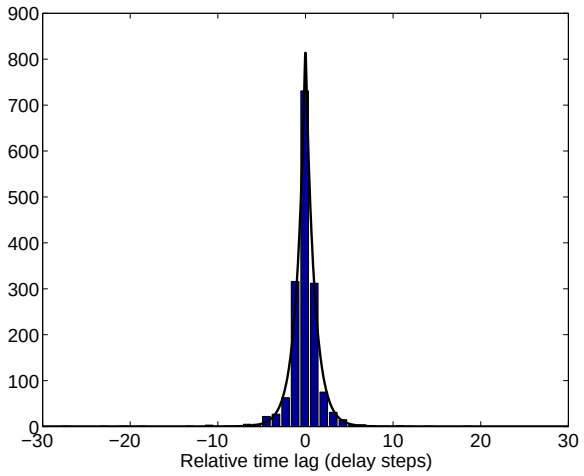

Fig. 4. Histogram and estimated distribution of relative time lags for a single pitch in channel 22. The bar graph represents the histogram and the solid line represents the estimated distribution.

the normalized correlogram in a particular channel supporting a pitch period hypothesis.

More specifically, consider channel *c*. We denote the true pitch period as *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 5 clean utterances of male and female speech, which is part of the sound mixture database collected by Cooke [5]. A true pitch track is estimated by running a correlogram-based PDA on clean speech before mixing, followed by manual correction. The speech signals are passed through the front-end and the channel/peak selection method described in Section III and IV. The statistics are collected for every channel separately from the selected channels across all voiced frames.

As an example, the histogram of relative time lags for channel 22 (center frequency: 264 Hz) is shown in Fig. 4. As can be seen, the distribution is sharply centered at zero, and can be modeled by a mixture of a Laplacian and a uniform distribution. The Laplacian represents the majority of channels "supporting" the pitch period and the uniform distribution models the "background noise" channels, 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 model. The Laplacian distribution with parameter λ_c has the formula

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

TABLE I
FOUR SETS OF ESTIMATED MODEL PARAMETERS

	Model parameters		
	a_0	a_1	q
One pitch (LF)	1.21	–0.011	0.016
One pitch (HF)	2.60	–0.008	0.063
Two pitches (LF)	1.56	–0.018	0.016
Two pitches (HF)	3.58	–0.016	0.108

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, multiple peaks may be selected and the average distance between the neighboring peaks is approximately the wavelength of the center frequency. As a result, we set the length of the range in the uniform distribution to this wavelength, that is, $\eta_c = \left(-F_s/(2F_c), F_s/(2F_c)\right)$, where F_s is the sampling frequency and F_c is the center frequency of channel *c*. In a high-frequency channel, however, ideally only one peak is selected. Therefore, $U(\delta;\eta_c)$ is the uniform distribution over all possible pitch periods. In other words, it is between 2 ms to 12.5 ms, or 32 to 200 lag steps, in our system.

The Laplacian distribution parameter λ_c and the partition parameter q can be estimated independently for each channel. However, some channels have too few data points to have accurate estimations. We observe that λ_c estimated this way decreases slowly as the channel center frequency increases. In order to have more robust and smooth estimations across all channels, we assume q to be constant across channels and a linear relationship between the frequency channel index and the Laplacian distribution parameter λ_c,

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

A maximum likelihood method is 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 I, where LF and HF indicate low- and high-frequency channels respectively. The estimated distribution of channel 22 is shown in Fig. 4. As can be seen, the distribution fits the histogram very well.

Similar statistics are extracted for time frames with two pitch periods. For a selected channel with signals coming from two different harmonic sources, we assume 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 method in Section IV. In this case, we define the relative time lags as relative to the pitch period of the dominant source. The statistics are extracted from the mixtures of the 5 speech utterances used earlier. For a particular time frame and channel, the dominant source is

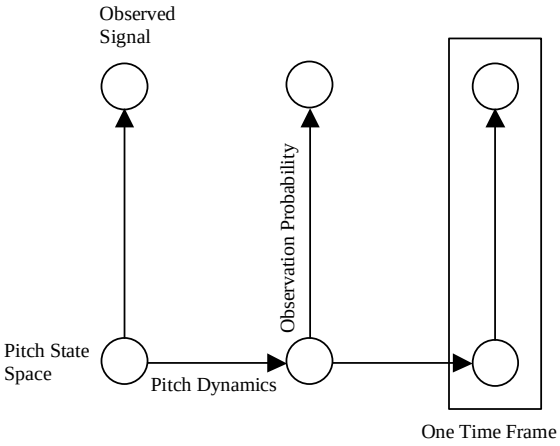

Fig. 5. Schematic diagram of an HMM for forming continuous pitch tracks. The hidden nodes represent possible pitch states in each time frame. The observation nodes represent the set of selected peaks in each frame. The temporal links in the Markov model represent the probabilistic pitch dynamics. The link between a hidden node and an observation node is called observation probability.

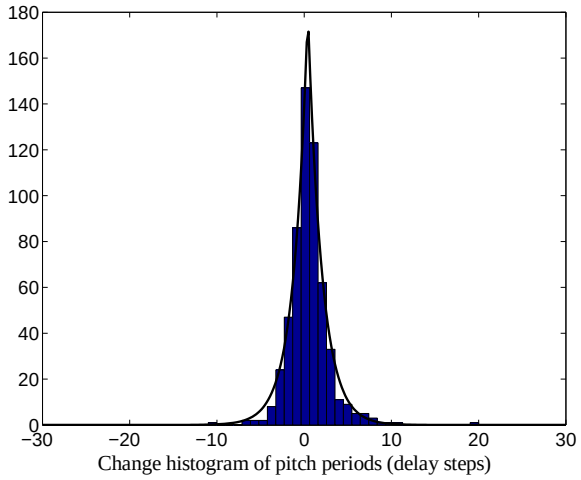

Fig. 6. Histogram and estimated distribution of pitch period changes in consecutive time frames. The bar graph represents the histogram and the solid line represents the estimated distribution.

decided by comparing the energy of 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 I. Likewise, LF and HF indicate low- and high-frequency channels respectively.

VI. Integration of Periodicity Information

As noted in Tokuda et al. [35], 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 $p(\Phi \mid x)$ of observing the set of selected peaks given a pitch state x.

The hypothesis of a single pitch period d is considered first. For a selected channel, the closest selected peak relative to the period d is identified and the relative time lag is 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

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

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 I. Note that, if a channel has not been selected, the probability of background noise is assigned.

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

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

where $C = 128$ is the number of all channels, the parameter $b = 6$ is the smoothing factor (see Section VIII for more discussion), and k is a normalization constant for probability definition.

Then we consider the hypothesis of two pitch periods, d_1 and d_2, corresponding to two different harmonic sources. Let d_1 correspond to the stronger source. The channels are labeled as the d_1 source if the relative time lags are small. More specifically, channel c belongs to the d_1 source 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

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

where

TABLE II
TRANSITION PROBABILITIES BETWEEN STATE SPACES OF PITCH

	$\to \Omega_0$	$\to \Omega_1$	$\to \Omega_2$
Ω_0	0.9250	0.0750	0.0000
Ω_1	0.0079	0.9737	0.0184
Ω_2	0.0000	0.0323	0.9677

$$p_2'(\Phi_c, d_1, d_2) = \begin{cases} q_2(c)U(0;\eta_c) & \text{if channel } c \text{ not selected} \\ p_c'(\Delta(\Phi_c, d_1)), & \text{if channel } c \text{ belongs to } d_1, \\ \max(p_c'(\Delta(\Phi_c, d_1)), p_c'(\Delta(\Phi_c, d_2))), & \text{otherwise} \end{cases} \quad (9)$$

with $q_2(c)$ denotes the parameter q of channel c estimated from two-pitch frames as shown in Table I.

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

$$p(\Phi \mid x_2) = k\alpha_2 \max[p_2(\Phi, d_1, d_2), p_2(\Phi, d_2, d_1)], \quad (10)$$

where $x_2 = ((d_1, d_2), 2) \in \Omega_2$ and $\alpha_2 = 1.7 \times 10^{-5}$.

Finally, we fix the probability of zero pitch,

$$p(\Phi \mid x_0) = k\alpha_0, \quad (11)$$

where $x_0 \in \Omega_0$ and $\alpha_0 = 2.3 \times 10^{-33}$.

VII. PITCH TRACKING USING AN HMM

We propose to use a hidden Markov model for approximating the generation process of harmonic structure in natural environments. The model is illustrated in Fig. 5. In each time frame, the hidden node indicates the pitch state space, and the observation node the observed signal. The temporal links between neighboring hidden nodes represent the probabilistic pitch dynamics. The link between a hidden node and an observation node describes observation probabilities, which have been formulated in the previous section (bottom-up pitch estimation).

Pitch dynamics have two aspects. The first is the dynamics of a continuous pitch track. The statistics of the changes of the pitch periods in consecutive time frames can be extracted from the true pitch contours of 5 speech utterances extracted earlier and their histogram is shown in Fig. 6. This is once again indicative of a Laplacian distribution. Thus, we model it by the following Laplacian distribution

$$p(\Delta) = \frac{1}{2\lambda} \exp(-\frac{|\Delta - m|}{\lambda}), \quad (12)$$

where Δ represents pitch period changes, and m and λ are distribution parameters. Using a maximum likelihood method, we have estimated that $\lambda = 2.4$ lag steps and $m = 0.4$ lag steps. A positive m indicates that, in natural speech, speech utterances have a tendency for pitch periods to increase; conversely, pitch frequencies tend to decrease. This is consistent with the declination phenomenon [27] that in natural speech pitch frequencies slowly drift down where no abrupt change in pitch occurs, which has been observed in many languages including English. The distribution is also shown in Fig. 6 and it fits the histogram very well.

The second aspect concerns jump probabilities between the state spaces of zero pitch, one pitch, and two pitches. We assume that a single speech utterance is present in the mixtures approximately half of the time and two speech utterances are present in the remaining time. The jump probabilities are estimated from the pitch tracks of the same 5 speech utterances analyzed above and the values are given in Table II.

Finally, the state spaces of one and two pitch are discretized and the standard Viterbi algorithm [15] is employed for finding the optimal sequence of states. Note that the sequence can be a mixture of zero, one, or two pitch states.

VIII. PARAMETER DETERMINATION

The frequency separating the low- and high-frequency channels is chosen according to several criteria. First, the separation frequency should be greater than possible pitch frequencies of speech, and the bandwidth of any high-frequency channels should be large enough to contain at least two harmonics of a certain harmonic structure so that amplitude modulation due to beating at the fundamental frequency is possible. Second, as long as such envelopes can be extracted, the normalized correlograms calculated from the envelopes give better indication of pitch periods than those calculated from the filtered signals directly. That is because envelope correlograms reveal pitch periods around the first peaks, whereas direct correlograms have many peaks in the range of possible pitch periods. Therefore, the separation frequency should be as low as possible so long as reliable envelopes can be extracted. By considering these criteria, we choose the separation frequency of 800 Hz.

In our model, there are a total of eight free parameters: four for channel/peak selection and four for bottom-up estimation of observation probability (their values are given). The parameters θ_1, θ_2, θ_3, and θ_4 are introduced in channel/peak selection method and they are chosen by examining the statistics from sample utterances mixed with interferences. The true pitch tracks are known for these mixtures. In every channel, the closest correlogram peak relative to the true pitch period is identified. If this peak is off from the true pitch period by more than 7 lag steps, we label this channel "noisy". Otherwise, the channel is labeled "clean". Parameter θ_1 is selected so that more than half of the noisy channels in low-frequency channels are rejected. Parameters θ_2 and θ_3 are chosen so that majority of the

TABLE III
CATEGORIZATION OF INTERFERENCE SIGNALS

	Interference signals
Category 1	White noise and noise bursts
Category 2	1 kHz tone, "cocktail party" noise, rock music, siren, and trill telephone
Category 3	Female utterance 1, male utterance and female utterance 2

noisy channels are rejected while minimizing the chance that a clean channel is rejected. Finally, parameter θ_4 is chosen so that, for almost all selected channels in high-frequency channels, the multiple peaks are removed.

Parameters β, α_0, α_2, and b are employed for bottom-up estimation of observation probability. Parameter β is used to specify the criterion for identifying the channels that belong to the dominant pitch period. It is chosen so that, in clean speech samples, almost all selected channels belong to the true pitch periods. Parameters α_0 and α_2 are employed to tune the relative strengths of the hypotheses of zero, one or two pitch periods. The smoothing factor b can be understood as tuning the relative influence of bottom-up and top-down processes. α_0, α_2, and b are optimized with respect to the combined total detection error for the training mixtures. We find that b can be chosen in a considerable range without influencing the outcome.

We note that in the preliminary version of this model [41], a different set of parameters has been employed and good results were obtained. In fact, there is a considerable range of appropriate values for these parameters, and overall system performance is not very sensitive to the specific parameter values used.

IX. EFFICIENT IMPLEMENTATION

The computational expense of the proposed algorithm can be improved significantly by employing several efficient implementations. First, a logarithm can be taken on both sides of Equation 6-11 and in the Viterbi algorithm [15]. Instead of computing multiplications and roots, which are time-consuming, only summations and divisions need to be calculated. Moreover, the number of pitch states is quite large and checking all of them using the Viterbi algorithm requires an extensive use of computational resources. Several techniques have been proposed in the literature to alleviate the computational load while achieving almost identical results [15]. 1) Pruning has been used to reduce the number of pitch states to be searched for finding the current candidates of a pitch state sequence. Since pitch tracks are continuous, the differences of pitch periods in consecutive time frames in a sequence can be restricted to a reasonable range. Therefore, only pitch periods within the range need to be searched. 2) Beam search has been employed to reduce the total number of pitch state sequences considered in evaluation. In every time frame, only a limited number of the most probable pitch state sequences are maintained and considered in the next frame. 3) The highest computational load comes from searching the pitch states corresponding to two pitch periods. In order to reduce the search effort, we only check the pitch periods in the neighborhood of the local peaks of bottom-up observation probabilities.

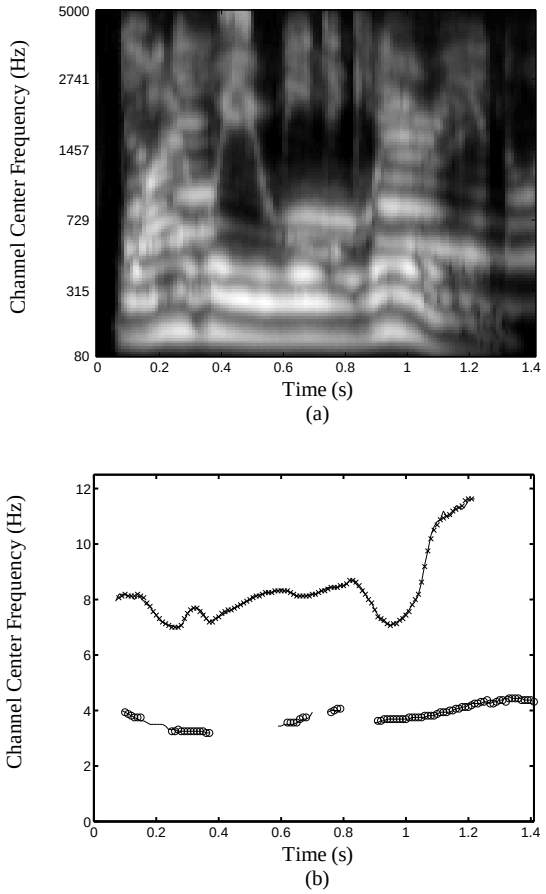

Fig. 7: (a) Time-frequency energy plot for a mixture of two simultaneous utterances of a male and a female speaker. The utterances are "Why are you all weary" and "Don't ask me to carry an oily rag like that." The brightness in a time-frequency cell indicates the energy of the corresponding gammatone filter output in the corresponding time frame. For better display, energy is plotted as the square of the logarithm. (b) Result of tracking the mixture. The solid lines indicate the true pitch tracks. The '×' and 'o' tracks represent the pitch tracks estimated by our algorithm.

By using the above efficient implementation techniques, we find that the computational load of our algorithm is drastically reduced. Meanwhile, our experiments show that the results from the original formulation and that derived for efficient implementation have negligible differences.

X. RESULTS AND COMPARISONS

A corpus of 100 mixtures of speech and interference [5], commonly used for CASA research [2] [8] [38], has been used for system evaluation and model parameter estimation. The mixtures are obtained by mixing 10 voiced utterances with 10 interference signals representing a variety of acoustic sounds. As shown in Table III, the interferences are further

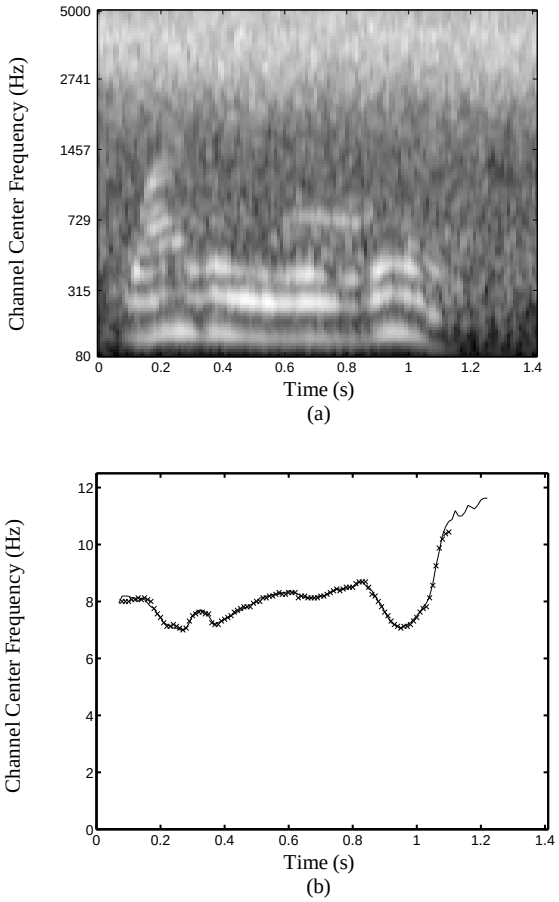

Fig. 8. (a) Time-frequency energy plot for a mixture of a male utterance and white noise. The utterance is "Why are you wary." The brightness in a time-frequency cell indicates the energy of the corresponding gammatone filter output in the corresponding time frame. For better display, energy is plotted as the square of logarithm. (b) Result of tracking the mixture. The solid lines indicate the true pitch tracks. The '×' tracks represent the pitch tracks estimated by our algorithm.

classified into three categories: 1) those with no pitch, 2) those with some pitch qualities, 3) other speech. Five speech utterances and their mixtures, which represent approximately half of the corpus, have been employed for model parameter estimation. The other half of the corpus is used for performance evaluation.

To evaluate our algorithm (or any algorithm for that matter) requires a reference pitch contour corresponding to true pitch. However, such a reference is probably impossible to obtain [13], even with instrument support [17]. Therefore, our method of obtaining reference pitch contours starts from pitch tracks computed from clean speech and is followed by a manual correction as mentioned before. Reference pitch contours obtained this way are far more accurate than those without manual correction, or those obtained from noisy speech.

To measure progress, it is important to provide a quantitative assessment of PDA performance. The guidelines for the performance evaluation of PDAs with single pitch track were established by Rabiner et al. [30]. However, there

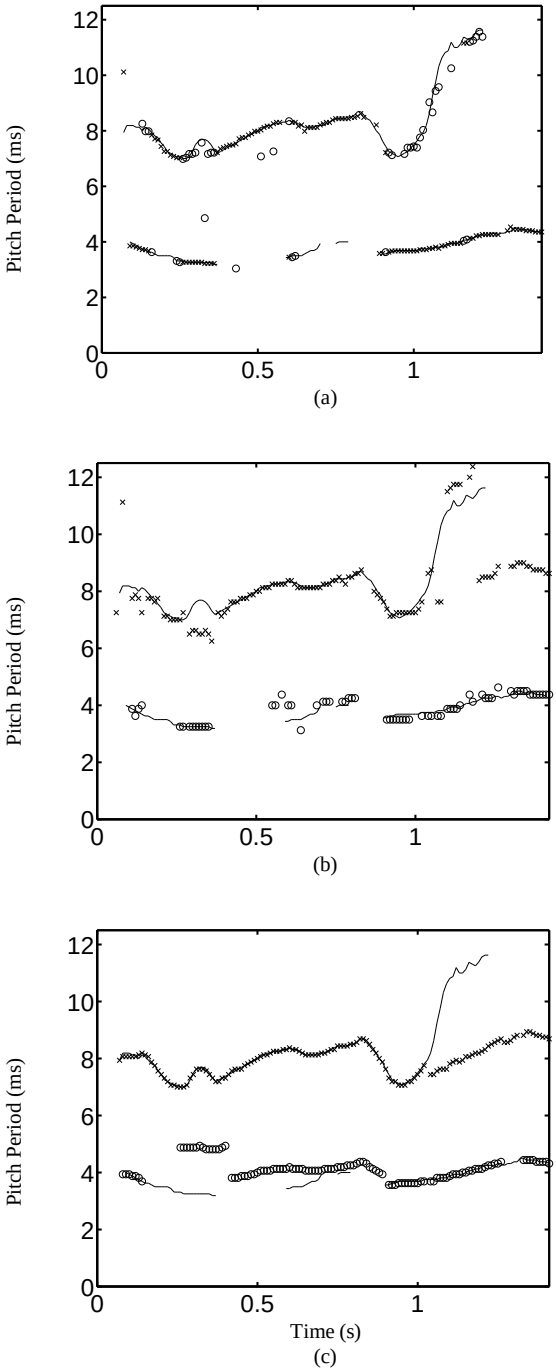

Fig. 9. Results of tracking the same signal as in Fig. 7 using (a) the TK PDA, (b) the GB PDA, and (c) the R-GB PDA. The solid lines indicate the true pitch tracks. The '×' and 'o' tracks represent the estimated pitch tracks.

are no generally accepted guidelines for multiple pitch periods that are simultaneously present. Extending the classical guidelines, we measure pitch determination errors separately for the three interference categories documented in Table III because of their distinct pitch properties. We denote $E_{x\to y}$ as the error rate of time frames where x pitch points are misclassified as y pitch points. The pitch frequency deviation Δf is calculated by

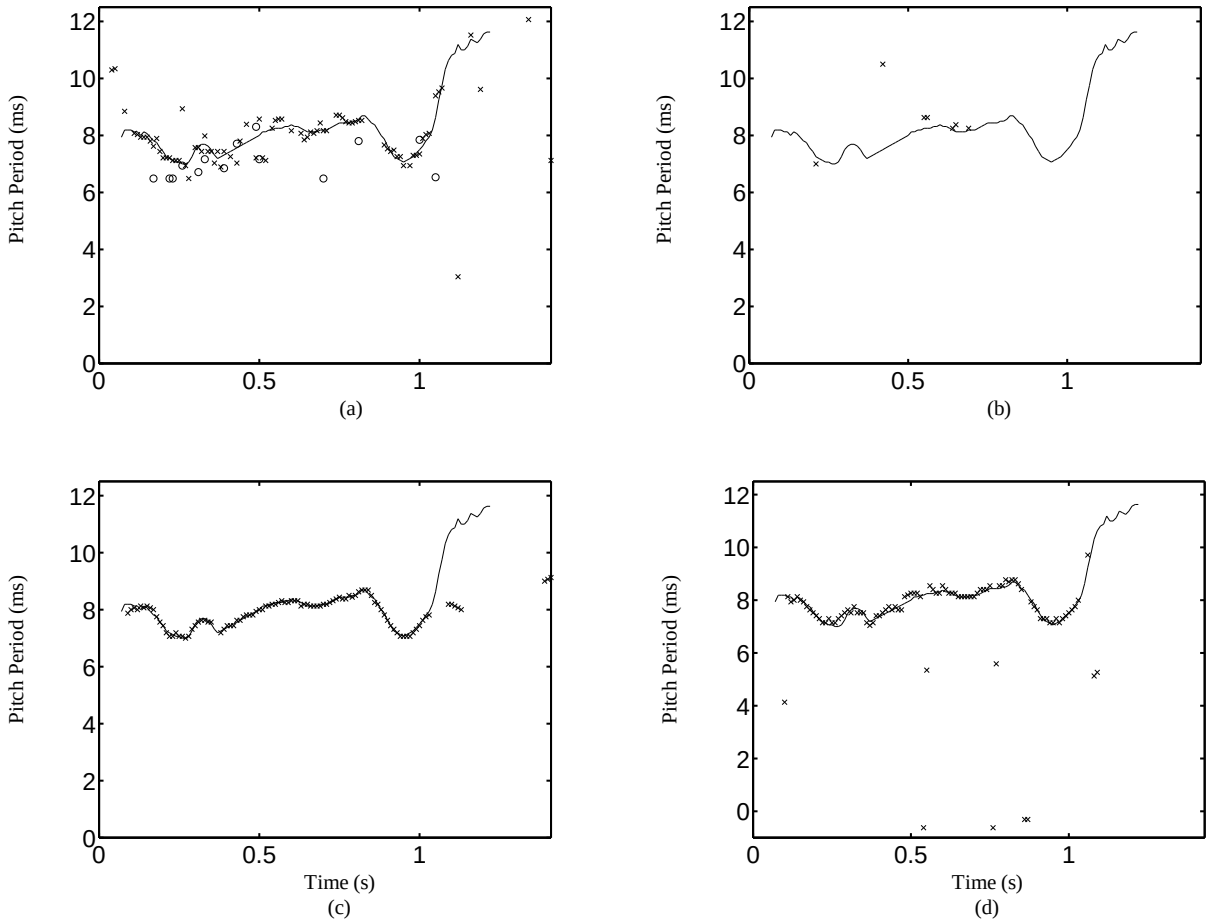

Fig. 10. Result of tracking the same signal as in Fig. 8 using (a) the TK PDA, (b) the GB PDA, (c) the R-GB PDA, and (d) the PDA proposed by Rouat et al. [31]. The solid lines indicate the true pitch tracks. The '×' and 'o' tracks represent the estimated pitch tracks. In subplot (d), time frames with negative pitch period estimates indicate the decision of voiced with unknown period.

$$\Delta f = \frac{\left| PDA_{output} - f_0 \right|}{f_0} \times 100\%, \tag{13}$$

where PDA_{output} is the closest pitch frequency estimated by the PDA to be evaluated and f_0 is the reference pitch frequency. Note that PDA_{output} may yield more than one pitch point for a particular time frame. The gross detection error rate E_{Gross} is defined as the percentage of time frames where $\Delta f > 20\%$ and the fine detection error E_{Fine} is defined as the average frequency deviation from the reference pitch contour for those time frames without gross detection errors.

For speech signals mixed with Category 1 interferences, a total gross error is indicated by

$$E_{Total} = E_{0\to 1} + E_{0\to 2} + E_{1\to 0} + E_{Gross}. \tag{14}$$

Since the main interest in many contexts is to detect the pitch contours of speech utterances, for Category 2 mixtures only $E_{1\to 0}$ is measured and the total gross error E_{Total} is indicated by the sum of $E_{1\to 0}$ and E_{Gross}. Category 3 interferences are also speech utterances and therefore all possible decision errors should be considered. For time frames with a single reference pitch, gross and fine determination errors are defined as earlier. For time frames with two reference pitches, a gross error occurs if either one exceeds the 20% limit, and a fine error is the sum of the two for two reference pitch periods. For many applications, the accuracy with which the dominating pitch is determined is of primary interest. Therefore, the total gross error E_{Gross}^{Dom} and the fine error E_{Fine}^{Dom} for dominating pitch periods are also measured.

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 two speakers. For instance, Fig. 7a shows the time-frequency energy plot for a mixture of two simultaneous utterances (a male speaker and a female speaker with signal-to-signal energy ratio = 9 dB) and Fig. 7b shows the result of tracking the mixture. As another example, Fig. 8a shows the time-frequency energy plot for a mixture of a male utterance and white noise (signal-to-noise ratio = –2 dB). Note here that the white noise is very strong. Fig. 8b shows the result of tracking the signal. In both cases, our algorithm robustly tracks either one or two pitches. Systemic performance of our algorithm for the three

TABLE IV
ERROR RATES (IN PERCENTAGE) FOR CATEGORY 1 INTERFERENCE

	$E_{0\to1}$	$E_{0\to2}$	$E_{1\to0}$	$E_{1\to2}$	E_{Gross}	E_{Total}	E_{Fine}
Proposed PDA	0.36	Nil	6.81	Nil	Nil	7.17	0.43
TK PDA	1.96	0.05	23.3	9.10	2.38	27.66	1.76
GB PDA	0.26	Nil	49.5	Nil	0.36	50.10	1.06
R-GB PDA	1.56	Nil	10.81	Nil	2.13	14.50	0.78

TABLE V
ERROR RATES (IN PERCENTAGE) FOR CATEGORY 2 INTERFERENCE

	$E_{1\to0}$	E_{Gross}	E_{Total}	E_{Fine}
Proposed PDA	3.18	0.32	3.50	0.44
TK PDA	7.70	4.53	12.23	1.41
GB PDA	22.10	2.10	24.21	2.20
R-GB PDA	5.94	4.48	10.04	0.70

TABLE VI
ERROR RATES (IN PERCENTAGE) FOR CATEGORY 3 INTERFERENCE

	$E_{0\to1}$	$E_{0\to2}$	$E_{1\to0}$	$E_{1\to2}$	$E_{2\to0}$	$E_{2\to1}$	E_{Gross}	E_{Fine}	E_{Gross}^{Dom}	E_{Fine}^{Dom}
Proposed PDA	0.68	Nil	0.88	0.16	Nil	27.08	0.21	0.33	0.93	0.21
TK PDA	0.47	0.10	2.64	4.55	1.19	26.84	2.33	0.99	4.28	0.69
GB PDA	0.41	Nil	2.65	4.20	4.20	34.54	3.89	2.04	7.70	1.34
R-GB PDA	0.57	Nil	2.28	2.78	0.57	11.80	9.09	2.11	3.63	0.53

interference categories is given in Tables IV-VI respectively. As can be seen, our algorithm achieves total gross errors of 7.17% and 3.50% for Category 1 and 2 mixtures respectively. For Category 3 interferences, a total gross error rate of 0.93% for the dominating pitch is obtained.

To put the above performance in perspective, we compare with two recent multpitch detection algorithms proposed by Tolonen and Karjalainen [36] and Gu and van Bokhoven [11]. In the Tolonen and Karjalainen model, the signal is first passed through a pre-whitening filter and then divided into two channels, below and above 1000 Hz. Generalized autocorrelations are computed in the low-frequency channel directly and those of the envelope are computed in the high-frequency channel. Then, enhanced summary autocorrelation functions are generated and the decisions on the number of pitch points as well as their pitch periods are based on the most prominent and the second most prominent peaks of such functions. We choose this study for comparison because it is a recent time-frequency domain algorithm based on a similar correlogram representation. We refer to this PDA as the TK PDA.

Gu and van Bokhoven's multpitch PDA is chosen for comparison because it is an HMM-based algorithm, and an HMM is also used in our system. The algorithm can be separated into two parts. The first part is a pseudo perceptual estimator [10] that provides coarse pitch candidates by analyzing the envelopes and carrier frequencies from the responses of a multi-channel front-end. Such pitch candidates are then fed into an HMM-based pitch contour estimator [10] for forming continuous pitch tracks. Two HMMs are trained for female and male speech utterances separately and are capable of tracking a single pitch track without voiced/unvoiced decisions at a time. In order to have voicing decisions, we add one more state representing unvoiced time frames to their original 3-state HMM. Knowing the number and types of the speech utterances presented in a mixture in advance (e.g. a mixture of a male and a female utterance) we can find the two pitch tracks by applying the male and female HMM separately. For a mixture of two male utterances, after the first male pitch track is obtained, the pitch track is subtracted from the pitch candidates and the second track is identified by applying the male HMM again. We refer to this PDA as the GB PDA.

Our experiments show that sometimes the GB PDA provides poor results, especially for speech mixed with a significant amount of white noise. Part of the problem is caused by its bottom-up pitch estimator, which is not as good as ours. To directly compare our HMM-based pitch track estimator with their HMM method, we substitute our bottom-up pitch estimator for theirs but still use their HMM model for forming continuous pitch tracks. The revised algorithm is referred as the R-GB PDA.

Fig. 9 shows the multipitch tracking results using the TK, the GB, and the R-GB PDAs, respectively, from the same mixture of Fig. 7. As can been seen, our algorithm performs significantly better than all those algorithms. Fig. 10a-c give the results of extracting pitch tracks from the same mixture of Fig. 8 using the TK, the GB, and the R-GB PDAs, respectively. As can be seen, our algorithm has much less detection error.

Quantitative comparisons are shown in Tables IV-VI. For Category 1 interferences, our algorithm has a total gross error of 7.17% while others have errors varying from 14.50% to 50.10%. The total gross error for Category 2 mixtures is 3.50% for ours, and for others it ranges from 10.04% to 24.21%. Our algorithm yields the total gross error rate of 0.93% for the dominating pitch. The corresponding error rates for the others range from 3.63% to 7.70%.

Note in Table VI that the error rate $E_{2\to1}$ of the R-GB PDA is considerably lower than ours. This, however, does not imply the R-GB PDA outperforms our algorithm. As shown in Fig. 9c, the R-GB PDA tends to mistake harmonics of the first pitch period as the second pitch period. As a result, the overall performance is much worse.

Finally, we compare our algorithm with a single-pitch determination algorithm for noisy speech proposed by Rouat

et al. [31][1]. Fig. 10d shows the result of tracking the same mixture as in Fig. 8. As can be seen, our algorithm yields less error. We do not compare with this PDA quantitatively because it is designed as a single-pitch tracker and cannot be applied to Category 2 and 3 interferences.

In summary, these results show that our algorithm outperforms the other algorithms significantly in almost all the error measures.

XI. Discussion

A common problem in PDAs is harmonic and subharmonic errors, in which the harmonics or subharmonics of a pitch are detected instead of the real pitch itself. Several techniques have been proposed to alleviate this problem. For example, a number of algorithms check sub-multiples of the time lag for the highest peak of the summary autocorrelations to ensure the detection of the real pitch period (for example, see [18]). Shimamura and Kobayashi [33] proposed a weighted autocorrelation method discounting the periodicity score of the multiples of a potential pitch period. The system by Rouat et al. [31] checks the sub-multiples of the two largest peaks in normalized summary autocorrelations and further utilizes the continuity constraint of pitch tracks to reduce these errors. Liu and Lin [23] compensate two pitch measures to reduce the scores of harmonic and subharmonic pitch periods. Medan et al. [24] disqualify such candidates by checking the normalized autocorrelation using a larger time window and pick the pitch candidate that exceeds a certain threshold and has the smallest pitch period.

In our time-frequency domain PDA, several measures contribute to alleviate these errors. First, the probabilities of subharmonic pitch periods are significantly reduced by selecting only the first correlogram peaks calculated from envelopes in high-frequency channels. Second, noisy channels tend to have random peak positions, which can reinforce harmonics or subharmonics of the real pitch. By eliminating these channels using channel selection, harmonic and subharmonic errors are greatly reduced. Third, the HMM for forming continuous pitch tracks contributes to decrease these errors.

The HMM in our model plays a similar role (utilizing pitch track continuity) as post-processing in many PDAs. Some algorithms, such as [31], employ a number of post-processing rules. These ad hoc rules introduce new free parameters. Although there are parameters in our HMM, they are learned from training samples. Also, in many algorithms (for example, see [37]), pitch tracking only considers several candidates proposed by the bottom-up algorithm and composed of peaks in bottom-up pitch scores. Our tracking mechanism considers all possible pitch hypotheses and therefore performs in a wider range of conditions.

There are several major differences in forming continuous pitch tracks between our HMM model and that of Gu and van Bokhoven [11]. Their approach is essentially for single pitch tracking while ours is for multipitch tracking. Theirs uses two different HMMs for modeling male and female speech while ours uses the same model. Their model needs to know the number and types of speech utterances in advance, and has difficulty tracking a mixture of two utterances of the same type (e.g. two male utterances). Our model does not have these difficulties.

Many models estimate multiple pitch periods by directly extending single-pitch detection methods, and they are called the one-dimensional paradigm. A common one-dimensional representation is a summary autocorrelation. Multiple pitch periods can be extracted by identifying the largest peak, the second largest peak, and so on. However, this approach is not very effective in a noisy environment, because harmonic structures often interact with each other. Cheveigné and Kawahara [7] have pointed out that a multi-step "estimate-cancel-estimate" approach is more effective. Their pitch perception model cancels the first harmonic structure using an initial estimate of the pitch, and the second pitch is estimated from the comb-filtered residue. Also, Meddis and Hewitt's [25] model of concurrent vowel separation uses a similar paradigm. A multi-dimensional paradigm is used in our model, where the scores of single and combined pitch periods are explicitly given. Interactions among the harmonic structures are formulated explicitly, and our results show that this multi-dimensional paradigm is effective for dealing with noise intrusions and mutual interference among multiple harmonic structures.

As stated previously, approximately half of the mixture database is employed for estimating (learning) relative time lag distributions in a channel (see Fig. 4) and pitch dynamics (see Fig. 6), while the other half is utilized for evaluation. It is worth emphasizing that such statistical estimations reflect general speech characteristics, not specific to either speaker or utterance. Hence, estimated distributions and parameters are expected to generalize broadly, and this is confirmed by our results. We have also tested our system on different kinds of utterance and different speakers, including digit strings from TIDigit [21], after the system is trained, and we observe equally good performance.

The proposed model can be extended to track more than two pitch periods. To do so, the union space described in Section VI would be augmented to include more than three pitch spaces. The conditional probability for the hypotheses of more than two pitch periods may be formulated using the same principles as for formulating up to two pitch periods.

In summary, we have shown that our algorithm performs reliably for tracking single and double pitch tracks in a noisy acoustic environment. A combination of several novel ideas enables the algorithm to perform well. First, an improved channel and peak selection method effectively removes corrupted channels and invalid peaks. Second, a statistical

[1] Results provided by J. Rouat.

integration method utilizes the periodicity information across different channels. Finally, an HMM realizes the pitch continuity constraint.

ACKNOWLEDGMENT

The authors thank Y.H. Gu for assisting us in understanding the details of her work, J. Rouat for providing the pitch tracking result using their PDA, and O. Fujimura for helpful discussion.

REFERENCES

[1] S. Ahmadi and A.S. Spanias, "Cepstrum-based pitch detection using a new statistical V/UV classification algorithm," *IEEE Trans. Speech Audio Processing*, vol. 7, no. 3, pp. 333-338, 1999.

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

[3] J. Cai and Z.-Q. Liu, "Robust pitch detection of speech signals using steerable filters," in *Proc. IEEE ICASSP*, 1997, vol. 2, pp. 1427-1430.

[4] D. Chazan, Y. Stettiner, and D. Malah, "Optimal multi-pitch estimation using the EM algorithm for co-channel speech separation", *in Proc. IEEE ICASSP*, 1993, pp. II-728-II-731.

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

[6] A. de Cheveigné, "Separation of concurrent harmonic sounds: Fundamental frequency estimation and a time-domain cancellation model of auditory processing," *J. Acoust. Soc. Am.*, vol. 93, pp. 3271-3290, 1993.

[7] A. de Cheveigné and H. Kawahara, "Multiple period estimation and pitch perception model," *Speech Communication*, vol. 27, pp. 175-185, 1999.

[8] L.A. Drake, "Sound source separation via computational auditory scene analysis (CASA)-enhanced beamforming," *Ph.D. Dissertation*, Northwestern University Department of Electrical Engineering, 2001.

[9] B. Gold and N. Morgan, *Speech and audio signal processing*, New York: John Wiley & Sons, 2000.

[10] Y.H. Gu, "Linear and nonlinear adaptive filtering and their application to speech intelligibility enhancement," *Ph.D. Dissertation*, Department of Electrical Engineering, Eindhoven University of Technology, Eindhoven, The Netherlands, 1992.

[11] 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.

[12] 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).

[13] W.J. Hess, *Pitch Determination of Speech Signals*, New York: Springer, 1983.

[14] 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.

[15] F. Jelinek, *Statistical methods for speech recognition*, Cambridge, Massachusetts U.S.A.: The MIT Press, 1997.

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

[17] A.K. Krishnamurthy and D.G. Childers, "Two-channel speech analysis," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 34, pp.730-743, 1986.

[18] 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 Processing*, vol. 39, no. 2, pp. 319-329, 1991.

[19] N. Kunieda, T. Shimamura, and J. Suzuki, "Pitch extraction by using autocorrelation function on the log spectrum", *Electronics and Communications in Japan*, Part 3, vol. 83, no. 1, pp. 90-98, 2000.

[20] 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.

[21] R.G. Leonard, "A database for speaker-independent digit recognition," in *Proc. IEEE ICASSP*, 1984, pp. 111-114.

[22] J.D.R. Licklider, "A duplex theory of pitch perception," *Experientia*, vol. 7, pp. 128-134, 1951.

[23] D.J. Liu and C.T. Lin, "Fundamental frequency estimation based on the joint time-frequency analysis of harmonic spectral structure," *IEEE Trans. Speech Audio Processing*, vol. 9, no. 6, pp. 609-621, 2001.

[24] Y. Medan, E. Yair, and D. Chazan, "Super resolution pitch determination of speech signals", *IEEE Trans. Signal Processing*, vol. 39, no. 1, pp. 40-48, 1991.

[25] R. Meddis and M.J. Hewitt, "Modeling the identification of concurrent vowels with different fundamental frequencies," *J. Acoust. Soc. Am.*, vol. 91, no. 1, pp. 233-244, 1992.

[26] H. Niemann, E. Nöth, A. Keiβling, R. Batliner, "Prosodic processing and its use in Verbmobil", in *Proc. IEEE ICASSP*, 1997, pp. 75-78.

[27] S. Nooteboom, "The prosody of speech: melody and rhythm," in *The Handbook of Phonetic Science*, W.J. Hardcastle and J. Laver (Eds.), Cambridge, MA: Blackwell Publishers, pp. 640-673, 1997.

[28] 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.

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

[30] L.R. Rabiner, M.J. Cheng, A.E. Rosenberg, and A. McGonegal, "A comparative study of several pitch detection algorithms," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 23, pp. 552-557, 1976.

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

[32] S.A. Shedied, M.E. Gadalah, and H.F. VanLandingham, "Pitch estimator for noisy speech signals", in *Proc. IEEE Int'l Conf. Sys., Man, Cybern.*, 2000, pp. 97-100.

[33] T. Shimamura and J. Kobayashi, "Weighted autocorrelation for pitch extraction of noisy speech," *IEEE Trans. Speech Audio Processing*, vol. 9, no. 7, pp. 727-730, 2001.

[34] 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.

[35] 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.

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

[37] L.M. Van Immerseel and J-P Martens, "Pitch and voiced/unvoiced determination with an auditory model," *J. Acoust. Soc. Am.*, vol. 91, no. 6, pp. 3511-3526, 1992.

[38] 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, pp. 684-697, 1999.

[39] C. Wang and S. Seneff, "Robust pitch tracking for prosodic modeling in telephone speech," in *Proc. IEEE ICASSP*, 2000, pp. 1343-1346.

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

[41] M. Wu, D.L. Wang and G.J. Brown, "Pitch tracking based on statistical anticipation," in *Proc. IJCNN*, 2001, vol. 2, pp. 866-871.