

# HMM-Based Multipitch Tracking for Noisy and Reverberant Speech

Zhaozhang Jin, *Student Member, IEEE*, and DeLiang Wang, *Fellow, IEEE*

**Abstract**—Multipitch tracking in real environments is critical for speech signal processing. Determining pitch in reverberant and noisy speech is a particularly challenging task. In this paper, we propose a robust algorithm for multipitch tracking in the presence of both background noise and room reverberation. An auditory front-end and a new channel selection method are utilized to extract periodicity features. We derive pitch scores for each pitch state, which estimate the likelihoods of the observed periodicity features given pitch candidates. A hidden Markov model integrates these pitch scores and searches for the best pitch state sequence. Our algorithm can reliably detect single and double pitch contours in noisy and reverberant conditions. Quantitative evaluations show that our approach outperforms existing ones, particularly in reverberant conditions.

**Index Terms**—Hidden Markov model (HMM) tracking, multipitch tracking, pitch detection algorithm (PDA), room reverberation.

## I. INTRODUCTION

PITCH determination is a fundamental problem that attracts much attention in speech analysis. A robust pitch detection algorithm (PDA) is needed for many applications including computational auditory scene analysis (CASA), prosody analysis, speech enhancement/separation, speech recognition, and speaker identification. Designing such an algorithm is challenging due to harmonic distortions brought about by acoustic interference and room reverberation.

Numerous PDAs have been developed to detect a single pitch track under clean or modestly noisy conditions (see [6] for a review). The presumption of a single pitch track, however, puts limitations on the background noise in which PDAs perform. A multipitch tracker is required when the interfering sound also contains harmonic structure (e.g., background music or another voice). A number of studies have investigated detecting multiple pitches simultaneously. Tolonen and Karjalainen [27] designed a two-channel multipitch analyzer with an enhanced

summary autocorrelation function. Wu *et al.* [32] modeled pitch period statistics on top of a channel selection mechanism and used a hidden Markov model (HMM) for extracting continuous pitch contours. Bach and Jordan [2] presented a model based on direct probabilistic modeling of the spectrogram of the signal using a factorial HMM for characterizing pitch. More recently, the mixture power spectrum was modeled as a sum of parametric source models that were trained from the voiced parts of speech [23]. Klapuri [15] proposed an “estimation and cancellation” model that iteratively detects pitch points for polyphonic music and speech signals. Hu and Wang [12] suggested a tandem algorithm to estimate pitch and segregate voiced speech jointly and iteratively.

Room reverberation smears the characteristics of pitch (i.e., harmonic structure) in speech and thus makes the task of pitch determination more difficult. The performance of existing systems is expected to degrade substantially in reverberant environments [3]. Little research has attempted to design and evaluate a multipitch tracker for reverberant speech signals, and what constitutes true pitch is even unclear in these conditions.

This paper proposes a multipitch tracking algorithm for both noisy and reverberant environments. First, we suggest a method to extract ground truth pitch for reverberant speech and use it as the reference for performance evaluation. After front-end processing, reliable channels are chosen based on cross-channel correlation and they constitute the summary correlogram for mid-level pitch representation. A pitch salience function is defined from which the pitch score of the observed correlogram given a pitch state is derived. The notion of ideal binary mask [30] is employed to divide selected channels into mutually exclusive groups, each corresponding to an underlying harmonic source. Finally, an HMM is utilized to form continuous pitch contours. The proposed method will be shown to be robust to room reverberation.

The paper is organized as follows. The next section discusses the question of what the pitch of reverberant speech should be. Sections III–V describe the detail of the proposed algorithm stage by stage. Results and comparisons are given in Section VI. We discuss related issues and conclude the paper in Section VII.

## II. WHAT SHOULD BE GROUND-TRUTH PITCH IN REVERBERANT SPEECH?

Before embarking on designing a multipitch tracker for reverberant speech, it is essential to establish a working definition of pitch in reverberant speech. This would not only point to what should be pursued, but also give a reference (or ground truth) pitch for evaluation purposes.

Manuscript received August 13, 2009; revised August 03, 2010; accepted September 05, 2010. Date of publication September 16, 2010; date of current version May 04, 2011. The work was supported in part by an Air Force Office of Scientific Research (AFOSR) under Grant FA9550-08-1-0155 and the National Science Foundation (NSF) under Grant IIS-0534707. A preliminary version has been published in the Proceedings of ICASSP 2010. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Richard C. Rose.

Z. Jin is with the Department of Computer Science and Engineering, The Ohio State University, Columbus, OH 43210 USA (e-mail: jinzh@cse.ohio-state.edu).

D. L. Wang is with the Department of Computer Science and Engineering and the Center of Cognitive Science, The Ohio State University, Columbus, OH 43210 USA (e-mail: dwang@cse.ohio-state.edu).

Digital Object Identifier 10.1109/TASL.2010.2077280

Pitch is a percept which is closely related to fundamental frequency. Following the tradition of computational literature (see, e.g., [10]), we use the term “pitch” to refer to fundamental frequency in later sections. In voiced speech, the fundamental frequency is usually defined as the rate of vibration of the vocal folds [33]. PDAs are then designed to estimate these glottal parameters directly from the speech signal which tends to be less periodic because of movements of the vocal tract that filters the excitation signal.

However, room reverberation causes the relationship between the excitation signal and the received speech signal to degrade due to the involvement of another filter which characterizes the room acoustics. According to the image model [1], the filtering effect can be modeled as an infinite number of image sources that are created by reflecting the actual source in room walls. Therefore, the reverberant speech is an aggregated signal from all image sources and no longer consistent with the glottal parameters in the original source. Several studies have attempted to extract the glottal information by counteracting the reverberation effects. Unoki *et al.* [28] utilized the concept of modulation transfer function and the source-filter model for complex cepstrum analysis. Prasanna and Yegnanarayana [21] predicted the location of glottal closure events using the Hilbert envelope of the linear prediction residual. Flego and Omologo [8] used a microphone array to remove channel variations for distant-talking speech. One result of doing so is that it creates a mismatch between the detected pitch and the actual periodicity of the received, reverberant speech, which may cause problems in applications. For example, a CASA system performing pitch-based speech segregation [14] would prefer a pitch estimate that is consistent with the harmonic structure of the reverberant speech rather than the rate of the glottal movements.

With these considerations, we consider the pitch in reverberant speech as the period of the quasi-periodic reverberant signal itself. Following this definition, we generate reference pitch contours for reverberant speech by adopting an interactive PDA [17]. This technique combines automatic pitch determination and human intervention. Specifically, it utilizes a simultaneous display (on the frame-by-frame basis) of the low-pass filtered waveform, the autocorrelation of the low-pass filtered waveform, and the cepstrum of the wideband signal. Each separate display has an estimate of the pitch period and the final decision is made by a knowledgeable user. More discussion is given in Section VI-A.

### III. FRONT-END PROCESSING

In this stage, our system decomposes the input signal into the time–frequency (T-F) domain and extracts correlogram and cross-channel correlation features.

#### A. Gammatone Filterbank

The input signal  $x(t)$  is first passed through a gammatone filterbank for time–frequency decomposition. This filterbank simulates cochlear filtering and is a standard model of the auditory periphery [20]. We use the fourth-order gammatone filterbank with 128 channels whose center frequencies are quasi-logarithmically spaced from 80 to 5000 Hz. The equivalent rectangular bandwidth (ERB) of each channel increases with the center frequency. The response  $x(c, t)$  of a filter channel  $c$  is further transduced by the Meddis model of auditory nerve transduction [18], which simulates the nonlinear characteristics of inner hair cells and produces firing activity in the auditory nerve, denoted by  $h(c, t)$ . Note that both  $x(c, t)$  and  $h(c, t)$  retain the original sampling frequency. In each channel, the output is then cut into 20-ms time frames with 10-ms frame shift. The resulting time–frequency representation is called a cochleagram and implementation details can be found in [31, Chap. 1]. We use  $u_{c,m}$  to denote a T-F unit for frequency channel  $c$  and time frame  $m$  in the cochleagram.

#### B. Correlogram

The normalized correlogram  $A(c, m, \tau)$  for T-F unit  $u_{c,m}$  of time frame  $m$  and channel  $c$  with a time delay  $\tau$  is computed by the normalized autocorrelation shown in (1) at the bottom of the page, where  $N$  denotes the frame length in samples. For the sampling frequency of 16 kHz, the frame size of 20-ms translates to  $N = 320$  samples; note that we allow samples from the neighboring frames to be used in calculating the correlogram. The denominator in (1) normalizes the correlogram to  $[0, 1]$ . The range of  $\tau$  should include the plausible pitch range.

Studies of pitch perception indicate that the pitch of complex sounds may be derived by combining information from both fine-structure phase-locking responses (resolved harmonics) in low-frequency channels and envelope-locking responses (unresolved harmonics) in high-frequency channels [4], [19]. Use of the neural underpinnings of pitch perception have proven to be useful in several CASA-based pitch detection models [11], [32]. However, in the reverberant case, pitch-related temporal-envelope cues are more degraded than fine-structure cues [24]. This is because the phase relationship of the harmonic components is randomized due to the filtering effect of reverberation, causing the complex sound reaching our ears to have a much less-modulated temporal envelope than the waveform of the sound source. In contrast to envelope responses, adding reverberation has little effect on temporal responses [24]. To make our system robust to room reverberation, we choose to only use the correlogram computed directly from the filter responses  $h(c, t)$ , rather than the temporal envelopes of  $h(c, t)$ .

$$A(c, m, \tau) = \frac{\sum_{n=-N/2}^{N/2} h(c, mN/2 + n)h(c, mN/2 + n + \tau)}{\sqrt{\sum_{n=-N/2}^{N/2} h^2(c, mN/2 + n)}\sqrt{\sum_{n=-N/2}^{N/2} h^2(c, mN/2 + n + \tau)}} \quad (1)$$

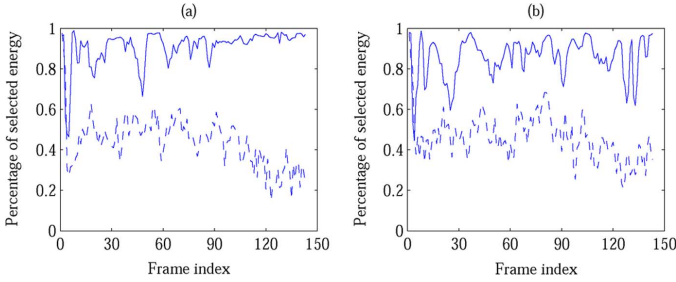


Fig. 1. Percentage of energy belonging to selected channels. (a)  $T_{60} = 0$  s. (b)  $T_{60} = 0.6$  s. The solid lines represent “speech + speech” cases and the dotted lines represent “speech + white noise” case.

### C. Cross-Channel Correlation

To detect pitch in noisy speech, it is suggested that selecting less corrupted channels from the correlogram improves the robustness of the system [22], [32]. In [32], for example, the maximum peak value at nonzero lags in  $A(c, m, \tau)$  is compared against a predetermined threshold to decide whether a low-frequency channel is noisy. A high-frequency channel is selected when the envelope of  $A(c, m, \tau)$  has a similar shape to that of another  $A(c, m, \tau)$  calculated using a bigger time window, but we find that it does not work well when reverberation is present. The main problem lies in high-frequency channels where envelope responses become highly degraded by reverberation.

We suggest the use of cross-channel correlation as an alternative method for channel selection. Due to their overlapping bandwidths, adjacent channels tend to have very similar patterns of periodicity in the correlogram if they are activated by a single harmonic source [25]. The cross-channel correlation between  $u_{c,m}$  and  $u_{c+1,m}$  is

$$C(c, m) = \frac{1}{L} \sum_{\tau=0}^{L-1} \hat{A}(c, m, \tau) \hat{A}(c+1, m, \tau) \quad (2)$$

where  $\hat{A}(c, m, \tau)$  is  $A(c, m, \tau)$  further normalized over  $\tau$  to have zero mean and unit variance, and  $L$  is the maximum delay in the plausible pitch range.  $C(c, m)$  gives a high value when a harmonic source has its strong presence and a low value when no harmonic source is present or background noise is dominant. Therefore, we select channels  $C_m$  in time frame  $m$  according to

$$C_m = \{c : C(c, m) > \theta_c\} \quad (3)$$

where  $\theta_c = 0.95$  is a threshold. Note that a relatively low threshold is used compared to [11] where the purpose is segmentation, not channel selection.

To demonstrate the robustness of channel selection, we calculate the percentage of energy belonging to selected channels in each frame as  $\sum_{c \in C_m} E(c, m) / \sum_c E(c, m)$ , where  $E(c, m)$  is the energy calculated as the sum of squares of the filter response within  $u_{c,m}$ . Fig. 1 displays this percentage of selected energy as a function of time frame in different types of interference under both anechoic and reverberant conditions. As shown in the figure, reverberation has little consequence on selected channels.

## IV. PITCH STATE SPACE

In this paper, we aim to track up to two pitches simultaneously; thus, the state space of pitch can be defined as a union space  $\mathcal{S}$  consisting of three subspaces with different dimensionalities [26], [32]

$$\mathcal{S} = \mathcal{S}_0 \cup \mathcal{S}_1 \cup \mathcal{S}_2 \quad (4)$$

where

$$\begin{aligned} \mathcal{S}_0 &= \{\emptyset\} \\ \mathcal{S}_1 &= \{\{\tau_1\} : \tau_1 \in [32, 200]\} \\ \mathcal{S}_2 &= \{\{\tau_1, \tau_2\} : \tau_1, \tau_2 \in [32, 200], \tau_1 \neq \tau_2\}. \end{aligned}$$

The three subspaces  $\mathcal{S}_0, \mathcal{S}_1, \mathcal{S}_2$  represent zero-, one-, and two-pitch hypotheses, respectively. We use the empty set  $\emptyset$  to indicate the absence of pitch, and time lags  $\tau_1$  and  $\tau_2$  to represent first and second pitch candidates. The range of pitch periods  $\tau_1$  and  $\tau_2$ , given in samples, corresponds to [2 ms, 12.5 ms] with the 16-kHz sampling frequency. This range translates to the pitch detection range from 80 to 500 Hz, a typical frequency range that covers both male and female speech in daily conversations.

### A. One-Pitch Hypothesis

When a pitch state  $s_1 \in \mathcal{S}_1$ , it is assumed that there is one and only one pitch in the current frame. To derive the pitch score  $p_m(s_1)$  in frame  $m$  given a pitch state  $s_1 = \{\tau_1\}$ , we first define the salience (or strength) of pitch candidate  $\tau_1$  within frame  $m$  as

$$f_m(\tau_1) = \begin{cases} \frac{\sum_{c \in C_m} A(c, m, \tau_1) \log E(c, m)}{\sum_{c \in C_m} \log E(c, m)}, & \text{if } C_m \neq \emptyset, \\ 0, & \text{else.} \end{cases} \quad (5)$$

The logarithmic operation acts like a pre-emphasis filter [13] which relieves the problem of high energy concentration in the low-frequency range for natural speech. The salience function  $f_m$  is essentially a weighted summary correlogram over the set of selected channels  $C_m$ . When a pitch exists, it is expected to have a predominant peak at the corresponding time delay and channel selection suppresses other “erroneous” peaks. Note that, if no channel is selected (e.g., in the case of pure noise), we set the salience function to zero for all pitch lags. In order to overcome subharmonic errors, when a peak in  $f_m$  is above a threshold  $\theta_p = 0.6$ , its related higher-order peaks are suppressed if they exist (see [7]).

The pitch score is then simply the modified salience

$$p_m(s_1) = f_m(\tau_1) \quad (6)$$

### B. Two-Pitch Hypothesis

When the noise has some periodic components or is another speech signal, we should capture both pitches—this is when the two-pitch hypothesis comes into play. In the following, we derive the pitch score  $p_m(s_2)$  given a pitch state  $s_2 = \{\tau_1, \tau_2\}$ .

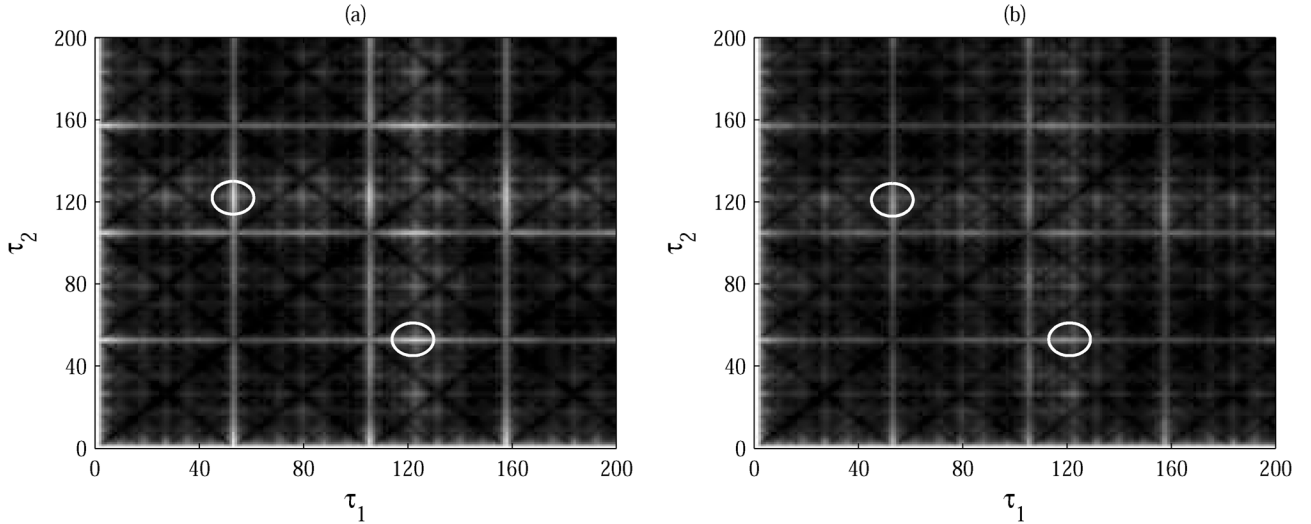


Fig. 2. Pitch salience function  $g_m$  in one time frame in a mixture of two speakers. Note that  $g_m$  is symmetric. The zero-setting step [as in (9)] is omitted in order to display the function smoothly. Plot (a) corresponds to the anechoic condition ( $T_{60} = 0$  s) and plot (b) to the reverberant condition ( $T_{60} = 0.6$  s). Brighter color indicates higher salience. The two panels show similar patterns and similar peak locations. The bright horizontal and vertical lines at multiples of 52 indicate one harmonic source with its pitch period at  $\tau_1 = 52$ . By searching for the global peaks, highlighted by white circles, the pitch period of the other source is found correctly at  $\tau_2 = 121$ .

It is not straightforward to design a pitch salience function in this situation because we are dealing with two pitches with the function expected to show a peak at or near the two true pitch periods. Since detecting multiple pitches is related to sound separation [31], we employ the notion of ideal binary mask [30] by assuming that each T-F unit is dominated by either one harmonic source or the other. Therefore, we divide the selected channels into two groups, each corresponding to one source:

$$C_{m,1} = C_m \cap \{c : A(c, m, \tau_1) \geq A(c, m, \tau_2)\} \quad (7)$$

and

$$C_{m,2} = C_m \cap \{c : A(c, m, \tau_1) < A(c, m, \tau_2)\}. \quad (8)$$

In other words, among all the selected channels, we assign a channel to source 1 if the correlogram has a higher value at  $\tau_1$  than  $\tau_2$  and source 2 otherwise. Note that  $C_{m,1} \cap C_{m,2} = \emptyset$  and  $C_{m,1} \cup C_{m,2} = C_m$ . Following this idea, we define a pitch salience function for  $s_2$  in each frame  $m$ , shown in (9) at the bottom of the page. The function is set to zero when either  $C_{m,1}$  or  $C_{m,2}$  is the empty set. We expect that this salience function generates a high peak near the two real pitch periods, since  $\tau_1$  and  $\tau_2$  should coincide with the peak locations in the channels from  $C_{m,1}$  and  $C_{m,2}$ , respectively.

An appealing property of  $g_m$  is that room reverberation hardly affects the peak formation near the real pitch periods. As we know, reverberation distorts the harmonic structure

and causes damped (less peaky) periodic patterns in the correlogram. However, the comparison between  $A(c, m, \tau_1)$  and  $A(c, m, \tau_2)$  are unlikely disrupted due to similar degradation in their values. Fig. 2 plots  $g_m$  in one same frame with and without room reverberation. The absolute value of salience  $g_m$  may be lower in reverberation, but the peak locations are robust across the two conditions. This feature is a key of our system.

We could have defined  $p_m(s_2)$  similarly to (6), but  $\mathcal{S}_2$  would dominate  $\mathcal{S}_1$  in this case. One way to fix this problem is to replace the numerator in (9) by

$$\sum_{c \in C_m} \max(A(c, m, \tau_1), A(c, m, \tau_2)) \log E(c, m) \quad (10)$$

It is clear from (10) that  $g_m(\tau_1, \tau_2)$  is greater than either  $f_m(\tau_1)$  or  $f_m(\tau_2)$ . In other words, the system would be prone to detecting a “spurious” pitch in the single pitch scenario. This problem, however, can be alleviated by scaling  $g_m$  and introducing a penalty term in  $p_m(s_2)$  as explained below.

To make  $\mathcal{S}_2$  and  $\mathcal{S}_1$  comparable, we define the pitch score by scaling  $g_m$  to the power of  $\gamma$ . Specifically,

$$p_m(s_2) = (g_m(\tau_1, \tau_2) + \delta_m)^\gamma - \delta_m \quad (11)$$

where  $\delta_m = 1 - \max_{\tau_1, \tau_2} g_m(\tau_1, \tau_2)$  and it ensures the scaling does not change the maximal peak of  $g_m$ . The scaling factor  $\gamma$  is set to 6 at which the marginal distribution of the scaled  $g_m$

$$g_m(\tau_1, \tau_2) = \begin{cases} \frac{\sum_{c \in C_{m,1}} A(c, m, \tau_1) \log E(c, m) + \sum_{c \in C_{m,2}} A(c, m, \tau_2) \log E(c, m)}{\sum_{c \in C_{m,1}} \log E(c, m) + \sum_{c \in C_{m,2}} \log E(c, m)}, & \text{if } C_{m,1} \neq \emptyset \text{ and } C_{m,2} \neq \emptyset \\ 0, & \text{else} \end{cases} \quad (9)$$

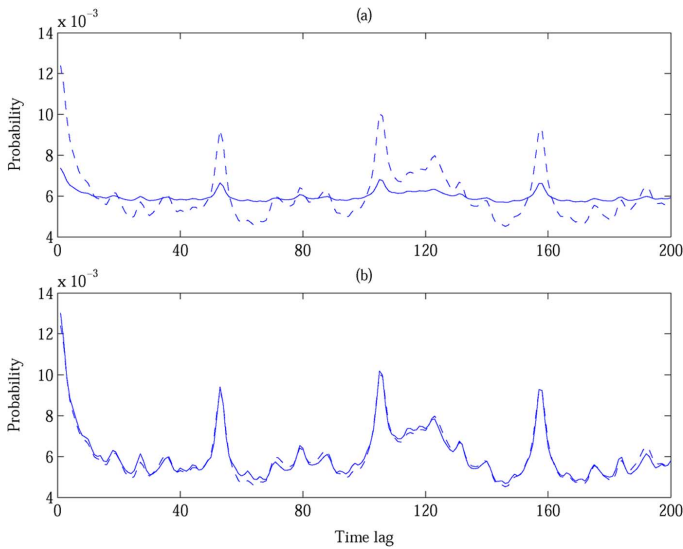


Fig. 3. Score matching. (a) Before scaling. (b) After scaling. The dotted lines represent the distribution derived from  $f_m$  and the solid lines represent the marginal distribution of  $g_m$  by summing across one  $\tau$ .

closely matches the distribution of  $f_m$ , as illustrated in Fig. 3. We find that the choice of  $\gamma$  is robust to reverberation.

### C. Zero-Pitch Hypothesis

When there is no pitch in one frame, i.e.,  $s_0 \in \mathcal{S}_0$ , it implies silence, unvoiced speech, noise, or a combination. Hence, we define its pitch score as

$$p_m(s_0) = \begin{cases} 1, & \text{if } \min(f_m) > \theta_s \\ \theta_p, & \text{else if } \text{var}(f_m) < \theta_b \\ 0, & \text{else.} \end{cases} \quad (12)$$

where  $\min(f_m)$  and  $\text{var}(f_m)$  are minimum and variance of  $f_m$  over  $\tau$ , respectively. In (12), the first case handles silence and some unvoiced speech. As shown in Fig. 4(a) and (b), for silence and high-frequency variations in unvoiced speech, their weighted summary correlograms  $f_m$  exhibit high values for all pitch lags. When all  $f_m$  values are greater than  $\theta_s = 0.5$ , a high score is assigned to  $\mathcal{S}_0$ . The second case covers broadband noise and the rest of unvoiced speech. When only this noise is present,  $f_m$  varies randomly and should have no prominent peaks [Fig. 4(c)]. In contrast, a harmonic source should exhibit a peaky distribution (high variance) in  $f_m$  [Fig. 4(d)]. Therefore, with the previously defined  $\theta_p$  and by choosing  $\theta_b = 0.01$ , we remove false pitch points from noise while still maintaining the ability to detect harmonicity buried in noise. In the third case, at least one pitch should exist, and hence the pitch score in (12) is set to zero.

### D. Parameters

In deriving pitch scores for different hypotheses, there are a total of five parameters with their values given where they are introduced. In general, the values of these parameters are chosen by examining the statistics from a small set of sample mixtures drawn in both anechoic and reverberant conditions.  $\theta_c$  is introduced in channel selection and it is determined so that a majority of the target-dominant channels are selected while minimizing the chance of including noise-dominant channels. The

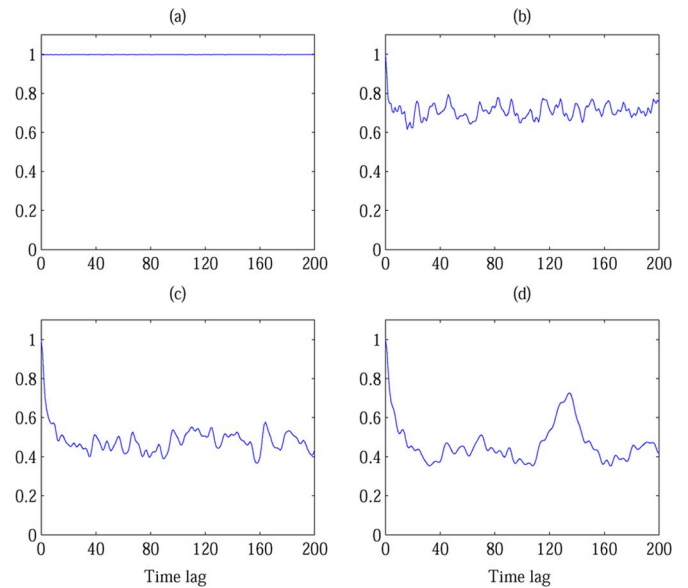


Fig. 4. Weighted summary correlogram in a frame. (a) Silence. (b) Unvoiced speech. (c) White noise. (d) Speech + white noise.

threshold  $\theta_p$  is used to identify strong salience peaks for correcting subharmonic errors and chosen to ensure that nearly all  $f_m$  values exceeding  $\theta_p$  occur at true pitch periods. Note that the same parameter is used in (12) to detect the existence of pitch. The scaling factor  $\gamma$  is chosen to balance  $f_m$  and  $g_m$ , as described in IV-B. By incorporating them into probability distributions, we find an optimal scaling factor when the marginal distribution of the scaled  $g_m$  is closest to the distribution of  $f_m$ . The parameters  $\theta_s$  and  $\theta_b$  in (12) differentiate silence, unvoiced and wideband noise frames from pitched ones. Their values are set by calculating the minimum of  $\min(f_m)$  in all silence/unvoiced frames and the maximum of  $\text{var}(f_m)$  in all wideband noise frames. Note that these parameter values need not be particular as indicated by their roundedness, and generalize well to the evaluation corpus used in Section VI even though they are selected from a small training set of 30 sample mixtures in each  $T_{60}$  outside the corpus.

## V. HMM TRACKING

### A. Model Specification

A hidden Markov model is employed as a stochastic framework to find the optimal sequence of hidden pitch states [32]. The HMM is described as follows.

- 1) Hidden states. Unlike many other practical applications, there is no ambiguity in defining the state space in our model. As discussed in the beginning of Section IV, the state space contains three subspaces corresponding to zero-, one-, and two-pitch hypotheses, respectively. We note that the cardinality (number of states) of this space is 28, 562 ( $1 + 169 + 169 \times 168$ ), which is a huge number. Later, we give ways to improve the computational efficiency. We denote the state in time frame  $m$  as  $q_m$ .
- 2) Observations. In time frame  $m$ , the observation  $\mathcal{O}_m$  is the correlogram. It is a  $128 \times 200$  matrix, with each element taking values in  $[0, 1]$  [see (1)].

TABLE I  
TRANSITION SCORES BETWEEN PITCH STATE SUBSPACES

	$\rightarrow \mathcal{S}_0$	$\rightarrow \mathcal{S}_1$	$\rightarrow \mathcal{S}_2$
$\mathcal{S}_0$	0.90	0.10	-
$\mathcal{S}_1$	0.01	0.97	0.02
$\mathcal{S}_2$	-	0.03	0.97

- 3) State transitions probability  $\mathcal{A}$ . We use a first-order HMM in which the current state only depends on the previous state. That is,  $\mathcal{A} = \{a_{q_{m-1}, q_m}\}$ . There are two aspects in  $a_{q_{m-1}, q_m}$ : The first is the score of jumping between the three pitch subspaces. To reduce search space, we assume that jumping can only take place between neighboring pitch subspaces. For example, if  $q_{m-1}$  is in  $\mathcal{S}_0$ ,  $q_m$  can be in  $\mathcal{S}_0$  or  $\mathcal{S}_1$ , but not  $\mathcal{S}_2$ . We assign jump scores in Table I. These numbers do not need to be exact as long as the diagonal probabilities are sufficiently high, and they are taken directly from [32] after rounding to the nearest hundredth. The second aspect is pitch continuity. As suggested in [32], it can be modeled by a Laplacian distribution

$$p_t(\Delta) = \frac{1}{2\sigma} \exp\left(-\frac{|\Delta - \mu|}{\sigma}\right) \quad (13)$$

where  $\Delta$  represents the change of pitch period from one frame to the next. We limit  $|\Delta| \leq 20$  to further reduce search space.  $\mu$  and  $\sigma$  are bias and spread, respectively. Following [32], we let  $\mu = 0.4$  and  $\sigma = 2.4$ . Note that all these coefficients may vary in different corpora and different reverberant environments, but they are not sensitive for pitch tracking results.

- 4) Observation probability distribution  $\mathcal{B}$  given a pitch state. As formulated in (6), (11), and (12), the pitch scores constitute  $\mathcal{B} = \{b_j(\mathcal{O}_m)\}$ , where

$$b_j(\mathcal{O}_m) = p_m(s_j), \quad 0 \leq j \leq 2. \quad (14)$$

Note that  $p_m(s_j)$  is not a probability in a strict sense. Here, we assimilate the pitch score a probability form in order to facilitate the discussion in the HMM context.

- 5) Initial state distribution  $\pi$ . We assume that every sentence starts with no pitch, i.e.,  $q_1 = \emptyset$  with probability one.

Given the above HMM,  $\Lambda = (\mathcal{A}, \mathcal{B}, \pi)$ , the task of pitch tracking is essentially to solve the following problem: given the observed correlogram sequence  $\mathcal{O} = \mathcal{O}_1 \mathcal{O}_2 \cdots \mathcal{O}_T$ , and the model  $\Lambda$ , find the most likely pitch state sequence  $\mathcal{Q}_{\max} = q_1 q_2 \cdots q_T$ . That is,

$$\begin{aligned} \mathcal{Q}_{\max} &= \operatorname{argmax}_{\mathcal{Q}} p(\mathcal{Q} | \mathcal{O}, \Lambda) \\ &= \operatorname{argmax}_{\mathcal{Q}} p(\mathcal{O}, \mathcal{Q} | \Lambda) \\ &= \operatorname{argmax}_{\mathcal{Q}} p(\mathcal{O} | \mathcal{Q}, \Lambda) p(\mathcal{Q} | \Lambda) \end{aligned} \quad (15)$$

where  $T$  is the total number of frames and  $\mathcal{Q}$  is a sequence of pitch states.  $p(\mathcal{O} | \mathcal{Q}, \Lambda)$  is defined by  $\mathcal{B}$  and  $p(\mathcal{Q} | \Lambda)$  is by  $\mathcal{A}$ . The

Viterbi algorithm provides a dynamic programming solution to the above problem and its time complexity is proportional to the size of the trellis. For efficient implementation of Viterbi search procedure, several considerations are suggested in [32].

- Remove from the trellis the least likely transition paths. This was discussed earlier in the section.
- Use beam search to reduce the total number of pitch state sequences maintained for comparison in a time frame.
- Trim the size of  $\mathcal{S}_2$  by only considering pitch candidates in the vicinity of the local peaks in (11).

These treatments are implemented and dramatically reduce the search time with almost identical results.

### B. Pitch State Tracking

The above HMM framework makes reasonably balanced decisions among different pitch hypotheses. However, biases can occur in some situations. For example, when speech is mixed with broadband noise, the level of spectral distortion breaks the balance, causing the HMM search process to be biased towards  $\mathcal{S}_2$ . To overcome this bias, we perform two independent Viterbi searches within different state spaces,  $\mathcal{S}_0 \cup \mathcal{S}_1$  and  $\mathcal{S}_0 \cup \mathcal{S}_1 \cup \mathcal{S}_2$ . Consequently, we obtain two pitch state sequences  $\mathcal{Q}_{\max,1}$  and  $\mathcal{Q}_{\max,2}$ , which can be viewed as two output candidates, each capable of tracking at most one or two pitches, respectively. Their corresponding log likelihoods are denoted by

$$\begin{aligned} l_1 &= \log p(\mathcal{Q}_{\max,1} | \mathcal{O}, \Lambda) / |\mathcal{Q}_{\max,1}| \\ l_2 &= \log p(\mathcal{Q}_{\max,2} | \mathcal{O}, \Lambda) / |\mathcal{Q}_{\max,2}| \end{aligned} \quad (16)$$

and they are normalized by the respective sequence length. Again, note that the likelihoods are calculated from pitch scores, thus not in the strict probabilistic sense.

Basically,  $l_1$  and  $l_2$  can be indicative of choosing as the final output between  $\mathcal{Q}_{\max,1}$  and  $\mathcal{Q}_{\max,2}$ . When broadband noise is present,  $l_1$  is expected to be relatively small due to reduced peak heights in  $f_m$ . On the other hand, when the interference has a periodic nature, the difference  $l_2 - l_1$  should be large due to the contribution of the second harmonic source in the likelihood score. These two aspects suggest a linear combination of  $l_1$  and  $l_2$  as a discriminant function which can be written as

$$l_2 = \alpha_0 + \alpha_1 l_1. \quad (17)$$

To find the parameters  $\alpha_0$  and  $\alpha_1$ , we use linear discriminant analysis (LDA) on the same set of sample mixtures as mentioned in Section IV-D. The resulting parameter values are  $\alpha_0 = -0.07$  and  $\alpha_1 = 0.7$ . Fig. 5 shows the scatter plot of  $(l_1, l_2)$ . The two kinds of mixtures are linearly separable. Note that the two parameters are not sensitive to different speech corpus and reverberant conditions, because of their derivation from likelihood scores which are defined by pitch salience functions and the inclusion of reverberant mixtures in the training set.

Finally, we select the output sequence  $\mathcal{Q}_{\max}$  according to

$$\mathcal{Q}_{\max} = \begin{cases} \mathcal{Q}_{\max,1}, & \text{if } l_2 < \alpha_1 l_1 + \alpha_0 \\ \mathcal{Q}_{\max,2}, & \text{else.} \end{cases} \quad (18)$$

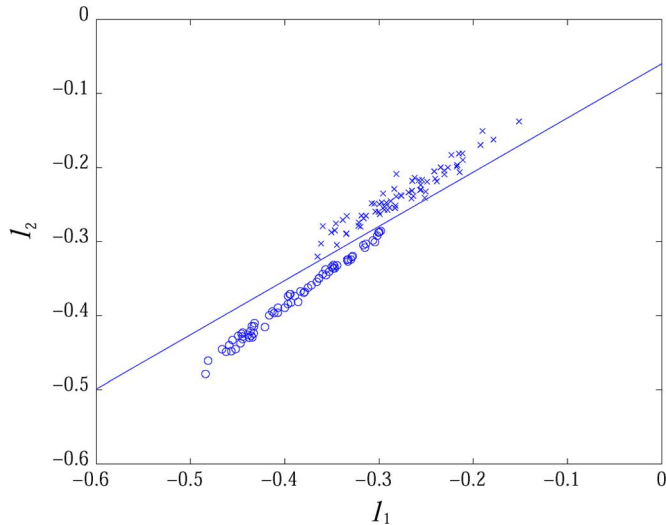


Fig. 5. Scatter plot of  $(I_1, I_2)$ . Crosses stand for mixtures with periodic interference and circles with broadband noise. The solid line is the discriminant plane from LDA.

TABLE II  
LIST OF TEN SELECTED TIMIT SENTENCES (M: MALE, F: FEMALE)

M1	Maybe they will take us.
M2	Alimony harms a divorced man's wealth.
M3	Norwegian sweaters are made of lamb's wool.
M4	Brush fires are common in the dry underbrush of Nevada.
M5	I just saw Jim near the new archeological museum.
F1	Quite often, honeybees form a majority on the willow catkins.
F2	Like enough we'll all be up on top by sundown.
F3	Challenge each general's intelligence.
F4	The water contained too much chlorine and stung his eyes.
F5	Gus saw pine trees and redwoods on his walk through Sequoia National Forest.

## VI. EXPERIMENTAL RESULTS

### A. Corpus and Reference Pitch

We construct an evaluation corpus by mixing ten randomly selected TIMIT utterances [9] with 15 different types of interference. Table II lists the selected utterances from five male and five female speakers. In Table III, the interferences are classified into three categories with five in each category: 1) those with no pitch; 2) those with some pitch qualities; and 3) other speech utterances, so that pitch tracking is evaluated differently in these categories (see Section VI-B for details). The interfering signals are compiled from NOISEX-92 [29], Cooke's corpus [5], and TIMIT.

To generate reverberant signals, we simulate room acoustics by using a MATLAB implementation [16] of the image model [1]. The model produces the room impulse response (RIR) when fed with room dimensions, wall reflection coefficients, and physical locations corresponding to sound sources and the microphone. To simulate both convolutive and additive distortions, we first specify in each configuration a random location for the microphone and then choose two locations

TABLE III  
CATEGORY OF INTERFERING SIGNALS

Category 1	White noise, pink noise, car noise, F16 cockpit noise, speech shape noise
Category 2	1 kHz tone, "cocktail party" noise, rock music, siren, trill telephone
Category 3	3 female utterances, 2 male utterances

for two sources (target and interference) randomly but control source-microphone distances to ensure that close-talking scenarios are eliminated and signal-to-reverberant energy ratios are roughly constant in each simulated room. Randomization is applied to length and width, with height fixed. Note that, even in the same room, the RIRs from different sources to the microphone differ significantly. Consequently, a reverberant mixture is constructed by convolving each source with its corresponding RIR and adding the two reverberant sources together at 0-dB signal-to-noise ratio. The resulting mixture has the sampling frequency of 16 kHz. More discussions can be found in [14].

To evaluate different reverberant conditions, we simulate two acoustic rooms with their reverberation time ( $T_{60}$ ) at 0.3 and 0.6 s, respectively. Within each room, we choose three configurations randomly and construct one reverberant mixture according to each of these configurations. More details are provided in Appendix A. Consequently, we generate a total of 1050 mixtures, with the original 150 mixtures in anechoic and  $2 \times 3 \times 150$  mixtures in reverberant conditions.

To obtain reference pitch contours, we run an interactive PDA [17] on reverberant speech signals before mixing, as described in Section II. This technique is not error free. However, as stated in Hess [10, p. 500], it is harmless to have some errors in the reference pitch contour if the PDA under evaluation will have a performance inferior to the reference PDA. This condition is met in our experiments because: 1) a pitch contour extracted from the premixed speech is expected to be more accurate than the one from the same speech mixed with interference; and 2) the manual labeling step in the reference PDA further reduces the chance of errors.

### B. PDA Performance Measure

To formulate a quantitative measure of PDA performance, we follow the metric used in [32] and extend it to reverberant cases. Generally, we use  $E_{x \rightarrow y}$  to denote the transition error rate of frames, where  $x$  pitch points are detected as  $y$  pitch points. The gross error  $E_{gs}$  is the percent of frames where the detected pitch differs with the true pitch by more than 20%. The fine error  $E_{fn}$  is defined as the average deviation from the reference pitch for those frames without gross errors.

Due to different scenarios of pitch detection in the three categories of interference, we consider each category individually:

- In Category 1, the total gross error  $E_{t1} = E_{0 \rightarrow 1} + E_{0 \rightarrow 2} + E_{1 \rightarrow 2} + E_{1 \rightarrow 0} + E_{gs}$ .
- In Category 2,  $E_{t1} = E_{0 \rightarrow 1} + E_{0 \rightarrow 2} + E_{1 \rightarrow 0} + E_{2 \rightarrow 0} + E_{gs}$ . In this category, the reference pitch of interference is not well defined for the following reasons. First, multiple harmonic sources recorded together may result in more than

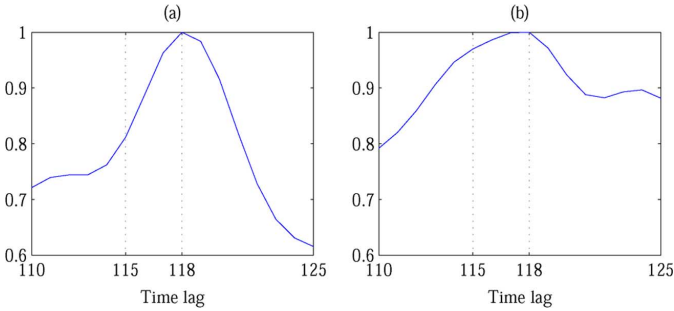


Fig. 6. Weighted summary correlogram normalized to value 1 at the true pitch period  $\tau_0 = 118$ . (a) Anechoic speech. (b) Reverberant speech.

one pitch point in a frame (e.g., cocktail party noise and rock music). Second, interference pitch may lie outside the range of voiced speech, forcing the algorithm to detect subharmonics instead (e.g., tone, siren and trill telephone). Due to these considerations, Wu *et al.* [32] only count missing pitch points for transition errors. To provide a broader indicator, we detect pitched frames in premixed interference and use them to give a measure of other transition errors. Note that  $E_{1 \rightarrow 2}$  and  $E_{2 \rightarrow 1}$  are reported but not included in the total error because it is possible that there is no error in detecting the pitch of speech in such transitions. Also,  $E_{1 \rightarrow 0}$  is measured only in the frames that contain a speech pitch point.

- In Category 3, since it is a two-talker case, all possible transition errors together with gross errors are considered. For the single reference pitch case, it is evaluated as described earlier. When two reference pitches exist in one frame, a gross error happens when the detection of either one exceeds 20% and the fine error is the sum of the two when applicable.

The above definition of fine error may not reflect well the accuracy of pitch determination in reverberant speech. Because multiple reflections are added to the original sound in a delayed and attenuated form, a single frame may fuse harmonic information from several preceding frames, resulting in a broader peak near the reference pitch in the correlogram. Fig. 6 illustrates the case, where the weighted summary correlograms are calculated for an anechoic speech signal and a reverberant speech signal in the same frame. The true pitch period  $\tau_0 = 118$ . Let the detected pitch  $\tau_1$  be 115. As shown in Fig. 6, in both of the conditions, the fine error is equal to 3 lag steps which does not manifest the different situations in the figure. A fine error may be more tolerable in reverberant space than the same error in the anechoic condition. Therefore, in addition to measuring the horizontal lag difference, we measure the percentage of vertical decrease in the summary correlogram. That is,

$$P_d = \frac{S_{wc}(\tau_0) - S_{wc}(\tau_1)}{S_{wc}(\tau_0)} \cdot 100\% \quad (19)$$

where  $S_{wc}$  is a weighted summary correlogram of all channels [cf., (5)]. Note that even though  $\tau_1$  might have a comparable or even higher value in  $S_{wc}$  than  $\tau_0$  (e.g., when  $\tau_1$  is a subharmonic of  $\tau_0$ ), it rarely happens within  $\tau_0$ 's 20% range. In case it happens, we treat it as correct and do not penalize it in the measure. Also note that  $S_{wc}$  is calculated from premixed speech

(i.e., without noise). It is worth pointing out that a vertical measure is usually used in pitch-based labeling in CASA [31, Ch. 3].

### C. Results and Comparison

We compare the proposed system with two multipitch tracking algorithms proposed by Wu *et al.* [32] and Klapuri [15]. Wu *et al.*'s framework is similar to ours, and it detects multiple pitches in three stages: auditory front-end processing, pitch statistical modeling, and HMM tracking. However, there are significant differences. Their algorithm uses a different channel selection strategy and pitch scores for different hypotheses are explicitly modeled from the statistical relationship between true pitch and selected peak locations. Due to the involvement of supervised training, the resulting pitch models are expected to degrade in mismatched conditions (e.g., room reverberation).

Klapuri's algorithm also starts with an auditory model. To analyze periodicity, it replaces the autocorrelation analysis with a DFT transform which is claimed to be more robust in multisource signals and have a wider pitch detection range. A so-called "summary spectrum" is computed and the pitch frequencies are iteratively detected by an estimation-and-cancellation procedure. Since it cannot detect the number of pitches in each frame reliably, the algorithm is provided with this number as prior knowledge.

Table IV gives the multipitch detection results of Wu *et al.*'s and our algorithm in different reverberant conditions. In Category 1, the proposed algorithm has lower fine errors than Wu *et al.*'s algorithm according to both  $E_{fn}$  and  $P_d$  measures. By performing a one-way ANOVA based significance test on both measures, we have verified that the above improvement is statistically significant (with  $p < 1 \times 10^{-10}$ ). However, our algorithm has higher total gross errors when  $T_{60} = 0.0$  s and 0.3 s, mainly because the peak selection/modeling strategy in Wu *et al.*'s algorithm is particularly effective for wideband noise. Notice that the gross errors have a decreasing trend as the level of reverberation increases, which does not happen in the other categories. The reason for this peculiar trend is, we believe, that, although reverberation distorts spectral shapes to make pitch determination generally more difficult, it can also strengthen some speech portions previously masked by noise, particularly when noise is wideband and stationary. In Category 2, the proposed algorithm produces lower total gross errors in reverberant conditions but not in anechoic conditions. For fine errors, our algorithm gives lower rates ( $p < 2 \times 10^{-4}$ ), particularly according to the  $P_d$  measure ( $p < 1 \times 10^{-10}$ ). In Category 3, the proposed algorithm yields significantly lower  $E_{t1}$  ( $p < 5 \times 10^{-8}$ ). Looking at individual transition errors, the main improvement comes from the sum of  $E_{1 \rightarrow 2}$  and  $E_{2 \rightarrow 1}$ , which shows that our algorithm does a better job in balancing between one- and two-pitch hypotheses. At the same time, both  $E_{fn}$  and  $P_d$  indicate that our algorithm has smaller fine errors in all three  $T_{60}$ 's ( $p < 1 \times 10^{-10}$ ).

In Fig. 7, we illustrate the pitch contours detected by Wu *et al.*'s and the proposed algorithm. Gross errors and transition errors are clearly seen in these plots. In the anechoic conditions, both systems can track pitch contours reliably. However, when reverberation is added, Wu *et al.*'s system loses its accuracy and



TABLE IV  
ERROR RATES (IN %) FOR THREE INTERFERENCE CATEGORIES

CATEGORY 1											
$T_{60}$ (s)	System	$E_{0 \rightarrow 1}$	$E_{0 \rightarrow 2}$	$E_{1 \rightarrow 0}$	$E_{1 \rightarrow 2}$	$E_{gs}$	$E_{tl}$	$E_{fn}$	$P_d$		
0.0	Wu <i>et al.</i>	1.77	0.00	15.09	0.27	1.53	18.66	0.87	3.71		
	Proposed	4.94	0.00	14.98	0.00	1.42	21.33	0.59	1.27		
0.3	Wu <i>et al.</i>	3.91	0.14	14.90	0.75	1.29	20.98	0.90	2.73		
	Proposed	7.74	0.60	12.18	0.93	2.20	23.65	0.67	0.89		
0.6	Wu <i>et al.</i>	1.83	0.08	13.56	1.08	0.86	17.41	1.20	2.88		
	Proposed	5.48	0.05	10.29	0.60	1.07	17.49	0.77	0.59		

CATEGORY 2											
$T_{60}$ (s)	System	$E_{0 \rightarrow 1}$	$E_{0 \rightarrow 2}$	$E_{1 \rightarrow 0}$	$E_{1 \rightarrow 2}$	$E_{2 \rightarrow 0}$	$E_{2 \rightarrow 1}$	$E_{gs}$	$E_{tl}$	$E_{fn}$	$P_d$
0.0	Wu <i>et al.</i>	3.02	0.15	2.78	0.29	1.64	37.42	1.38	8.97	0.99	3.13
	Proposed	2.07	0.99	3.42	0.60	1.31	16.03	1.75	9.54	0.83	1.52
0.3	Wu <i>et al.</i>	2.29	0.13	4.44	0.56	4.01	44.00	3.04	13.91	1.18	3.09
	Proposed	1.19	1.52	3.61	2.07	1.72	23.50	3.92	11.96	1.05	1.57
0.6	Wu <i>et al.</i>	0.64	0.04	7.16	1.16	5.83	48.74	3.39	17.06	1.50	3.40
	Proposed	0.92	0.88	5.73	1.90	2.71	32.29	3.49	13.73	1.34	1.85

CATEGORY 3											
$T_{60}$ (s)	System	$E_{0 \rightarrow 1}$	$E_{0 \rightarrow 2}$	$E_{1 \rightarrow 0}$	$E_{1 \rightarrow 2}$	$E_{2 \rightarrow 0}$	$E_{2 \rightarrow 1}$	$E_{gs}$	$E_{tl}$	$E_{fn}$	$P_d$
0.0	Wu <i>et al.</i>	4.86	0.26	1.27	1.60	0.32	25.44	0.64	34.39	0.98	3.38
	Proposed	2.40	1.48	0.85	3.67	0.07	15.13	0.38	23.98	0.73	1.15
0.3	Wu <i>et al.</i>	2.53	0.18	1.91	1.98	0.81	36.52	1.31	45.24	1.30	3.75
	Proposed	1.14	2.37	0.51	8.09	0.25	20.56	3.90	36.83	1.05	1.82
0.6	Wu <i>et al.</i>	0.95	0.14	1.75	2.76	2.93	44.73	4.54	57.81	2.03	4.83
	Proposed	0.40	0.52	0.61	7.39	0.39	21.22	7.07	37.60	1.48	2.32

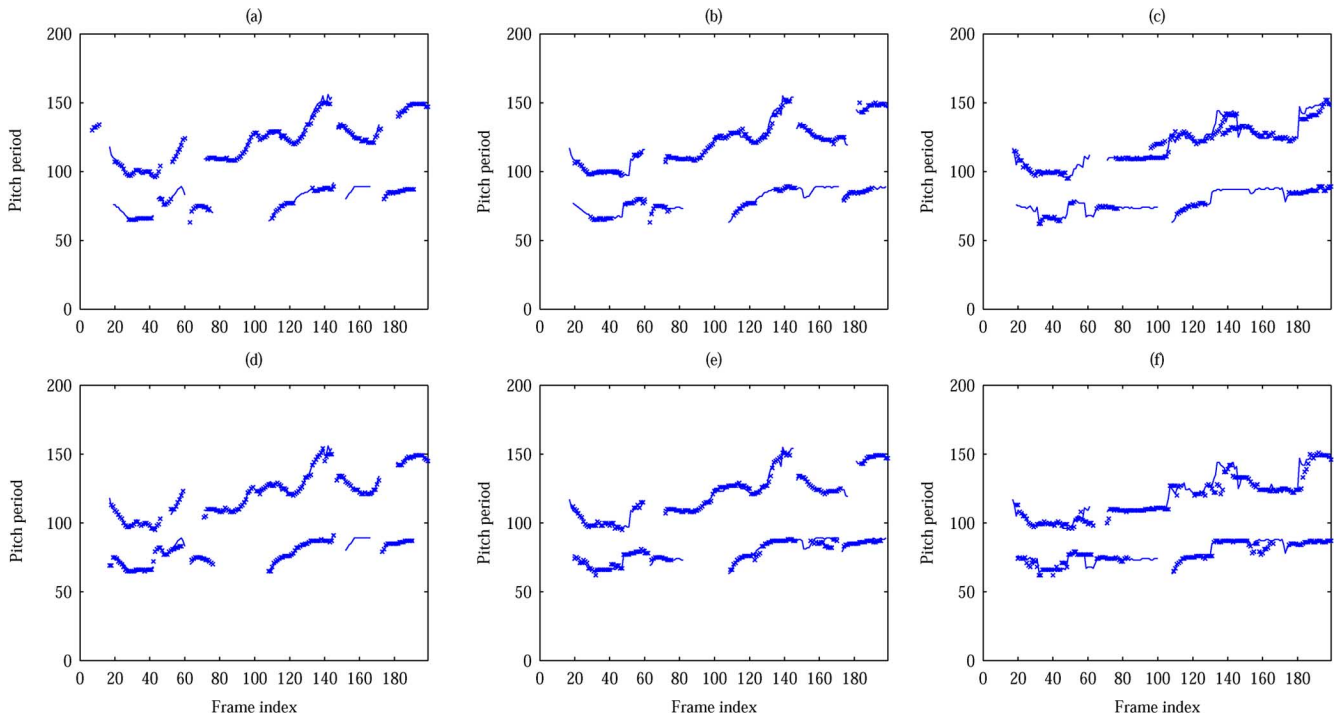


Fig. 7. Pitch tracking results for a mixture of one male and one female utterance. (a)–(c) plot detected pitch contours from Wu *et al.*'s algorithm, and (d)–(f) are from the proposed algorithm. Each column from left to right corresponds to  $T_{60} = 0.0, 0.3$  and  $0.6$  s, respectively. The solid lines indicate the reference pitch tracks. The “x” tracks represent the estimated pitch contours.

starts to make many transition and gross errors. Our algorithm performs well even in the presence of strong reverberation.

As mentioned earlier, Klapuri's algorithm requires prior information of the number of pitches in each frame. In this case,

TABLE V  
ERROR RATES (IN %) WITH PRIOR PITCH NUMBER  
FOR TWO INTERFERENCE CATEGORIES

CATEGORY 1				
$T_{60}$ (s)	System	$E_{gs}$	$E_{fn}$	$P_d$
0.0	Wu <i>et al.</i>	1.73	1.60	5.56
	Klapuri	7.89	1.26	3.18
	Proposed	2.39	1.29	3.04
0.3	Wu <i>et al.</i>	1.86	1.55	4.07
	Klapuri	13.30	1.32	2.47
	Proposed	2.13	1.21	1.91
0.6	Wu <i>et al.</i>	1.55	1.68	3.55
	Klapuri	9.86	1.37	2.10
	Proposed	0.68	1.18	1.24
CATEGORY 3				
$T_{60}$ (s)	System	$E_{gs}$	$E_{fn}$	$P_d$
0.0	Wu <i>et al.</i>	13.38	1.43	4.55
	Klapuri	6.92	1.20	3.03
	Proposed	2.50	1.06	2.26
0.3	Wu <i>et al.</i>	17.83	2.03	5.31
	Klapuri	25.38	1.54	3.34
	Proposed	8.60	1.39	2.61
0.6	Wu <i>et al.</i>	28.43	2.86	6.05
	Klapuri	28.55	1.87	3.56
	Proposed	12.85	1.78	2.96

there will be no transition errors and only gross and fine errors. For a fair comparison, we provide this prior knowledge to both Wu *et al.*'s and the proposed algorithms by disabling unrelated pitch states in the search space and ensure no transition errors are made in the results. Table V lists the error rates from all three systems. Note that only the first and the third categories of noise are evaluated because the pitch numbers are hard to determine for Category 2 interference for the reasons given in Section VI-B. For gross errors, the proposed algorithm shows a clear advantage over the other two algorithms in Category 3 ( $p < 1 \times 10^{-10}$ ), which reflects the effectiveness of the proposed salience functions  $f_m$  and  $g_m$  in one- and two-pitch hypotheses. In Category 1, Wu *et al.*'s algorithm has the lowest gross errors when  $T_{60} = 0.0$  s and 0.3 s, but it is outperformed by our algorithm when  $T_{60} = 0.6$  s. Klapuri's algorithm almost always performs the worst and lacks robustness to handle reverberation, which indicates that the summary spectrum method has limitations for wideband noise and is more susceptible to reverberation. In terms of fine errors, the proposed algorithm yields the best results in both categories and all reverberant conditions ( $p < 1 \times 10^{-10}$ ). Klapuri's system ranks second and Wu *et al.*'s almost always has the largest fine errors. It is worth noting that the above comparison of fine errors should not be taken independently as a lower rate of gross errors may make it harder to avoid fine errors. Taking this into account, we have also evaluated for each algorithm fine errors only for the same set of frames in which fine errors occur in all three algorithms. With this measure, all the algorithms yield lower fine errors while our algorithm reduces fine errors the most.

We have also implemented a version using a 64-channel gammatone filterbank that covers the same frequency range as the

original 128-channel filterbank. By doing so, the computation time is reduced roughly by half. Only one parameter needs to be adjusted to accommodate this change:  $\theta_c = 0.85$ , due to the less overlapping between adjacent channels. The 64-channel version of our algorithm yields comparable performance, with about one to two percentage points more total errors in all categories. The differences in fine error are negligible.

## VII. DISCUSSION

The impact of noise and reverberation on speech signals poses a major problem for pitch determination. The noise aspect has been studied before, but reverberation has been little investigated together with interference. A PDA that performs robustly in everyday listening environments has many applications. This paper has proposed a multipitch tracking system for reverberant conditions.

A number of novel considerations are given to the robustness of our algorithm to reverberation. First, in the front-end processing, we avoid using the temporal envelopes of filter responses to compute the correlogram in high-frequency channels (as in [32]) because they are expected to be very sensitive to reverberation. A new mechanism of channel selection is utilized to ensure the effectiveness of noise removal in reverberant conditions. Second, our formulation of pitch salience functions underlies robust derivation of pitch scores. This is worth elaborating. The use of the summary correlogram from only selected channels improves local signal-to-noise ratio and limits the influence from broadband noise. In addition, the pitch salience function for two-pitch hypothesis is defined in a robust way. The idea of assigning two disjoint groups of channels to two corresponding pitch periods is closely related to speech separation and offers an effective framework to predict how well these two pitch candidates explain the observed correlogram. As mentioned in Section IV-B, a prominent peak almost always appears near the true pitch period in different reverberant conditions. This feature affords our algorithm a considerable benefit for two-talker mixtures.

Third, one subtle but important aspect of our HMM tracking is that it not only smooths pitch contours but also plays a key role in choosing between one- and two-pitch hypotheses. From (5) and (10), we find that the maximum peak of  $p_m(s_2)$  is always greater than that of  $p_m(s_1)$  without the penalty term. Therefore, before Viterbi tracking takes place, our algorithm detects two pitches in all time frames. During the tracking process, the feature of pitch continuity can force the algorithm to switch to a single-pitch hypothesis if the detected pitch periods in neighboring frames are far apart. It is worth pointing out that, when there exists only one true pitch, the second pitch period is usually detected at a random location, unlikely near the second pitch period in the previous frame. This does not occur in the case of two true pitches. Therefore, our formulation of pitch scores allows the HMM to choose correct pitch hypotheses, which happens naturally in our formulation. This is, however, not the case for Wu *et al.*'s system where the pitch hypotheses are largely decided before HMM tracking by assigning explicit weights. These weights are obtained through training and become sensitive to different reverberant conditions.

TABLE A.I  
SETTINGS OF TWO ACOUSTIC ROOMS (L: LENGTH, W: WIDTH, H: HEIGHT)

ROOM NO.	L × W × H (m)	Reflection Coeff.	$T_{60}$ (s)
1	6 × 4 × 3	0.73	0.3
2	9 × 5 × 3	0.87	0.6

TABLE A.II  
MICROPHONE AND SOURCE LOCATIONS ( $x, y, z$ ) (IN METERS)

ROOM 1 ( $T_{60} = 0.3$ s)			
Config.	Microphone	Target	Interference
1	(2.00, 1.00, 1.30)	(0.59, 2.41, 1.30)	(3.51, 2.31, 1.30)
2	(3.32, 1.33, 1.30)	(1.38, 1.82, 1.30)	(4.82, 2.65, 1.30)
3	(4.95, 1.88, 1.30)	(4.46, 3.81, 1.30)	(2.98, 2.21, 1.30)
ROOM 2 ( $T_{60} = 0.6$ s)			
Config.	Microphone	Target	Interference
1	(1.10, 0.70, 1.30)	(2.97, 1.42, 1.30)	(1.77, 2.58, 1.30)
2	(3.04, 3.20, 1.30)	(1.10, 2.72, 1.30)	(1.05, 3.05, 1.30)
3	(6.31, 4.47, 1.30)	(4.68, 3.30, 1.30)	(5.85, 2.52, 1.30)

Table A.III  
SIGNAL-TO-REVERBERANT ENERGY RATIOS (IN dB)

ROOM 1 ( $T_{60} = 0.3$ s)		
Config.	SRR (Target)	SRR (Interference)
1	-4.37	-5.55
2	-4.10	-3.93
3	-4.21	-4.50
ROOM 2 ( $T_{60} = 0.6$ s)		
Config.	SRR (Target)	SRR (Interference)
1	-8.74	-8.66
2	-7.91	-8.23
3	-8.80	-8.28

Like many other PDAs, the proposed algorithm can be readily extended to detect more than two pitches simultaneously. The pitch state space needs to be expanded and pitch scores could be formulated using the same principle as for the two-pitch hypothesis. However, for the application of speech separation, two dominant pitches are usually enough for segregating foreground and background streams.

#### APPENDIX A ROOM CONFIGURATION SPECIFICATIONS

Tables A.I–A.III provide specifications on the acoustic rooms, microphone/source locations, and signal-to-reverberant energy ratios (SRR), respectively, used in our evaluations.

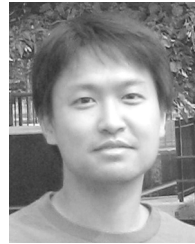
#### ACKNOWLEDGMENT

The authors would like to thank A. Klapuri for providing his pitch tracking code, and four anonymous reviewers for their useful suggestions/criticisms.

#### REFERENCES

- [1] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Amer.*, vol. 65, pp. 943–950, 1979.
- [2] F. Bach and M. Jordan, "Discriminative training of hidden Markov models for multiple pitch tracking," in *Proc. IEEE ICASSP*, 2005, pp. 489–492.
- [3] G. J. Brown and K. J. Palomäki, "Reverberation," in *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*, D. L. Wang and G. J. Brown, Eds. Hoboken, NJ: Wiley/IEEE Press, 2006, pp. 209–250.
- [4] P. A. Cariani and B. Delgutte, "Neural correlates of the pitch of complex tones. I. Pitch and pitch salience," *J. Neurophysiol.*, vol. 76, pp. 1698–1716, 1996.
- [5] M. P. Cooke, *Modeling Auditory Processing and Organization*. Cambridge, U.K.: Cambridge Univ. Press, 1993.
- [6] A. de Cheveigné, "Multiple  $F_0$  estimation," in *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*, D. L. Wang and G. J. Brown, Eds. Hoboken, NJ: Wiley/IEEE Press, 2006, pp. 45–78.
- [7] A. de Cheveigné and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music," *J. Acoust. Soc. Amer.*, vol. 111, pp. 1917–1930, 2002.
- [8] F. Flego and M. Omologo, "Robust  $F_0$  estimation based on a multi-channel periodicity function for distant-talking speech," in *Proc. EU-SIPCO*, 2006.
- [9] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, "DARPA TIMIT Acoustic Phonetic Continuous Speech Corpus," CDROM 1993 [Online]. Available: <http://www ldc.upenn.edu/Catalog/LDC93S1.html>
- [10] W. Hess, *Pitch Determination of Speech Signals*. Berlin, Germany: Springer-Verlag, 1983.
- [11] G. Hu and D. L. Wang, "Monaural speech segregation based on pitch tracking and amplitude modulation," *IEEE Trans. Neural Netw.*, vol. 15, no. 5, pp. 1135–1150, Sep. 2004.
- [12] G. Hu and D. L. Wang, "A tandem algorithm for pitch estimation and voiced speech segregation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 8, pp. 2067–2079, Nov. 2010.
- [13] X. Huang, A. Acero, and H. Hon, *Spoken Language Processing*. Upper Saddle River, NJ: Prentice Hall, 2001.
- [14] Z. Jin and D. L. Wang, "A supervised learning approach to monaural segregation of reverberant speech," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 4, pp. 625–638, May 2009.
- [15] A. Klapuri, "Multipitch analysis of polyphonic music and speech signals using an auditory model," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 2, pp. 255–266, Feb. 2008.
- [16] E. Lehmann, "Image-source method: Matlab code implementation," 2008 [Online]. Available: [http://www.eric-lehmann.com/ism\\_code.html](http://www.eric-lehmann.com/ism_code.html)
- [17] C. A. McGonegal, L. R. Rabiner, and A. E. Rosenberg, "A semiautomatic pitch detector (SAPD)," *IEEE Trans. Audio, Speech, Signal Process.*, vol. ASSP-23, no. 6, pp. 570–574, Dec. 1975.
- [18] R. Meddis, "Simulation of auditory-neural transduction: Further studies," *J. Acoust. Soc. Amer.*, vol. 83, pp. 1056–1063, 1988.
- [19] R. Meddis and L. P. O'Mard, "A unitary model of pitch perception," *J. Acoust. Soc. Amer.*, vol. 102, pp. 1811–1820, 1997.
- [20] R. D. Patterson, I. Nimmo-Smith, J. Holdsworth, and P. Rice, "An efficient auditory filterbank based on the gammatone function," in *Appl. Psychol. Unit*, Cambridge, U.K., 1988, APU Rep. 2341.
- [21] S. R. M. Prasanna and B. Yegnanarayana, "Extraction of pitch in adverse conditions," in *Proc. IEEE ICASSP*, 2004, pp. 109–112.
- [22] J. Rouat, Y. C. Liu, and D. Morissette, "A pitch determination and voice/unvoiced decision algorithm for noisy speech," *Speech Commun.*, pp. 191–207, 1997.
- [23] J. L. Roux, H. Kameoka, N. Ono, A. de Cheveigné, and S. Sagayama, "Single and multiple  $F_0$  contour estimation through parametric spectrogram modeling of speech in noisy environments," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 4, pp. 1135–1145, May 2007.

- [24] M. Sayles and I. M. Winter, "Reverberation challenges the temporal representation of the pitch of complex sounds," *Neuron*, vol. 58, pp. 789–801, 2008.
- [25] S. A. Shamma, "Speech processing in the auditory system I: The representation of speech sounds in the responses of the auditory nerve," *J. Acoust. Soc. Amer.*, vol. 78, pp. 1613–1621, 1985.
- [26] 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, pp. 229–232.
- [27] T. Tolonen and M. Karjalainen, "A computationally efficient multipitch analysis model," *IEEE Trans. Speech Audio Process.*, vol. 8, no. 6, pp. 708–716, Nov. 2000.
- [28] M. Unoki, M. Furukawa, K. Sakata, and M. Akagi, "An improved method based on the MTF concept for restoring the power envelope from a reverberant signal," *Acoust. Sci. Technol.*, vol. 25, pp. 232–242, 2004.
- [29] A. Varga and H. J. M. Steeneken, "Assessment for automatic speech recognition II: NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Commun.*, vol. 12, pp. 247–251, 1993.
- [30] D. L. Wang, "On ideal binary mask as the computational goal of auditory scene analysis," in *Speech Separation by Humans and Machines*, P. Divenyi, Ed. Norwell, MA: Kluwer, 2005, pp. 181–197.
- [31] , D. L. Wang and G. J. Brown, Eds., *Computational Auditory Scene Analysis: Principles, Algorithms and Applications*. Hoboken, NJ: Wiley-IEEE Press, 2006.
- [32] M. Wu, D. L. Wang, and G. J. Brown, "A multipitch tracking algorithm for noisy speech," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 3, pp. 229–241, May 2003.
- [33] W. A. Yost, *Fundamentals of Hearing*. San Diego, CA: Academic, 2000.



**Zhaozhang Jin** (S'06) received the B.S. degree in electrical engineering from Shanghai Jiao Tong University, Shanghai, China, and the M.S. degree in computer science and engineering from The Ohio State University, Columbus, where he is currently pursuing the Ph.D. degree.

His research interests include signal processing, machine learning, and computational auditory scene analysis.



**DeLiang Wang** (M'90–SM'01–F'04) received the B.S. and M.S. degrees from Peking (Beijing) University, Beijing, China, in 1983 and 1986, respectively, and the Ph.D. degree from the University of Southern California, Los Angeles, in 1991, all in computer science.

From July 1986 to December 1987, he was with the Institute of Computing Technology, Academia Sinica, Beijing. Since 1991, he has been with the Department of Computer Science and Engineering and the Center for Cognitive Science at The Ohio State University, Columbus, where he is currently a Professor. From October 1998 to September 1999, he was a Visiting Scholar in the Department of Psychology, Harvard University, Cambridge, MA. From October 2006 to June 2007, he was a Visiting Scholar at Oticon A/S, Denmark. His research interests include machine perception and neurodynamics.

Dr. Wang received the National Science Foundation Research Initiation Award in 1992, the Office of Naval Research Young Investigator Award in 1996, and the Helmholtz Award from the International Neural Network Society in 2008. He also received the 2005 Outstanding Paper Award from the IEEE TRANSACTIONS ON NEURAL NETWORKS.