Technical Report OSU-CISRC-8/09-TR38
Department of Computer Science and Engineering
The Ohio State University
Columbus, OH 43210-1277

Ftpsite: **ftp.cse.ohio-state.edu**
Login: **anonymous**
Directory: **pub/tech-report/2009**
File: **TR38.pdf**
Website: **http://www.cse.ohio-state.edu/research/techReport.shtml**

Multipitch Tracking for Noisy and Reverberant Speech

Zhaozhang Jin
Department of Computer Science and Engineering
The Ohio State University, Columbus, OH 43210, USA
jinzh@cse.ohio-state.edu

DeLiang Wang
Department of Computer Science and Engineering & Center for Cognitive Science
The Ohio State University, Columbus, OH 43210, USA
dwang@cse.ohio-state.edu

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 the conditional probability given each pitch state, which estimates the likelihood of the observed periodicity features given pitch candidates. A hidden Markov model integrates these probabilities 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 significantly outperforms existing ones, particularly in reverberant conditions.

Index Terms – Multipitch tracking, pitch detection algorithm, room reverberation, HMM tracking.

1 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 signal 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 [25] designed a two-channel multipitch analyzer with an enhanced summary autocorrelation function. Wu *et al.* [29] 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 [21]. Klapuri [14] proposed an "estimation and cancelation" model that iteratively detects pitch points for polyphonic music and speech signals. Hu and Wang [11] 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 conditional probability of the observed correlogram given a pitch state is derived. The notion of ideal binary mask [27] 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. Section 3, 4 and 5 describe the detail of the proposed algorithm stage by stage. Results and comparisons are given in Section 6. We discuss related issues and conclude the paper in Section 7.

2 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.

Pitch, which originally refers to a percept, has been widely used in computational literature to equate fundamental frequency (or period). So, in the following discussion, we will use these terms interchangeably. For voiced speech, the fundamental frequency is usually defined as the rate of vibration of the vocal folds [7]. 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.* [26] utilized the concept of modulation transfer function and the source-filter model for complex cepstrum analysis. Prasanna and Yegnanarayana [19] 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 information in the received speech, which may cause problems in applications. For example, a CASA system performing pitch-based speech segregation [13] 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 fundamental period of the quasi-periodic reverberant signal itself. Following this definition, we generate reference pitch contours for reverberant speech by adopting an interactive PDA [15]. 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 6.1.

3 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.

3.1 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 [18]. We use the 4-th order gammatone filterbank with 128 channels whose center frequencies are quasi-logarithmically spaced from 80 Hz 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 [16], 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 divided 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 [28] (Chap. 1). We use $u_{c,m}$ to denote a T-F unit for frequency channel c and time frame m in the cochleagram.

3.2 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 τ is computed by the following normalized autocorrelation

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

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. The denominator in (1) normalizes the correlogram to $[0,1]$. The range of τ 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, 17]. This neural underpinnings of pitch perception have proven to be useful in several CASA based pitch detection models [10, 29]. However, in the reverberant case, pitch-related temporal-envelope cues are more degraded than fine-structure cues [22]. 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 [22]. To make our system robust to room reverberation, we choose to only use the correlogram computed directly from the filter responses, rather than the response envelopes.

3.3 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 [20, 29]. The approach in [29] was to identify highly corrupted channels using different criteria in low- and high-frequency ranges. 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 to use 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 [23]. The cross-channel correlation between $u_{c,m}$ and $u_{c+1,m}$ is

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

where $\hat{A}(c,m,\tau)$ is $A(c,m,\tau)$ further normalized 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

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

where $\theta_c = 0.95$ is a threshold. Note that a relatively low threshold is used compared to [10] 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

$$\xi_m = \frac{\sum_{c\in C_m} E(c,m)}{\sum_c E(c,m)} \tag{4}$$

where $E(c,m)$ is the energy calculated as the sum of squares of the filter response within $u_{c,m}$. Fig. 1 displays ξ_m as a function of time frame in different types of interference under both anechoic and reverberant conditions. As can be observed, reverberation has little consequence on ξ_m. It is interesting to note that different types of interference vary ξ_m significantly. This effect is later utilized to discriminate broadband noise from others when formulating pitch conditional probabilities.

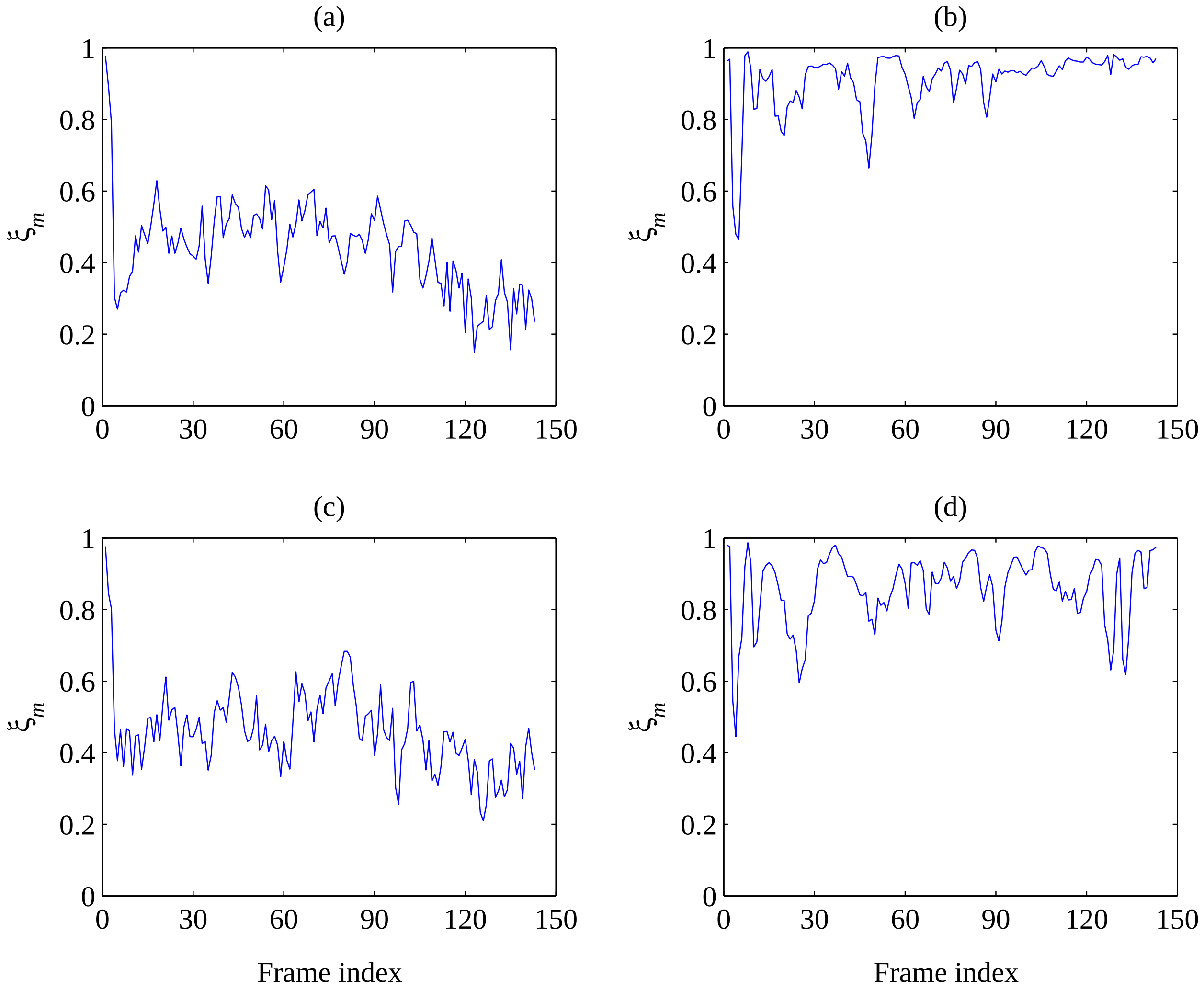

Figure 1: Percentage of energy belonging to selected channels. (a) Male speech + white noise (anechoic). (b) Male speech + female speech (anechoic). (c) Male speech + white noise (reverberant). (d) Male speech + female speech (reverberant).

4 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 [24, 29]

$$\mathcal{S} = \mathcal{S}_0 \cup \mathcal{S}_1 \cup \mathcal{S}_2 \tag{5}$$

where

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

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 τ_1 and τ_2 to represent first and second pitch candidates. The range of pitch periods τ_1 and τ_2 is set to $[32, 200]$, indicating that the pitch detection range of our algorithm is from 80 Hz to 500 Hz, a typical frequency range that covers both male and female speech in daily conversations.

4.1 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 conditional probability $p(\mathcal{O}_m|s_1)$ of observing the correlogram in frame m, $\mathcal{O}_m$, given a pitch state $s_1 = \{\tau_1\}$, we first define the salience (or strength) of pitch candidate τ_1 within frame m as

$$f_m(\tau_1) = \begin{cases} \dfrac{\sum_{c \in C_m} A(c,m,\tau_1) \log E(c,m)}{\sum_{c \in C_m} \log E(c,m)} & \text{if } C_m \neq \emptyset, \\ 0 & \text{else.} \end{cases} \tag{6}$$

The logarithmic operation acts like a pre-emphasis filter [12] 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.

The conditional probability can then be defined as

$$p(\mathcal{O}_m|s_1) = \kappa f_m(\tau_1) \tag{7}$$

where κ is a normalization coefficient for the definition of a probability measure.

4.2 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 conditional probability $p(\mathcal{O}_m|s_2)$ given a pitch state $s_2 = \{\tau_1, \tau_2\}$.

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 [28], we employ the notion of ideal binary mask [27] 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:

$$C_{m,1} = C_m \cap \Big\{c : A(c,m,\tau_1) \geq A(c,m,\tau_2)\Big\} \tag{8}$$

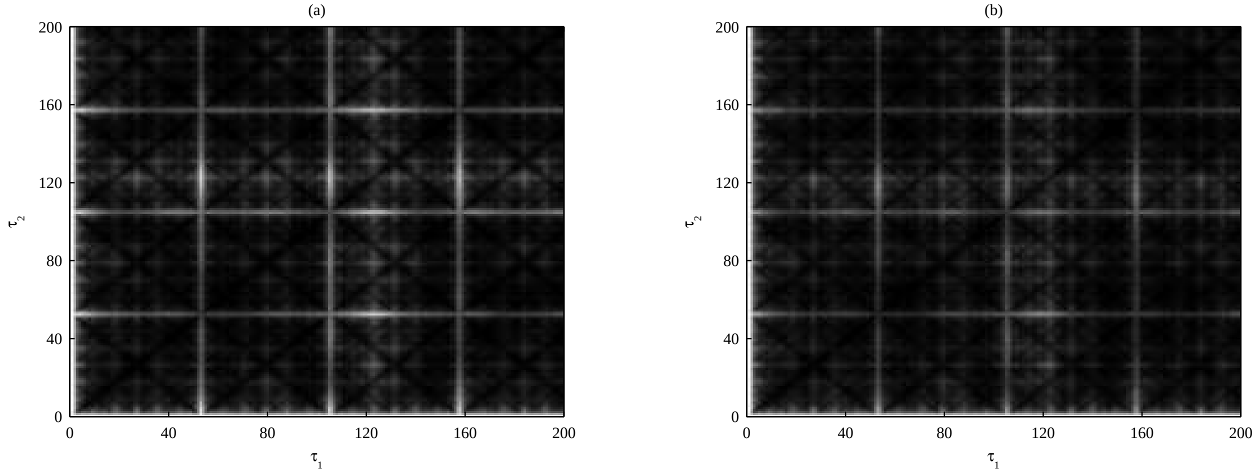

Figure 2: The pitch salience function g_m in one time frame in a mixture of two speakers. The zero-setting step (as in (10)) is omitted in order to display the function smoothly. Plot (a) corresponds to the anechoic condition and plot (b) the reverberant condition. Brighter color indicates higher salience. The two plots show a similar pattern and similar peak locations.

and

$$C_{m,2} = C_m \cap \Big\{ c : A(c,m,\tau_1) < A(c,m,\tau_2) \Big\}. \tag{9}$$

In other words, among all the selected channels, we assign a channel to source 1 if the correlogram has a higher value at τ_1 than τ_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 in (10):

$$g_m(\tau_1,\tau_2) = \begin{cases} \dfrac{\sum_{c\in C_{m,1}} A(c,m,\tau_1)\log E(c,m) + \sum_{c\in C_{m,2}} A(c,m,\tau_2)\log E(c,m)}{\sum_{c\in C_{m,1}} \log E(c,m) + \sum_{c\in C_{m,2}} \log E(c,m)} & \text{if } C_{m,1} \neq \emptyset \text{ and } C_{m,2} \neq \emptyset, \\ 0 & \text{else.} \end{cases} \tag{10}$$

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 τ_1 and τ_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) sinusoidal patterns in the correlogram. However, the comparison between $A(c,m,\tau_1)$ and $A(c,m,\tau_2)$ should not be disrupted because their values would degrade similarly and their order would remain unchanged. 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 $p(\mathcal{O}_m|s_2)$ similarly to (7), but $\mathcal{S}_2$ would dominate $\mathcal{S}_1$ in this case.

One way to elucidate this problem is to rewrite the numerator in (10) as

$$\textstyle\sum_{c \in C_m} \max(A(c,m,\tau_1), A(c,m,\tau_2)) \log E(c,m) \tag{11}$$

It is clear from (11) 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 can be alleviated by scaling g_m and introducing a penalty term in $p(\mathcal{O}_m|s_2)$ as explained below.

To make $\mathcal{S}_2$ and $\mathcal{S}_1$ comparable, we scale g_m by a power of γ. Specifically,

$$g'_m(\tau_1, \tau_2) = (g_m(\tau_1, \tau_2) + \delta_m)^\gamma - \delta_m \tag{12}$$

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 γ is set to 6 at which the marginal distribution of g'_m closely matches the distribution of f_m, as illustrated in Fig. 3. We find that the choice of γ is robust to reverberation.

Finally, we define the conditional probability as

$$p(\mathcal{O}_m|s_2) = \kappa(g'_m(\tau_1, \tau_2) - H(\beta - \xi_m) \cdot \lambda) \tag{13}$$

where it penalizes g'_m when $\xi_m \leq \beta$. $H(\cdot)$ is the Heaviside step function

$$H(x) = \begin{cases} 1 & \text{if } x \geq 0, \\ 0 & \text{else.} \end{cases} \tag{14}$$

and $\lambda = 0.05$ is the amount of penalty. As mentioned in Section 3.3, ξ_m is a good indicator of different types of interference. When speech is mixed with broadband noise, the process of channel selection tends to keep ξ_m low by excluding most of the noise energy. On the contrary, when the interference has a periodic nature, channel selection includes the energy of both sources, resulting in a high ξ_m value. We find $\beta = 0.65$ is appropriate to discriminate the above cases (see Fig. 1). Therefore, by penalizing $\mathcal{S}_2$ in the presence of broadband noise, $\mathcal{S}_1$ can compete with $\mathcal{S}_2$ in an unbiased way. Also, there is a third case in which the interference is absent. In this case, ξ_m should also display a high value which disables the penalty term. Fortunately, penalizing $\mathcal{S}_2$ is not necessary here and $\mathcal{S}_1$ automatically dominates through Viterbi tracking (more discussion in Section 7).

4.3 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 conditional probability as

$$p(\mathcal{O}_m|s_0) = \kappa \cdot \begin{cases} 1 & \text{if } \min(f_m) > \theta_s, \\ \eta & \text{else if } \mathrm{var}(f_m) < \theta_b, \\ 0 & \text{else.} \end{cases} \tag{15}$$

In (15), the first case handles silence and unvoiced speech. As shown in Fig. 4(a) and 4(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 probability is assigned to $\mathcal{S}_0$. The second case covers broadband noise. 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, by choosing $\eta = 0.6$ and $\theta_b = 0.01$, we remove false pitch points from noise while still maintain the ability to detect harmonicity buried in noise. In the third case, at least one pitch should exist, and hence the conditional probability in (15) is set to zero. Note that the choices of all these parameters are robust to different reverberant conditions.

5 HMM Tracking

A hidden Markov model is employed as a stochastic framework to find the optimal sequence of hidden pitch states [29]. The HMM is described below:

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 4, 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 $N = 28,562$, 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 128×200 matrix, with each element taking values in $[0, 1]$ (see (1)).

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 probability 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 probabilities in Table 1. These numbers do not need to be exact as long as the diagonal probabilities are sufficiently high, and they are taken directly from [29] after rounding to the nearest hundredth.

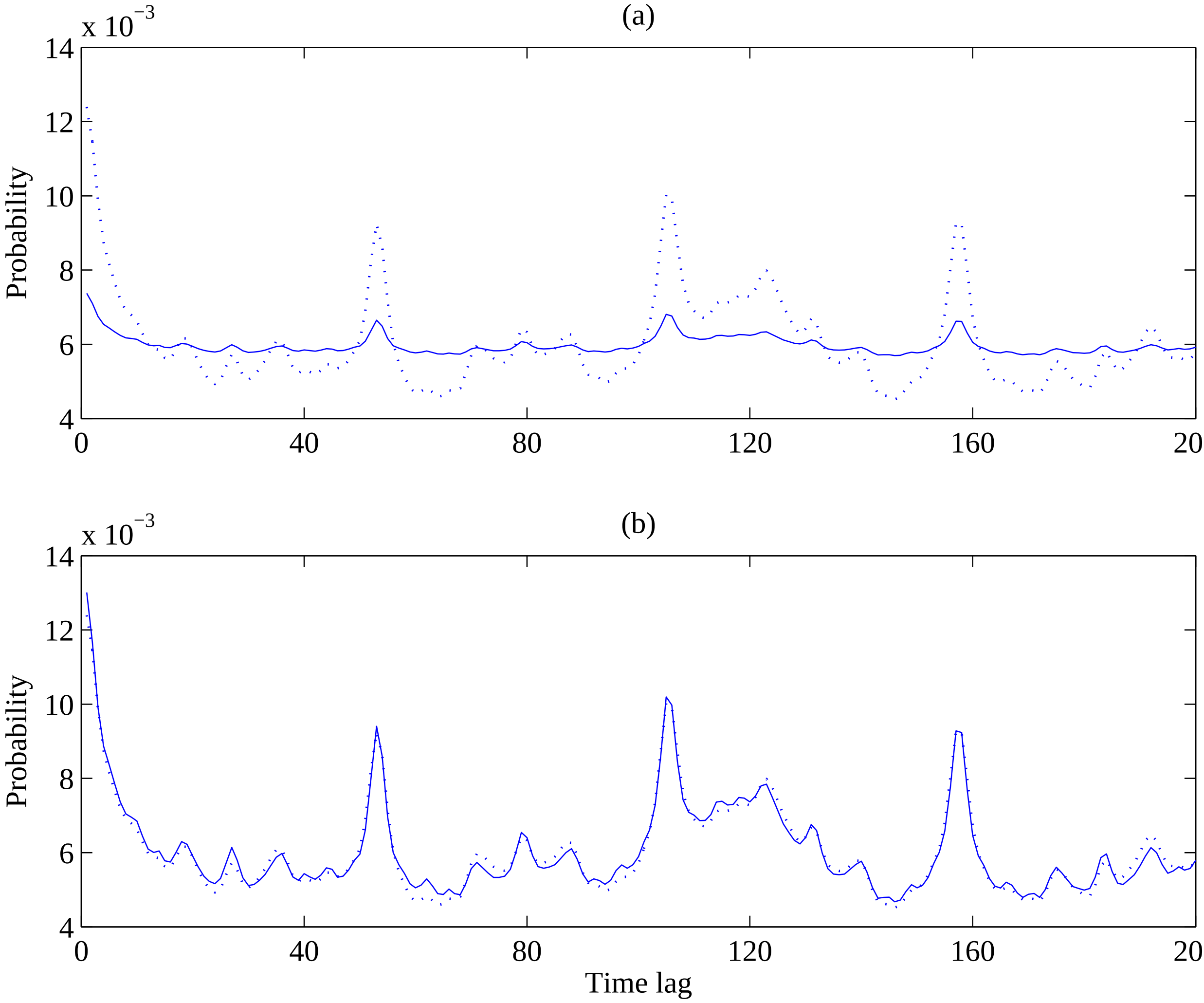

Figure 3: Probability matching. (a) Before scaling. (b) After scaling. The dotted lines represent the probability distribution derived from f_m and the solid lines represent the marginal distribution of g_m.

The second aspect is pitch continuity. As suggested in [29], it can be modeled by a Laplacian distribution

$$p_t(\Delta) = \frac{1}{2\sigma} \exp\left(- \frac{|\Delta - \mu|}{\sigma} \right) \tag{16}$$

where Δ represents the change of pitch period from one frame to the next. We limit $|\Delta| \leq 20$ to further reduce search space. μ and σ are bias and spread, respectively. Following [29], 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 (7), (13) and (15), the conditional probability distribution $\mathcal{B} = \{b_j(\mathcal{O}_m)\}$, where

$$b_j(\mathcal{O}_m) = p(\mathcal{O}_m | s_j), \quad 0 \leq j \leq 2. \tag{17}$$

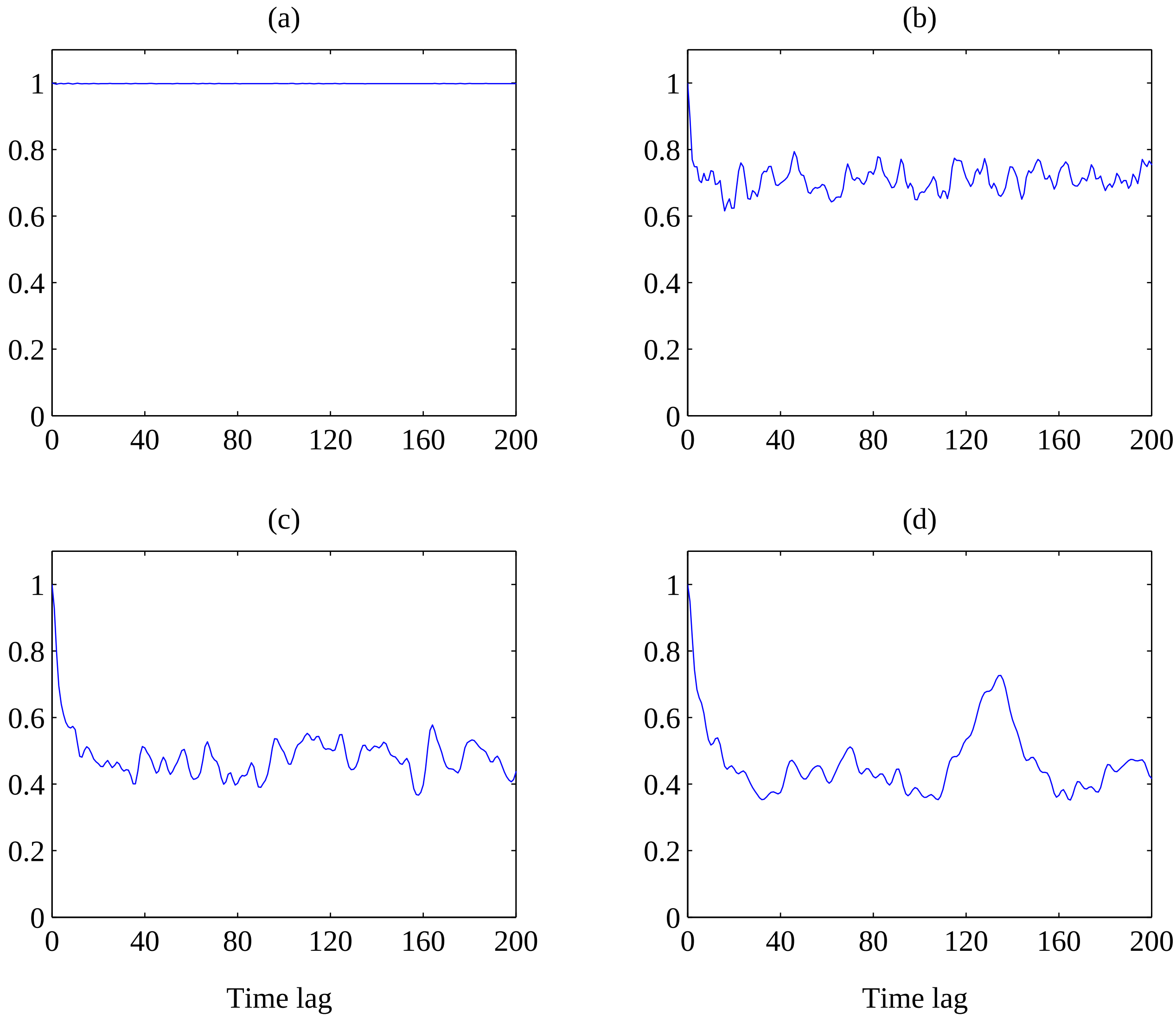

Figure 4: Weighted summary correlogram in a frame. (a) Silence. (b) Unvoiced speech. (c) White noise. (d) Speech + white noise.

5) Initial state distribution π. 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 $\mathcal{O} = \mathcal{O}_1\mathcal{O}_2...\mathcal{O}_T$, and the model Λ, find the most likely pitch state sequence $\mathcal{Q}_{max} = q_1q_2...q_T$. That is,

$$\begin{aligned} \mathcal{Q}_{max} &= \text{argmax}_{\mathcal{Q}} p(\mathcal{Q}|\mathcal{O}, \Lambda) \\ &= \text{argmax}_{\mathcal{Q}} p(\mathcal{O}, \mathcal{Q}|\Lambda) \\ &= \text{argmax}_{\mathcal{Q}} p(\mathcal{O}|\mathcal{Q}, \Lambda) p(\mathcal{Q}|\Lambda) \end{aligned} \tag{18}$$

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

Table 1: Transition probabilities 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
$\mathcal{S}_0$	-	0.03	0.97

Table 2: Category of interfering signals

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

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 [29]:

- 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 (13).

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

6 Experimental Results

6.1 Corpus and Reference Pitch

We use Cooke's corpus [5], which contains 100 noisy utterances constructed by mixing 10 voiced speech utterances with 10 different types of interference signals. This corpus is commonly used for evaluating PDA performance [21,29]. In Table 2, the interferences are classified into three categories: 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 6.2 for details).

To generate reverberant recordings, we simulate room acoustic by using the image model [1]. The model produces the room impulse response (RIR) when fed with room dimensions, wall refection coefficients and physical locations corresponding to sound sources and the microphone. To simulate both convolutive and additive distortions, we specify in each *configuration* two locations for two sources (target and interference) and another location for the microphone. Note that 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 SNR. The resulting mixture has a sampling frequency of 16 kHz.

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*. Consequently, we generate a total of 700 mixtures, with the original 100 mixtures in anechoic and $2 \times 3 \times 100$ mixtures in reverberant conditions.

To obtain reference pitch contours, we run an interactive PDA [15] on reverberant speech signals before mixing, as described in Section 2. This technique is not error free. However, as stated in Hess [9] (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.

6.2 PDA Performance Measure

To formulate a quantitative measure of PDA performance, we follow the metric used in [29] and extend it to reverberant cases. Generally, we use $E_{x \to y}$ 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 $E_{tl} = E_{0 \to 1} + E_{0 \to 2} + E_{1 \to 0} + E_{gs}$. Note that $E_{1 \to 2}$ is not counted in E_{tl} because we aim to detect a single pitch contour for the target speech in this category.

- In Category 2, $E_{tl} = E_{1 \to 0} + E_{gs}$. Due to the uncertainty of pitch in this category of interference, we only consider missing pitch points for transition errors.

- In Category 3, since it is a two-talker case, all possible transition errors together with

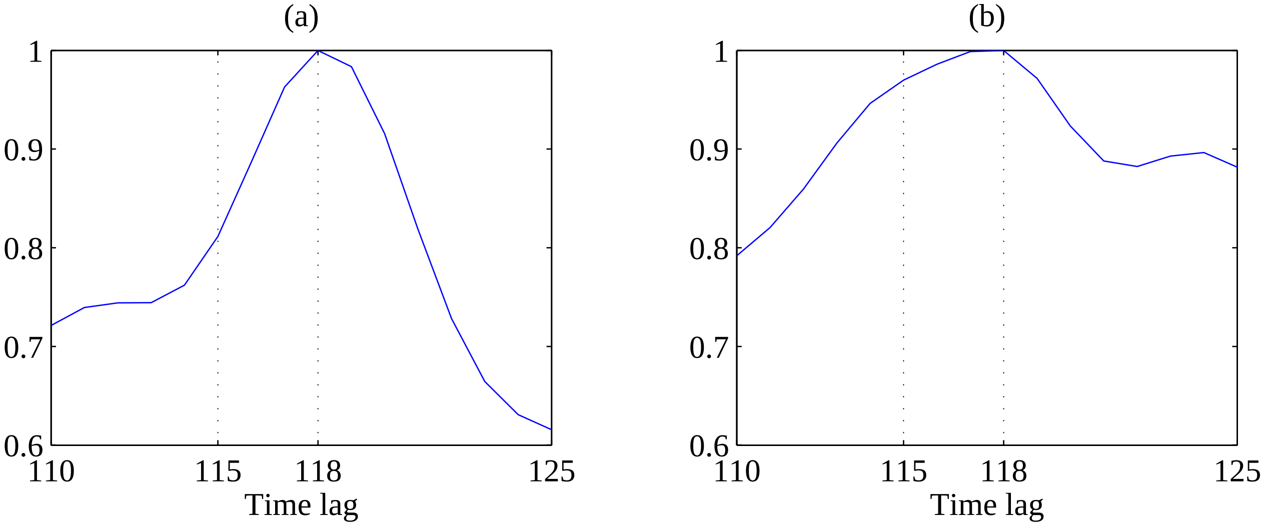

Figure 5: Weighted summary correlogram normalized to value 1 at the true pitch period $\tau_0 = 118$. (a) Anechoic speech. (b) Reverberant speech.

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. 5 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 τ_1 be 115. As shown in Fig. 5, 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,

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

where S_{wc} is a weighted summary correlogram of all channels (cf. (6)). Note that even though τ_1 might have a comparable or even higher value in S_{wc} than τ_0 (e.g. when τ_1 is a subharmonic of τ_0), it rarely happens within τ_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

Table 3: Error rates (in %) for three interference categories

Category 1

T_{60}(s)	System	$E_{0\to1}$	$E_{0\to2}$	$E_{1\to0}$	$E_{1\to2}$	E_{gs}	E_{tl}	E_{fn}	P_d
0.0	Wu *et al.*	1.01	0.00	6.23	0.06	0.00	7.24	1.21	3.23
	Proposed	1.47	0.49	7.66	4.96	0.00	9.62	1.22	2.21
0.3	Wu *et al.*	0.84	0.01	10.39	1.80	0.31	11.55	1.32	3.49
	Proposed	1.44	0.66	8.40	8.90	0.12	10.63	1.58	2.93
0.6	Wu *et al.*	0.25	0.00	14.56	8.95	0.51	15.32	1.69	4.23
	Proposed	0.85	0.53	8.53	11.34	0.44	10.35	2.06	3.31

Category 2

T_{60}(s)	System	$E_{1\to0}$	E_{gs}	E_{tl}	E_{fn}	P_d
0.0	Wu *et al.*	5.19	0.60	5.79	1.27	3.23
	Proposed	2.72	0.54	3.26	1.44	2.21
0.3	Wu *et al.*	6.56	1.67	8.23	1.54	3.63
	Proposed	2.30	1.80	4.09	1.80	2.46
0.6	Wu *et al.*	12.69	2.27	14.96	1.89	4.16
	Proposed	4.08	1.59	5.67	2.48	2.85

Category 3

T_{60}(s)	System	$E_{0\to1}$	$E_{0\to2}$	$E_{1\to0}$	$E_{1\to2}$	$E_{2\to0}$	$E_{2\to1}$	E_{gs}	E_{tl}	E_{fn}	P_d
0.0	Wu *et al.*	1.08	0.00	0.94	1.29	0.18	21.26	0.00	24.75	1.01	2.54
	Proposed	1.02	0.12	0.51	1.88	0.10	10.54	0.22	14.39	0.94	0.91
0.3	Wu *et al.*	0.80	0.05	1.36	1.17	0.93	33.16	0.87	38.34	1.29	3.37
	Proposed	1.01	0.41	0.30	4.85	0.19	14.91	3.04	24.71	1.22	1.41
0.6	Wu *et al.*	0.41	0.00	1.83	2.36	2.42	38.05	9.04	54.11	2.18	4.92
	Proposed	0.72	0.18	0.45	5.80	0.20	16.26	9.78	33.39	1.89	2.23

labeling in CASA [28].

6.3 Results and Comparison

We compare the proposed system with two multipitch tracking algorithms proposed by Wu *et al.* [29] and Klapuri [14]. 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 training, the resulting pitch models may 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-cancelation procedure. Since it cannot detect the number of pitches in each frame reliably, the algorithm is provided with this number as prior knowledge.

Table 3 gives the multipitch detection results of Wu *et al.*'s and our algorithm in different reverberant conditions. In Category 1 and 2, the proposed algorithm almost always has a lower rate of total gross error and the margin of difference grows with the increasing level of reverberation. For fine errors, our algorithm is superior according to the P_d measure but not the E_{fn} measure. As discussed earlier, E_{fn} may not be as relevant a measure for reverberant speech. Also, E_{fn} is lower for Wu *et al.*'s algorithm because it explicitly models statistics of pitch period differences used in this measure. In Category 3, the proposed algorithm yields a significantly lower E_{tl}. In the anechoic condition (T_{60} = 0.0 s), our algorithm outperforms Wu *et al.*'s by 10 percentage points. This advantage doubles in the most reverberant case (T_{60} = 0.6 s). At the same time, both E_{fn} and P_d indicate that our algorithm has smaller fine errors in all three T_{60}'s.

In Fig. 6, we plot 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 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, 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 4 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. The proposed algorithm yields the lowest gross error rate in both categories and all reverberant conditions. Klapuri's algorithm performs similarly to Wu *et al.*'s in the anechoic condition but degrades more rapidly with increasing level of reverberation. This indicates that the summary spectrum used in Klapuri's algorithm is more susceptible to reverberation. Our algorithm also yields the lowest fine errors in all conditions. 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, the proposed algorithm reduces the fine error, but the relative performance between the three algorithms is about the same.

We have also implemented a version using a 64-channel gammatone filterbank that covers

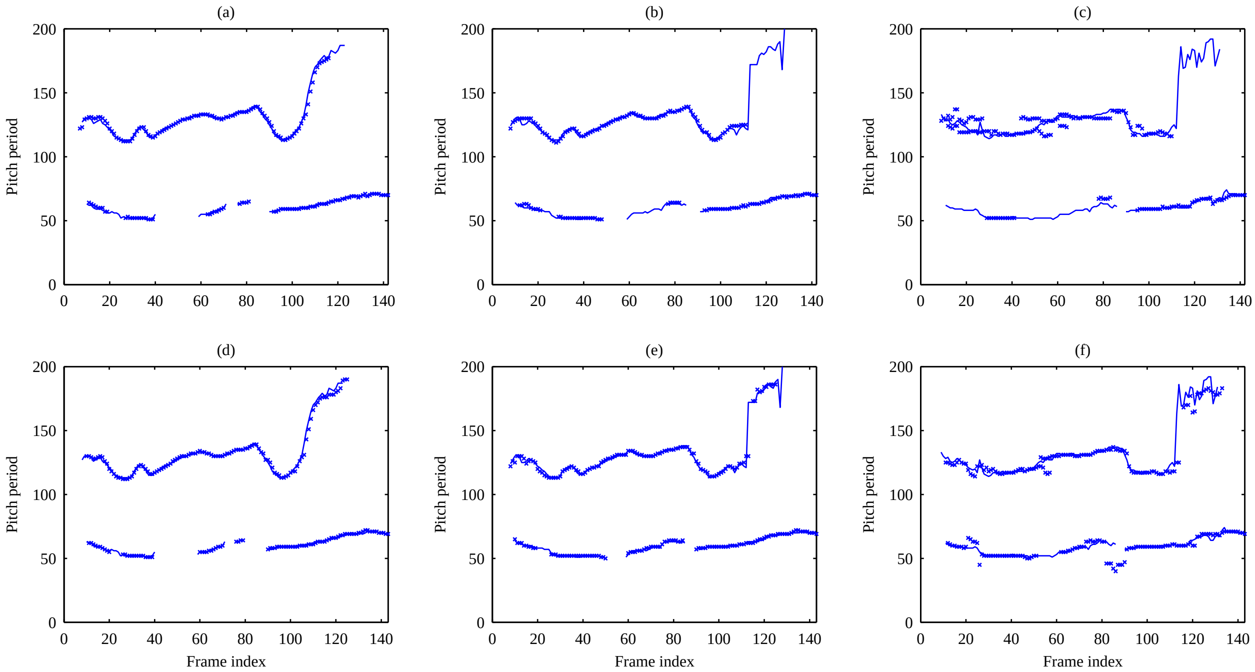

Figure 6: Pitch tracking results for a mixture of one male and one female utterance. (a)–(c) plot detected pitch contours from Wu *et al.*'s algorithm, and (d)–(f) are from the proposed algorithm. Each column from left to right corresponds to $T_{60} = 0.0$, 0.3 and 0.6 s, respectively. The solid lines indicate the reference pitch tracks. The "×" tracks represent the estimated pitch contours.

the same frequency range as the original 128-channel filterbank. By doing so, the computation time is reduced roughly by half. Three parameters are adjusted to accommodate this change: $\theta_c = 0.85$, $\beta = 0.6$, and $\lambda = 0.1$. The 64-channel version of our algorithm yields comparable performance, with about one percentage point fewer total errors in Category 1 and two to three percentage points more total errors in the other two categories. The differences in fine error are negligible.

7 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 considerations are given to the robustness of our algorithm to reverberation. First, in the front-end processing, we avoid using envelope responses to compute the correlogram in high-frequency channels because they are expected to be very sensitive to reverberation. A new mechanism of channel selection is utilized to ensure the effectiveness of

Table 4: 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 *et al.*	1.16	1.45	3.46
	Klapuri	0.74	1.57	4.09
	Proposed	0.09	1.61	2.75
0.3	Wu *et al.*	2.62	1.90	4.22
	Klapuri	5.16	1.93	3.97
	Proposed	0.50	2.13	3.64
0.6	Wu *et al.*	4.11	2.48	4.84
	Klapuri	7.17	2.68	3.86
	Proposed	1.34	2.56	3.69

Category 3

T_{60}(s)	System	E_{gs}	E_{fn}	P_d
0.0	Wu *et al.*	2.80	1.40	3.32
	Klapuri	4.82	1.37	3.05
	Proposed	0.59	1.10	1.10
0.3	Wu *et al.*	7.20	2.00	4.50
	Klapuri	21.00	1.74	3.08
	Proposed	5.10	1.48	1.74
0.6	Wu *et al.*	18.48	3.18	6.01
	Klapuri	29.12	2.51	3.64
	Proposed	11.92	2.25	2.55

noise removal in reverberant conditions. Second, our formulation of pitch salience functions underlies robust estimation of pitch conditional probabilities. 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 4.2, 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 smoothes pitch contours but also plays a key role in choosing between one- and two-pitch hypotheses. From (6) and (11), we find that the maximum peak of $p(\mathcal{O}_m|s_2)$ is always greater than that of $p(\mathcal{O}_m|s_1)$ 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 switch to a single-pitch hypothesis if the detected pitch periods in neighboring frames are far apart. It is worth pointing 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 probabilities 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.

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 conditional probabilities 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.

Acknowledgment

The authors would like to thank A. Klapuri for providing his pitch tracking code. The research described in this paper was supported in part by an AFOSR grant (FA9550-08-1-0155) and an NSF grant (IIS-0534707).

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, UK: 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 F0 estimation based on a multichannel periodicity function for distant-talking speech,” in *EUSIPCO*, 2006.

[9] W. Hess, *Pitch Determination of Speech Signals.* Berlin: Springer-Verlag, 1983.

[10] G. Hu and D. L. Wang, “Monaural speech segregation based on pitch tracking and amplitude modulation,” *IEEE Trans. Neural Networks*, vol. 15, pp. 1135–1150, 2004.

[11] ——, “A tandem algorithm for pitch estimation and voiced speech segregation,” Dept. Comp. Sci. & Eng., The Ohio State Univ., Tech. Rep. 52, 2009.

[12] X. Huang, A. Acero, and H. Hon, *Spoken Language Processing.* Upper Saddle River, NJ: Prentice Hall PTR, 2001.

[13] Z. Jin and D. L. Wang, “A supervised learning approach to monaural segregation of reverberant speech,” *IEEE Trans. Audio, Speech, and Language Processing*, vol. 17, pp. 625–638, 2009.

[14] A. Klapuri, “Multipitch analysis of polyphonic music and speech signals using an auditory model,” *IEEE Trans. Audio, Speech, and Language Processing*, vol. 16, pp. 255–266, 2008.

[15] C. A. McGonegal, L. R. Rabiner, and A. E. Rosenberg, “A semiautomatic pitch detector (SAPD),” pp. 570–574, 1975.

[16] R. Meddis, “Simulation of auditory-neural transduction: Further studies,” *J. Acoust. Soc. Amer.*, vol. 83, pp. 1056–1063, 1988.

[17] R. Meddis and L. P. O’Mard, “A unitary model of pitch perception,” *J. Acoust. Soc. Amer.*, vol. 102, pp. 1811–1820, 1997.

[18] R. D. Patterson, I. Nimmo-Smith, J. Holdsworth, and P. Rice, “An efficient auditory filterbank based on the gammatone function,” Appl. Psychol. Unit, Cambridge, UK, APU Rep. 2341, 1988.

[19] S. R. M. Prasanna and B. Yegnanarayana, “Extraction of pitch in adverse conditions,” in *Proc. IEEE ICASSP*, 2004, pp. 109–112.

[20] J. Rouat, Y. C. Liu, and D. Morissette, “A pitch determination and voice/unvoiced decision algorithm for noisy speech,” *Speech Comm.*, pp. 191–207, 1997.

[21] J. L. Roux, H. Kameoka, N. Ono, A. de Cheveigné, and S. Sagayama, “Single and multiple F0 contour estimation through parametric spectrogram modeling of speech in noisy environments,” *IEEE Trans. Audio, Speech, and Language Processing*, vol. 15, pp. 1135–1145, 2007.

[22] M. Sayles and I. M. Winter, “Reverberation challenges the temporal representation of the pitch of complex sounds,” *Neuron*, vol. 58, pp. 789–801, 2008.

[23] 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.

[24] 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.

[25] T. Tolonen and M. Karjalainen, “A computationally efficient multipitch analysis model,” *IEEE Trans. Speech Audio Processing*, vol. 8, pp. 708–716, 2000.

[26] 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,” *Acoustical Science and Technology*, vol. 25, pp. 232–242, 2004.

[27] 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 Academic, 2005, pp. 181–197.

[28] D. L. Wang and G. J. Brown, *Ed. Computational auditory Scene Analysis: Principles, Algorithms and Applications.* Hoboken, NJ: Wiley-IEEE Press, 2006.

[29] M. Wu, D. L. Wang, and G. J. Brown, “A multipitch tracking algorithm for noisy speech,” *IEEE Trans. Speech and Audio Processing*, vol. 11, pp. 229–241, 2003.