A MULTI-PITCH TRACKING ALGORITHM FOR NOISY SPEECH

Mingyang Wu and DeLiang Wang

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

Guy J. Brown

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

ABSTRACT

We present a robust algorithm for multi-pitch tracking of noisy speech. Our approach integrates an improved channel and peak selection method, a new integration method for extracting periodicity information across different frequency channels, and a hidden Markov model (HMM) for forming continuous pitch tracks, and as a result, our algorithm can reliably track single and double pitch tracks in a noisy environment. The proposed algorithm is evaluated on a database of speech utterances mixed with various interferences and the results show that our algorithm outperforms existing algorithms significantly.

1. INTRODUCTION

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

Among the numerous PDAs proposed, some have been specifically designed for detecting a single pitch track with voiced/unvoiced decisions in noisy speech. The majority of these algorithms were tested on clean speech or speech mixed with different levels of white noise. Some systems also have been tested in other speech and noise conditions. For example, the system designed by Rouat et al. [5] was tested on telephone speech, vehicle speech, and speech mixed with white noise. Other single pitch track PDAs were also tested on speech mixed with pink noise, music, and a male voice. In these studies, however, the multi-pitch nature of the signals is ignored and a single pitch decision is given.

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

The tracking of multiple pitches also has been investigated. For examples, Gu and van Bokhoven [2] proposed an algorithm for detecting up to two pitch periods for co-channel speech separation. A model by Tolonen and Karjalainen [6] was tested on musical chords and a mixture of two vowels. However, these multi-pitch trackers were designed for and tested on clean music signals or mixtures of single-pitch signals with little or no background noise interference. Their performance on tracking speech mixed with significant broadband interference such as white noise is not clear.

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

2. MODEL DESCRIPTION

The proposed algorithm consists of four stages. In the first stage, the front-end, the input is filtered into frequency channels and the envelopes in high-frequency channels are extracted. Then, the normalized correlograms are computed for each channel at every 10 ms. Section 2.1 gives the detail of this stage.

Channel and peak selection comprises the second stage. In noisy speech, some channels are significantly corrupted by noise. By only selecting the less corrupted channels, the robustness of the system is enhanced. Channel selection was first implemented on mid- and high-frequency channels by Rouat et al. [5]. We extend this idea to low-frequency channels and propose an improved method for all channels. Furthermore, we broaden the idea of peak selection. Generally speaking, peaks in normalized correlograms suggest periodicity of the signals. However, some peaks give misleading periodicity information and should be removed. Section 2.2 gives the detail of this stage.

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

relationship between ideal pitch periods and time lags of selected peaks is obtained in Section 2.3 and the integration method is described in Section 2.4.

The last stage of the algorithm is to form continuous pitch tracks using an HMM. In several previous studies, HMMs have been employed to model pitch track continuity. Weintraub [7] utilized a Markov model to determine whether zero, one or two pitches were present. Gu and van Bokhoven [2] used an HMM to group pitch candidates proposed by a bottom-up PDA and form continuous pitch tracks. In these studies, pitch is treated as the observation and the HMM must be trained. In our formulation, the pitch is explicitly modeled as the hidden states and hence there is no training needed. Finally, optimal pitch tracks are obtained by using the Viterbi algorithm. This stage is described in Section 2.5.

2.1. Multi-channel front-end

The input signals are sampled at 16 kHz and then passed through a bank of 128 fourth-order "gammatone" filters [4]. The frequency channels are further classified into two categories. Channels with center frequencies lower then 800 Hz (channels 1-55) are called low-frequency channels. Others are called high-frequency channels (channels 56-128). Envelopes are extracted in high-frequency channels. Finally, the normalized correlogram S is computed by running normalized autocorrelation using a window size of 16 ms.

2.2. Channel and peak selection

In a low-frequency channel, for a quasi-periodic signal with period T, the greater the normalized correlogram is at time lag T, the stronger the periodicity of the signal. Therefore, the maximum value of all peaks at non-zero lags measures the noise level of this channel. If the maximum value is greater than a threshold (0.945), the channel is considered "clean" and thus selected.

For high-frequency channels, as suggested by Rouat et al. [5], if a channel is not much corrupted by noise, the original normalized correlogram computed using a window size of 16 ms and the normalized correlogram S' using a longer window size of 30 ms should have similar shapes. For every local peak of S, we search for the closest local peak in S'. If the difference between the two time lags is greater then 2 delay steps, the channel is removed.

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

A high-frequency channel responds to multiple harmonics, and the nature of beats and combination tones dictates that the response envelope fluctuates at the fundamental frequency [3]. Therefore, the occurrence of strong peaks at time lag T and its multiples in a high-frequency channel suggests a fundamental period of T. Thus, for the second method of peak selection, if the value of the peak at the first non-zero time lag is greater than 0.6, all the multiple peaks are removed. The second method for peak
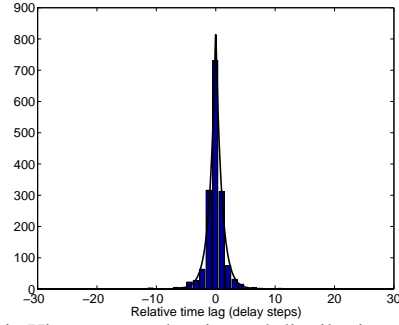
Figure 1: Histogram and estimated distribution of relative time lags for single pitch in channel 22. The bar graph represents the histogram and the solid line represents the estimated distribution.

selection is critical for reducing errors caused by multiple and sub-multiple pitch peaks in autocorrelation functions.

2.3. Pitch period and time lags of selected peaks

By studying the difference between the ideal pitch period and the time lag from the closest selected peak, we can derive the evidence of the normalized correlogram in a particular channel supporting a hypothesis of a pitch delay.

More specifically, the relative time lag Δ is defined as the distance from the ideal pitch delay to the closest peak and the statistics of the relative time lag Δ are collected from the selected channels across all voiced frames of clean speech utterances for every channel separately. As an example, the histogram of relative time lags for channel 22 is shown in Fig. 1. As can be seen, the distribution is sharply centered at zero. Thus, a mixture of a Laplacian and a uniform distribution is employed for modeling the distribution in channel c:

$$p_c(\Delta) = (1-q)\frac{1}{2\lambda_c}\exp\left(-\frac{|\Delta|}{\lambda_c}\right) + qU(\Delta;\eta_c), \qquad (1)$$

where $0 < q < 1$ is a partition coefficient of the mixture and λ_c is the Laplacian distribution parameter.

$U(\Delta;\eta_c)$ is a uniform distribution with range η_c. In a low-frequency channel, we set the length of the range as the wavelength of the center frequency. In high-frequency channels, however, $U(\Delta;\eta_c)$ is the uniform distribution over all possible pitch periods (between 2 ms and 12.5 ms in our system).

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

$$\lambda_c = a_0 + a_1 c. \qquad (2)$$

The maximum likelihood method is utilized to estimate the three parameters a_0, a_1, and q in low- and high-frequency channels separately. As can be seen in Fig. 1, the estimated distribution fits the histogram very well.

Likewise, similar statistics are extracted for time frames with two pitch periods. We redefine the relative time lags as relative to the pitch period of the dominant source in a channel. The probability distribution of relative time lags with two pitch periods is denoted as $p'_c(\Delta)$.

Table 1: Categorization of interference signals.

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

2.4. Integration of periodicity information

The state space of pitch is a union-space Ω consisting of three subspaces:

$$\Omega = \Omega_0 \cup \Omega_1 \cup \Omega_2 , \tag{3}$$

where Ω_0, Ω_1, Ω_2 are zero, one, and two dimensional spaces representing zero, one, and two pitches, respectively. This section derives the conditional probability $p(\Phi \,|\, x)$ given a pitch state x observing the set of selected peaks Φ.

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

The channel conditional probability is derived as

$$p(\Phi_c \,|\, x_1) = \begin{cases} p_c\big(\Delta(\Phi_c, d)\big), & \text{if channel } c \text{ selected} \\ q_1(c)U(0; \eta_c), & \text{otherwise} \end{cases} , \tag{4}$$

where $x_1 \in \Omega_1$ and $q_1(c)$ is the parameter q of channel c estimated from one-pitch frames.

We propose the following formula to combine the information across the channels:

$$p(\Phi \,|\, x_1) \propto \sqrt[r]{\prod_{c=1}^{C} p(\Phi_c \,|\, x_1)} , \tag{5}$$

where $C = 128$ is the number of all channels and the parameter $r = 6$ is the smoothing factor.

Now we consider the hypothesis of two pitch periods, d_1 and d_2. The observation probability is defined as

$$p(\Phi \,|\, x_2) \propto \alpha_2 \max\left(\sqrt[r]{\prod_{c=1}^{C} p_2(\Phi_c, d_1, d_2)}, \sqrt[r]{\prod_{c=1}^{C} p_2(\Phi_c, d_2, d_1)} \right),$$
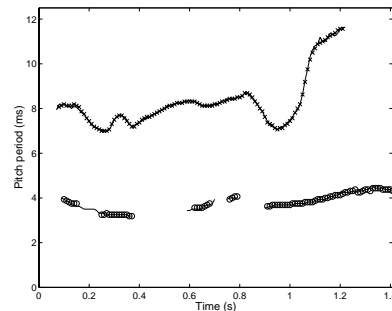
where

$$p_2(\Phi_c, d_1, d_2) =$$
$$\begin{cases} q_2(c)U(0; \eta_c), & \text{if channel } c \text{ not selected} \\ p_c'\big(\Delta(\Phi_c, d_1)\big), & \text{if } |\Delta(\Phi_c, d_1)| < \beta\lambda_c \\ \max\big(p_c'(\Delta(\Phi_c, d_1)), p_c'(\Delta(\Phi_c, d_2))\big), & \text{otherwise} \end{cases} \tag{6}$$

with $x_2 \in \Omega_2$, $\beta = 5.0$, $\alpha_2 = 1.7 \times 10^{-5}$ and $q_2(c)$ denotes the parameter q of channel c estimated from two-pitch frames.

Finally, we fix the probability of zero pitch,

$$p(\Phi \,|\, x_0) \propto \alpha_0 , \tag{7}$$

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

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

2.5. Pitch tracking using an HMM

Our approach utilizes a hidden Markov model for approximating the generation process of harmonic structure in natural environments. The hidden nodes represent possible pitch states in every time frame. The observation nodes represent the set of selected peaks in each time frame. The temporal links in the Markov model represent the probabilistic pitch dynamics. The links between a hidden node and an observation node are called observation probabilities, which have been formulated in the last section representing bottom-up pitch estimation.

There are two parts of probabilistic pitch dynamics. The first part is the dynamics of a continuous pitch track. The changes of the pitch periods in consecutive time frames can be modeled as a Laplacian distribution estimated from pitch tracks of natural speech utterances using the maximum likelihood method. The second part is the probabilities of jumping between the state spaces of zero pitch, one pitch, and two pitch and they can also be estimated from pitch tracks of speech signals and their mixtures.

Finally, the state spaces of one and two pitch are discretized and the Viterbi algorithm is employed for finding the optimal sequence of states.

In our model, there are a total of eight free parameters: four for channel/peak selection and four more employed for bottom-up estimation of observation probability (their values were given earlier). We note that there is a considerable range from which to choose appropriate values for these parameters, and consequently the system performance is not sensitive to specific parameter values.

3. RESULTS AND COMPARISON

A corpus of 100 mixtures of speech and interference [1] commonly used for CASA research has been used for system evaluation and model parameter estimation. The mixtures are obtained by mixing 10 voiced speech utterances with 10 interference signals representing a variety of acoustic sounds. As shown in Table 1, the interference signals are further classified into three categories: 1) signals with no pitch, 2) signals with some pitch qualities, and 3) speech. Half of the mixtures and clean speech utterances are employed for parameter estimation described in the last section. The other half of the mixtures are

Table 2a: Results of speech mixed with Category 1 interference. All error measures are in percentage.

	$E_{0\to1}$	$E_{0\to2}$	$E_{1\to0}$	$E_{1\to2}$	E_{Gross}	E_{Total}	E_{Fine}
Proposed PDA	0.36	Nil	6.81	Nil	Nil	7.17	0.43
TK PDA	1.96	0.05	23.3	9.10	2.38	27.66	1.76

Table 2b: Results of speech mixed with Category 2 interference. (Measured in %)

	$E_{1\to0}$	E_{Gross}	E_{Total}	E_{Fine}
Proposed PDA	3.18	0.32	3.50	0.44
TK PDA	7.70	4.53	12.23	1.41

Table 2c: Results of speech mixed with Category 3 interference. (Measured in %)

	$E_{0\to1}$	$E_{0\to2}$	$E_{1\to0}$	$E_{1\to2}$	$E_{2\to0}$	$E_{2\to1}$	E_{Gross}	E_{Fine}	E_{Gross}^{Dom}	E_{Fine}^{Dom}
Proposed PDA	0.68	Nil	0.88	0.16	Nil	27.08	0.21	0.33	0.93	0.21
TK PDA	0.47	0.10	2.64	4.55	1.19	26.84	2.33	0.99	4.28	0.69

used in performance evaluation. Our results show that the proposed algorithm reliably tracks pitch points in various situations, such as one speaker, speech mixed with other acoustic sources, and multiple speakers. As an example, Fig. 2 shows our result of tracking two simultaneous utterances of a male speaker and a female speaker.

To measure progress, it is important to provide a quantitative assessment of PDA performance. Due to the number of cases involved, it is a nontrivial task to develop appropriate pitch determination measures. As a result, we measure determination errors separately for the three interference categories because of their distinct pitch properties. Manual pitch tracks obtained from clean speech utterances are used as ground truth. We denote $E_{x\to y}$ as the error rate of time frames where x pitch periods are misclassified as y pitch periods. The gross detection error rate E_{Gross} is defined as the percentage of time frames in which detected pitch frequencies are more than 20% from the ground truth. Fine detection error E_{Fine} is defined as the average frequency deviation from the ground truth for those time frames without gross detection error. The error measures for the three interference categories are shown in Table 2a-c respectively.

To put our result in perspective, we compare with a recent multi-pitch detection algorithm proposed by Tolonen and Karjalainen [6]. Their model generates enhanced summary autocorrelation functions (ESAFs) and the decisions on the pitch periods as well as the number of them are based on the most and the second most prominent peaks of the ESAFs. We refer to this PDA as the TK PDA.

For speech signals mixed with Category 1 interferences, a total gross error E_{Total} is defined as the sum of $E_{0\to1}$, $E_{0\to2}$, $E_{1\to0}$ and E_{Gross}. Our algorithm has the total gross error of 7.17% while theirs has 27.66%.

Since our main interest is to produce the pitch contours of speech utterances, for Category 2 mixtures, only $E_{1\to0}$ is measured and the total gross detection error E_{Total} is the sum of $E_{1\to0}$ and E_{Gross}. As can be seen, for our algorithm the total gross error is 3.50%, and theirs is 12.23%.

Category 3 interferences are also speech utterances. For some applications, such as CASA, the determination accuracy of dominating pitch is of primary interest. Therefore, total gross error E_{Gross}^{Dom} and fine error E_{Fine}^{Dom} of dominating pitch periods are also shown in Table 2c. Our algorithm yields the total gross error

rate of 0.93% for dominating pitch. Their corresponding error rate is 4.28%.

These results show that our algorithm outperforms the TK algorithm significantly. We also compared the performance of our algorithm with that of the HMM-based system by Gu and Bokhoven [2]. The improvement of the performance is equally large.

4. CONCLUSION

We have proposed a novel algorithm that can reliably produce single and double pitch tracks in noisy acoustic environments. A combination of several ideas enables our algorithm to perform robustly. First, an improved channel and peak selection method removes the corrupted channels and invalid peaks. Second, a statistical integration method utilizes the periodicity information across different channels. Finally, an HMM is employed for realizing the pitch continuity constraint.

REFERENCES

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

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

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

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

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

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

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