

A MULTIPITCH TRACKING ALGORITHM FOR NOISY AND REVERBERANT SPEECH

Zhaozhang Jin and DeLiang Wang

Department of Computer Science and Engineering
& Center for Cognitive Science
The Ohio State University
{jinzh, dwang}@cse.ohio-state.edu

ABSTRACT

Determining multiple pitches in noisy and reverberant speech is an important and challenging task. We propose a robust multipitch tracking algorithm in the presence of both background noise and room reverberation. A new channel selection method is utilized in conjunction with an auditory front-end to extract periodicity features in the time-frequency space. These features are combined to formulate frame level conditional probabilities given each pitch state. A hidden Markov model is then applied to integrate these probabilities and search for the most likely pitch state sequences. The proposed approach can reliably detect up to two simultaneous pitch contours in noisy and reverberant conditions. Quantitative evaluations show that our system significantly outperforms existing ones, particularly in reverberant environments.

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 ([1], Ch. 2). 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 attempted to detect multiple pitches simultaneously. Wu *et al.* [2] modeled pitch period statistics on top of a channel selection mechanism and used a hidden Markov model (HMM) to extract continuous pitch contours. More recently, Klapuri [3] proposed an “estimation and cancelation” model that iteratively detects pitch points for polyphonic music and speech signals.

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 ([1], Ch. 7). 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 [4] 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, followed by a conclusion section.

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

Pitch, which originally refers to a percept, has been widely used in computational literature to equate fundamental frequency (or period). For voiced speech, the fundamental frequency is usually defined as the rate of vibration of the vocal folds. 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 [5], 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.

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 [6]. 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.

3. FRONT-END PROCESSING

The input signal $x(t)$ is first passed through a gammatone filterbank with 128 channels whose center frequencies are quasi-logarithmically spaced from 80 Hz to 5000 Hz. The response of each filter channel is further transduced by the Meddis model of auditory nerve transduction. In each channel, the output is then divided into 20-ms time frames with 10-ms frame shift. We use $u_{c,m}$ to denote a time-frequency (T-F) unit for frequency channel c and time frame m . The normalized correlogram $A(c, m, \tau)$ for T-F unit $u_{c,m}$ with a time delay τ is then computed by the running normalized autocorrelation. 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.

To select less corrupted channels from the correlogram for robust pitch analysis [7], we define the cross-channel correlation between $u_{c,m}$ and $u_{c+1,m}$ as $C(c, m)$, which 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

$$C_m = \{c : C(c, m) > \theta_c\} \quad (1)$$

where $\theta_c = 0.95$ is a threshold.

We also calculate the percentage of energy belonging to selected channels in each frame as

$$\xi_m = \sum_{c \in C_m} E(c, m) / \sum_c E(c, m) \quad (2)$$

where $E(c, m)$ is the energy calculated as the sum of squares of the filter response within $u_{c,m}$. We observe that 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.

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 [2]

$$\mathcal{S} = \mathcal{S}_0 \cup \mathcal{S}_1 \cup \mathcal{S}_2 \quad (3)$$

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 τ_1 and τ_2 to represent first and second pitch candidates.

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

$$f_m(\tau_1) = \frac{\sum_{c \in C_m} A(c, m, \tau_1) \log E(c, m)}{\sum_{c \in C_m} \log E(c, m)}. \quad (4)$$

The logarithmic operation acts like a pre-emphasis filter 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

$$p(\mathcal{O}_m | s_1) = \kappa f_m(\tau_1) \quad (5)$$

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 $s_2 = \{\tau_1, \tau_2\}$.

Since detecting multiple pitches is related to sound separation [1], we employ the notion of ideal binary mask [4] 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:

$$\begin{aligned} C_{m,1} &= C_m \cap \{c : A(c, m, \tau_1) \geq A(c, m, \tau_2)\}, \\ C_{m,2} &= C_m \cap \{c : A(c, m, \tau_1) < A(c, m, \tau_2)\}. \end{aligned} \quad (6)$$

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 (7):

$$\begin{aligned} g_m(\tau_1, \tau_2) &= \\ &= \frac{\sum_{c \in C_{m,1}} A(c, m, \tau_1) \log E(c, m) + \sum_{c \in C_{m,2}} A(c, m, \tau_2) \log E(c, m)}{\sum_{c \in C_{m,1}} \log E(c, m) + \sum_{c \in C_{m,2}} \log E(c, m)}. \end{aligned} \quad (7)$$

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. One appealing property of g_m is that room reverberation hardly affects the peak formation near the real pitch periods, which is illustrated by Fig. 1.

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

$$g'_m(\tau_1, \tau_2) = (g_m(\tau_1, \tau_2) + \delta_m)^\gamma - \delta_m \quad (8)$$

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 . We find that the choice of γ is robust to reverberation.

Finally, we define the conditional probability as

$$p(\mathcal{O}_m | s_2) = \kappa (g'_m(\tau_1, \tau_2) - H(\beta - \xi_m) \cdot \lambda) \quad (9)$$

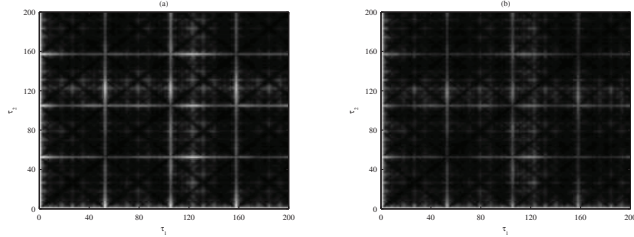


Fig. 1. The pitch salience function g_m in one time frame in a mixture of two speakers. Plot (a) corresponds to the anechoic condition and plot (b) the reverberant condition. Brighter color indicates higher salience. The two plots show a similar pattern and similar peak locations.

where it penalizes g'_m when $\xi_m \leq \beta$. $H(\cdot)$ is the Heaviside step function and $\lambda = 0.05$ is the amount of penalty. As mentioned in Section 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. We find $\beta = 0.65$ is appropriate to discriminate the above cases. 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.

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

$$p(\mathcal{O}_m | s_0) = \kappa \cdot \begin{cases} 1 & \text{if } \min(f_m) > \theta_s, \\ \eta & \text{else if } \text{var}(f_m) < \theta_b, \\ 0 & \text{else.} \end{cases} \quad (10)$$

In (10), the first case handles silence and unvoiced speech. 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. In contrast, a harmonic source should exhibit a peaky distribution (high variance) in f_m . 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 (10) 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 [2]. The hidden states are from the state space defined earlier in Section 4. The state transition probabilities have two aspects: jump probabilities and pitch continuity. We use the same set of parameters as in [2] for all reverberant cases because they do not need to be exact and work well within a considerable range. The observation probability distributions are already given in (5), (9) and (10).

Table 1. Error rates (in %) for three interference categories

T_{60} (s)/System	CATEGORY 1		CATEGORY 2		CATEGORY 3	
	E_{tl}	E_{fn}	E_{tl}	E_{fn}	E_{tl}	E_{fn}
0.0 Wu <i>et al.</i>	7.24	1.21	5.79	1.27	24.75	1.01
	Proposed	9.62	1.22	3.26	1.44	14.39
0.3 Wu <i>et al.</i>	11.55	1.32	8.23	1.54	38.34	1.29
	Proposed	10.63	1.58	4.09	1.80	24.71
0.6 Wu <i>et al.</i>	15.32	1.69	14.96	1.89	54.11	2.18
	Proposed	10.35	2.06	5.67	2.48	33.39

6. EXPERIMENTAL RESULTS

We use Cooke's corpus [8], 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. Following [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.

To generate reverberant recordings, 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, each is specified by two locations for two sources (target and interference) and another location for the microphone. 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 [6] on reverberant speech signals before mixing, as described in Section 2. This technique is not error free. However, in our experiment, it is harmless to have some errors in the reference pitch contour since the PDA under evaluation will have a performance inferior to the reference PDA [9].

We follow [2] to formulate a quantitative measure of PDA performance. Only total and fine errors (E_{tl} and E_{fn}) are reported: the former is the combination of transition errors and gross errors (E_{gs} , with a 20% criterion), and the latter is defined as the average deviation from the reference pitch for those frames without gross errors. Table 1 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 not superior according to the E_{fn} measure. This is because a lower rate of total errors makes it harder to avoid fine errors. Also, E_{fn} can be 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, E_{fn} indicates that our algorithm has smaller fine errors in all three T_{60} 's. Fig. 2 plots the pitch contours detected by Wu *et al.*'s and the proposed algorithm. 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.

To compare with Klapuri's algorithm requires prior information

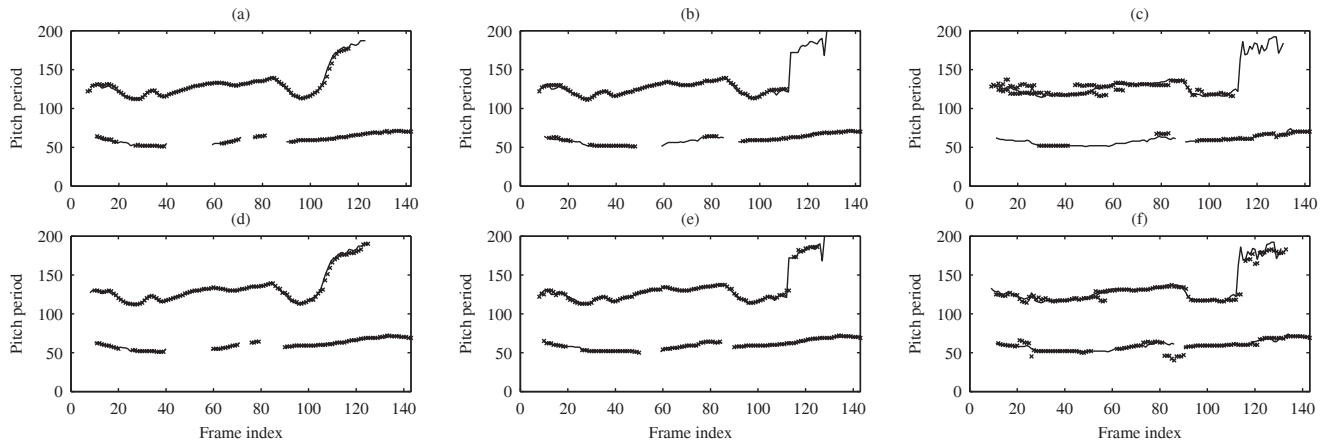


Fig. 2. Pitch tracking results for a mixture of one male and one female utterance. (a)–(c) plot detected pitch contours from Wu *et al.*'s algorithm, and (d)–(f) are from the proposed algorithm. Each column from left to right corresponds to $T_{60} = 0.0, 0.3$ and 0.6 s, respectively. The solid lines indicate the reference pitch tracks. The “x” tracks represent the estimated pitch contours.

of the number of pitches in each frame since it lacks the ability to do so reliably. 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. Table 2 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. For fine errors, three algorithms have comparable results in the first category. However, in Category 3, our algorithm 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.

7. CONCLUSION

This paper has proposed a multipitch tracking system for both noisy and reverberant conditions. Several ideas contribute to achieve the robust performance. Firstly, reliable channel selection is carried out. Secondly, the salience functions are formulated to model the likelihood of the observed correlogram being explained by each given pitch state. Finally, an HMM is responsible for choosing appropriate pitch hypotheses as well as forming continuous pitch tracks.

Acknowledgements. 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).

8. REFERENCES

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

[2] 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.

Table 2. Error rates (in %) with prior pitch number for two interference categories

		CATEGORY 1		CATEGORY 3	
T_{60} (s)/System		E_{gs}	E_{fn}	E_{gs}	E_{fn}
0.0	Wu <i>et al.</i>	1.16	1.45	2.80	1.40
	Klapuri	0.74	1.57	4.82	1.37
	Proposed	0.09	1.61	0.59	1.10
0.3	Wu <i>et al.</i>	2.62	1.90	7.20	2.00
	Klapuri	5.16	1.93	21.00	1.74
	Proposed	0.50	2.13	5.10	1.48
0.6	Wu <i>et al.</i>	4.11	2.48	18.48	3.18
	Klapuri	7.17	2.68	29.12	2.51
	Proposed	1.34	2.56	11.92	2.25

[3] 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.

[4] 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.

[5] 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.

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

[7] 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.

[8] M. P. Cooke, *Modeling auditory processing and organization*. Cambridge, UK: Cambridge Univ. Press, 1993.

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