J. King Saud University, Vol. 22, Comp. & Info. *Sci*., pp. 13-28,Riyadh (1431H./2010)

Noise-Robust Pitch Detection using Auto-correlation Function with Enhancements

Ghulam Muhammad
Dept. of Computer Engineering, College of Computer and Information Sciences
King Saud University, P.O. Box: 51178, Riyadh 11543, Saudi Arabia.
Email: ghulam@ksu.edu.sa

(Received 04/05/2008; accepted for publication 06/05/2009)

Keywords: Pitch, autocorrelation, voiced/unvoiced, speech analysis.

Abstract. An efficient noise-robust pitch detection algorithm is proposed in this paper. The algorithm is based on time domain autocorrelation function (ACF). A bank of band-pass filters is used for competitive contribution of periodicity to select primary pitch candidates. A weighting criterion that involves both increase and decrease in merit is applied to the candidates by exploiting the presence or the absence of pitch harmonics. Finally, a simple enhancement is integrated to smooth the pitch contour. The proposed algorithm is evaluated on TIMIT database in different types and levels of noise in terms of pitch and voice activity detection. The experimental results show the superiority of the proposed method over well known other methods.

1. Introduction

The fundamental frequency of a sound, whose percept is called pitch, has great importance in many areas. Pitch detection is one of the oldest, yet unsolved topic among the researchers of speech and music. Accurate pitch detection is essential to areas such as speech coding [1], speech synthesis [2], to more recent topic of speaker emotion recognition [3].

The automatic tracking of pitch has multiple applications in the field of speech processing and speech technology. Pitch contour is useful in assisting hearing impaired people [4]. Pitch determination might facilitate the diagnosis of aphasia and dysarthria as well as be integrated in computer aided pronunciation teaching systems. One could continue to enumerate many other potential applications based on the automatic determination of pitch [5].

Consequently, a wide range of perceptual models and algorithms using a variety of techniques and a varying degree of accuracy to extract pitch exist [6]. However, the pitch detection algorithms (PDAs) face a real challenge in the presence of noise [7].

There are three types of PDAs in the literature: *time-domain, frequency domain,* and *time-frequency domain*. Time domain method includes the short-time average magnitude difference function (AMDF) [8], short-term autocorrelation function (ACF) [6], etc; frequency domain method includes harmonics enhancement based on instantaneous frequency [9], and cepstrum analysis [10]; while pitch detection based on Hilbert-Huang transform [11] falls in time-frequency domain method. Among all the methods, ACF-based algorithms are simpler to implement and robust against noise. However, AMDF-based algorithms have less computational complexity and are used in real time processing. In this paper, we focus on an ACF-based pitch detection algorithm for its accuracy against noise. Because of the periodic nature of voiced speech, its ACF is also periodic with period equal to the pitch value. ACF shows peaks at pitch and its harmonics locations. Natural speech is not absolutely periodic, rather it is quasi periodic. Hence ACF produces the highest peak at pitch period, and gradually decreasing peaks at its harmonics. Thereby, the highest peak other than zero location in ACF corresponds to the pitch period. However conditions like presence of noise, quasi periodic nature of speech signal, peaks due to detailed formant structure of vocal tract affect the location of the highest peak in ACF. A weighted ACF method using AMDF has been proposed in [7], but this method suffers from double pitch error at low signal to noise ratio (SNR). One bit ACF based on AMDF can be found in [12], though it is not evaluated in noisy condition. A modification to the basic autocorrelation is termed as normalized cross correlation function (NCCF) and it is introduced in [13]. As reported, NCCF is better suited for pitch detection than the standard ACF as the peaks are more prominent and less affected by the rapid variations in the signal amplitude. In [14], the performance of NCCF is further enhanced by exploiting the existing of a second large peak at double the true pitch position. AMDF-based methods have shortcoming associated with the falling trend in minima, and have degraded performance in noisy environments. Several modifications of AMDF have also been proposed in literature. For example, high resolution AMDF (HRAMDF) [1] and circular AMDF (CAMDF) [15] conquer the falling trend, but at the expense of introducing new double pitch error. Despite many methods have been proposed so far to extract true pitch after tackling these issues [7, 14, 16, 17, and 18], a more noise-robust and efficient pitch tracking method is necessary for advanced speech processing algorithms.

In order to decrease the error rate in pitch detection as well as to increase the performance of voice activity detection (VAD) under severe noisy condition, this paper proposes a new ACF based pitch detection with several enhancements. The novelty of this paper is as follows: (a) addition of ACFs from different band-pass filters (BPFs), (b) increasing or decreasing the weight of a pitch candidate corresponding to the presence or the absence of pitch harmonics at its multiple locations, and (c) a sophisticated smoothing of pitch particularly at the beginning and the ending of a voice segment. Different BPFs are used to contribute relative amount of periodicity at different frequency ranges. Increase or decrease of weight is supposed to increase the possibility of finding the true pitch, while suppressing double or half pitch error. The weight assignment uses the fact that a true pitch has other peaks at its multiple integer locations in ACF. Under noisy condition, pitch information is buried towards the both ends of a voiced segment and hence a smoothing technique is applied considering nearby pitch information. A preliminary work related to this can be found in [20], and the current version is an extensive and more elaborative one comparing to that.

This paper is organized as follows: Section 2 describes some basic features of pitch that include time domain processing. Section 3 presents the proposed pitch detection algorithm, and Section 4 gives experimental results with discussion. Finally, Section 5 draws some conclusion.

2. Basic Features of Pitch

2.1. Pitch definition

Pitch or fundamental frequency is the lowest frequency component of a signal that excites to a system (for example, vocal system). The pitch period, which is the inverse of fundamental frequency, is the smallest repeating unit of a signal. One such period describes the periodic signal (voiced part of speech) completely.

2.2. Extracting pitch in time-domain

The fact that variations in voiced signal are so evident suggests that the *time-domain* techniques should be capable in detecting pitch period of a voiced signal. Most of the time-domain pitch period estimation techniques use auto-correlation function (ACF).

2.2.1. Autocorrelation function

The basic idea of correlation-based pitch tracking is that the correlation signal will have a peak of large magnitude at a lag corresponding to the pitch period.

A short-time ACF for a signal $s[m]$ is computed as:

$$R[k] = \sum_{m=0}^{N-k-1} s[m]s[m+k] \qquad (1)$$

where, N is total number of samples in a window, and k is the lag index. The choice of window length N for calculating $R[k]$ has conflicting requirements:

- N should be as small as possible to show time variation;
- N should be large enough to cover at least 2 periods so that periodicity can be captured by $R[k]$.

Properties of $R[k]$ include:

a. Same periodicity as the s[m].
b. Maximum value at k = 0 and R[0] is equal

to energy of deterministic signal.

c. If s[*m*] is periodic with period of *P* samples, *R*[*k*] has maximum at *k* = 0, ±*P*, ±2*P*,….

2.2.2. Center clipping

The ACF may contain too much information, most of which is not related to the fundamental frequency. For pitch detection, speech signal is usually pre-processed to make the periodicity more prominent and to suppress other distracting features. Such techniques are often called *spectrum flattening*. Center clipping is the most popular spectrum flattening technique, and can be expressed as Eq. (2). A choice of clipping level (C in Eq. (2)) should fulfill the following criterion:

- should be high enough to eliminate all distracting peaks, but
- cannot be too high so as not to lose desirable peaks.

$$C\{s(n)\} = \begin{cases} s(n), & s(n) > +C \\ 0, & -C \le s(n) \le +C \\ s(n) & s(n) < -C \end{cases} \quad (2)$$

Fig. 1. Block diagram of the preprocessing steps of the proposed pitch detection algorithm (PDA).

Usually, the clipping level is chosen to be 60%-80% of the maximum amplitude and is adaptively adjusted according to the signal level.

3. The Proposed Pitch Detection Algorithm

3.1 Preprocessing using autocorrelation function

Figure 1 shows preprocessing steps for the proposed PDA. Input speech is at first passed through a bank of band-pass filters. Center frequencies of the lowest and the highest bands are 50 Hz and 1 kHz, respectively. It is known that pitch periodicity cannot be observed in high frequency channels, and hence the frequency components above 1 kHz are filtered out. Each filter output is then half-wave rectified and center-clipped. Half-wave rectification mimics phase-locking property of human auditory system. Center clipping is performed to simulate spectrum flattening. For center clipping, the minimum of the maximum amplitudes of the first one-third samples and the last one-third samples in a frame is determined. Then the clipping level is set to 75% of that minimum value. Figure 2 shows a demonstration of half-wave rectification and center clipping.

An ACF, shown in Eq. (3), where s_i is the *i*-th filter output of speech signal, *N* is total number of samples in a frame, *m* and *M* are the lag index and the total number of lag position, respectively, is then applied to the center-clipped output to give an auto-correlogram. A summary auto-correlogram is obtained by summing up all the auto-correlograms using Eq. (4), where I_p corresponds to the total number of filters used for pitch detection. The summary auto-correlogram is normalized by dividing the auto-correlogram values by that at lag 0 (Eq. (5)). The value at lag 0 corresponds to energy level and it is the maximum in a frame. A noise-robust, non-delayed PDA is then applied to summary auto-correlogram to extract pitches. In the algorithm, if the pitch is equal to zero for a certain frame, the frame is considered to be an unvoiced/silent frame, otherwise the frame is a voiced frame.

Fig. 2. A demonstration of half-wave rectification and center clipping.

$$r_i[m] = \frac{1}{N-m}\sum_{n=0}^{N-m-1} s_i[n] \times s_i[n+m], \; m = 0,1,\ldots,M \quad (3)$$

$$R[m] = \sum_{i=1}^{I_P} r_i[m] \quad (4)$$

$$\overline{R}[m] = \frac{R[m]}{R[0]} \quad (5)$$

Figure 3 shows an example of the summary auto-correlogram and the auto-correlogram obtained when ACF is applied directly to the speech signal without filtering. From the figure, we can see that the summary auto-correlogram has clearly distinguished peaks at the lags multiple to the pitch period. However, the auto-correlogram which is obtained when ACF is directly applied to input speech has many spurious peaks that can have negative impact in pitch detection. Multi-band can fully exploit the contribution of each channel to pitch, and thereby can enhance the true pitch and suppress the false candidates. This is why multi-channel summary auto-correlogram is preferred to single-channel auto-correlogram.

Fig. 3. Illustration of advantage of using the summary auto-correlogram, spanned over several filters, rather than using ACF directly on the input speech without filtering.

3.2 Basic steps of the proposed PDA

The basic idea of auto-correlation based pitch tracking is that the correlogram will have a large peak at the lag corresponding to the pitch period. However, in actual cases, many large peaks may exist at lags corresponding to half or double pitch periods or at random locations. A noise-robust PDA is proposed to overcome the shortcomings for pitch detection. A flow chart of the proposed PDA is given in Fig. 4. Possible pitch candidates are extracted from the summary auto-correlogram by using the four basic steps (steps i-iv). An actual pitch is then detected by using enhancement blocks (A) and (B). The frame length is set to 35 ms to capture at least one large peak at 2nd multiple lag corresponding to the true pitch period. The proposed PDA is designed to extract any pitch within the range between 2.5 ms and 16 ms.

Fig. 4. Flow chart of the proposed pitch detection algorithm.

The basic steps of the proposed PDA are as follows:

i. Find local maximums (L_{max}s) from the summary auto-correlogram, while ignoring the peaks for the first 2.5 ms. Any value in the summary auto-correlogram, which is greater than the values at ±3 lags, and higher than a threshold, θ_1, is extracted as L_{max}.

ii. Find global maximums (G_{max}s) from the L_{max}s. Any L_{max} which is greater than the L_{max}s with some threshold, θ_2, at ±2 ms, is selected as G_{max}. The L_{max} s that are below θ_2% of G_{max} within ±2 ms are 'deleted'.

iii. G_{max}s are assigned *merit* values or weight as confidence of their pitch candidacy. Initially, the merit values of all the G_{max}s are set to zero. Then, increase the merit of those G_{max}s that have other G_{max}s at their multiple (up to 4th multiple) lags with an offset. The merit is increased by $4/w$, where w is an integer (w = 2, 3, 4) that corresponds to 2nd multiple, 3rd multiple, or 4th multiple. The presence of the G_{max}s at the multiple lags is evidence that the G_{max} at lower lag is a strong candidate for true pitch. The offset is set to $(p_c/2)\times(w/f_s)$, where p_c is the lag for the G_{max} that is a pitch candidate, and f_s is the sampling rate of the speech signal in kHz. For example, for a p_c located at lag 159, the merit is increased by $4/3$, if there is a G_{max} at lag $(159\times3)\pm((159/2)\times(3/16))$ [between lag 462 to lag 492], when the speech signal is sampled at 16 kHz rate.

   The proposed PDA makes a good use of the peaks located at multiple lags of a pitch candidate. An intensive observation shows that, for noisy data, a false pitch candidate may have other G_{max}s at its 2nd, 3rd, or 4th multiple lags. To correctly detect a pitch candidate, the proposed PDA not only increases the candidacy of a G_{max} by corresponding merit, regarding to the presence of other G_{max}s at its multiple lag, but also decreases the candidacy in the case where no G_{max} is available at any of its multiple lag. For example, if a candidate does not have a G_{max} at its 2nd multiple lag, but has a G_{max} at its 3rd multiple lag, the merit of that candidate will be updated by $(-(4/2)+(4/3))$.

iv. Find the G_{max} with the maximum merit value. The lag of that G_{max} corresponds to the pitch period.

An example of extracting L_{max}s, G_{max} s and pitch location is shown in Fig. 5. Fig. 5(a) shows a speech segment with SNR = 5 dB. Corresponding summary auto-correlogram, local maximums, global maximums, and the extracted pitch are shown in Fig. 5 (b), (c), (d), and (e), respectively.

Fig. 5. Illustration of the basic steps of the proposed pitch detection algorithm.

3.3 Enhancements on the proposed PDA

The basic steps (steps i-iv) alone do not necessarily provide much accuracy in pitch detection, particularly in noisy environments. These steps can do a fair job if we use information of only a single frame. However, incorporating the result obtained from the previous frame may further help. Hence, some enhancements are adopted into the proposed PDA. The enhancement procedures are shown in blocks marked (A) and (B) in Fig. 4.

The decision block marked (A) checks for any undesirable pitches resulted from noise, or for any half-pitch or double-pitch errors. If any half-pitch or double-pitch error is found by comparing pitch period of the previous frame, P_{t-1}, the pitch is adjusted by taking twice or half the current lag, respectively. We choose pitch information of only one previous frame to reduce complexity.

The decision block marked (B) eliminates the possibility of finding 'no pitch' towards the end of voiced segments. If a pitch is found in the previous frame, but no pitch in the current frame, P_t, this block checks whether there is a large G_{max} in the current frame at around (or twice) the lag similar to the pitch position in the previous frame. If such a large G_{max} is found, then a pitch is set for the current frame. Figure 6 illustrates two examples of the enhancements described above.

The proposed PDA adopts a non-delayed approach. It means that while determining pitch at frame *t*, it does not check for any information of succeeding frames, i.e., frames *t*+1, *t*+2, etc. As pitches do not change abruptly in successive frames, the proposed PDA checks only the pitch of the previous frame, *t*-1.

4. Experiments

4.1 Database

Ten English sentences spoken by 2 male speakers and 2 female speakers each from dialect region 6 (dr6: New York City) in TIMIT Acoustic-Phonetic Continuous Speech Corpus [19] are used for the evaluation. White Gaussian noise is added to the clean speech at SNR = 10 dB, 5 dB, and 0 dB. The sampling rate is 16 kHz.

There are a total of 4212 frames of which 2992 are voiced and the rest are unvoiced / silent in the test dataset.

4.2 Experimental setup

Six FIR Hamming BPFs of order 61 are used to calculate the summary auto-correlogram. The center frequencies of the filters range from 50 Hz to 1 kHz, and are uniformly spaced on the Bark scale. Frame length is set to 32 ms, and frame rate is 10 ms. The proposed PDA is then applied on the summary auto-correlogram to find out pitch periods. If a pitch cannot be found for a certain frame, then the frame is considered to be unvoiced or silent (U/S), otherwise the frame is voiced (V). The reference pitch is extracted manually from clean speech after a semi automated method generates a gross approximation of the pitch. The reference pitch is cross checked by four individuals for a final decision. The values of θ_1 and θ_2 described in Section 3.3, are obtained by varying the parameters with the range of 0.01 ~ 0.20 and 70% ~ 95%, respectively.

To verify the effects on the performance of different enhancements of the proposed PDA, we performed experiments using the following variations of the PDA along with a baseline algorithm, which we call RAPT (Robust Algorithm for Pitch Tracking) [17].

Figure 6: Examples for the enhancements of pitch detection. (a) The pitch candidate, located at lag 100, does not have G_{max} near its 3rd multiple lag 300, and hence its merit is decreased. On the other hand, the pitch candidate at lag 196 has a peak near its 2nd multiple lag 400, and 3rd multiple lag is out of frame. So its merit is increased. Also, pitch candidate at lag 196 locates closely to the pitch of previous frame (lag 191). The enhancement eliminates half-pitch error and detects the correct pitch. (b) A large G_{max} is located around double the previous pitch lag, and it helps to detect the current pitch at lag 202.

1) The proposed PDA.
2) PDA without the enhancement blocks (A) and (B) from the proposed PDA. This approach does not check pitch value of the previous frame. The pitch of the current frame solely depends on the information at the current frame.
3) PDA without blocks (A), (B) and 'decrease merit' in step iii (see Fig. 4) from the proposed PDA. The merit of a candidate is only increased in presence of other G_{max}s at multiple lags. This approach becomes a rather conventional auto-correlation-based pitch extraction method [14].
4) RAPT: The KTH's WaveSurfer implementation of a robust algorithm for pitch tacking [13], a method based on normalized cross-correlation and dynamic programming.

4.3 Experimental results and discussion

The experimental results are shown in Tables 1 and 2. The results are given in terms of %Gross Error. 'Gross error' is an error when the generated pitch is not within 1 ms of the reference pitch for a certain frame [6]. For example, if the proposed PDA generated pitch is 10 ms for a frame, while the reference pitch is 3 ms, then a gross error is reported for that frame. Gross error also includes error of detecting voiced frames as unvoiced/silent frames (V to U/S). The values of θ_1 = 0.03 and θ_2 = 90% are found to give the best results and are fixed for the evaluation.

Tables 1 and 2 depict the strength of the proposed PDA for male and female speech, respectively. For male clean speech, the proposed PDA misclassifies only 0.11% of voiced frames as unvoiced or silent frames, while it is only 2.78% for noisy speech with SNR = 0 dB. Without enhancement blocks (A) and (B), the performance is poor even in clean environment. The performance is greatly affected in voiced segments. Also, 'decrease merit', which is a novelty of the proposed PDA, shows positive effect in the experiment. For example, 5.97% of voiced frames are found to have detected as unvoiced or silent frames for variation (2), which includes 'decrease merit' for noisy speech with SNR = 0 dB, while 6.96% of voiced frames are reported to have detected as unvoiced/silent frames for variation (3) that does not include 'decrease merit'. The proposed method also shows superiority over the RAPT algorithm. From Table 1, we can see that the proposed method has only 2.65% gross error compared to 8.11% obtained by the RAPT algorithm at SNR = 0 dB. Similarly, Table 2 shows corresponding improvements of the proposed method for female speech. If we compare Tables 1 and 2, we can find that the PDA performs better for female speech. For example, in SNR = 0 dB condition, the proposed method has 2.65% gross error for male speech, while that for female speech is 2.54%. Figure 7 summarizes average % gross error for male and female speech for different noisy conditions. From Fig. 7 we can see that the proposed PDA has average 2.60% gross error comparing to 8.03% obtained by RAPT.

Table 1. Performance of different methods of pitch detection for male speech. Gross error, V to U/S (voiced to unvoiced/silent error), and U/S to V (unvoiced/silent to voiced error) are given in %.

Table 2. Performance of different methods of pitch detection for female speech. Gross error, V to U/S (voiced to unvoiced/silent error), and U/S to V (unvoiced/silent to voiced error) are given in %.

Fig. 7. Average %Gross Error of male and female speech for different methods.

Fig. 8. Mean FPE (fine pitch error) of different methods for male speech.

Fig. 9. Mean FPE (fine pitch error) of different methods for female speech.

Fig. 10. Comparison of extracted pitch between RAPT and the proposed method. The utterance is contaminated with white noise at SNR = 5 dB.

Figure 8 and 9 show mean fine pitch error (FPE) of male and female speech, respectively, at different SNR. FPE is termed as pitch error of less than 1 ms, and it is defined in Hz. From these figures, we can see that FPE is more for female speech than for male speech. Female voice pitch ranges from 250 Hz to 500 Hz, while that for male voice is from 60 Hz to 180 Hz. Therefore, it is obvious that FPE for female voice pitch will be greater than FPE for male voice pitch. These figures clearly indicate that the proposed method outperforms all other methods in terms of FPE at different noise levels.

Figure 10 shows a comparison of detected pitch of a full length TIMIT sentence (dr6, SA1.wav) contaminated with white noise at SNR = 5 dB. From the figure we see better accuracy of the proposed method in terms of pitch value and boundary detection of voiced segments over RAPT.

5. Conclusion

A noise-robust pitch detection algorithm based on autocorrelation function has been introduced. The method includes several band pass filters to exploit different level of periodicity at different frequency range, weight assignment to pitch candidates based on presence or absence of pitch harmonics in autocorrelation function, and a smoothing technique that suppresses any abrupt changes of pitch in successive frames. Experimental results conclude the followings:

a) Weight assignment is important not only in terms of ‘increase merit’ when there is a presence of peak at integral multiple position of a candidate, but also in terms of ‘decrease merit’ when there is absence of a peak at its integral multiple position.

b) Smoothening over multiple frames is necessary for finding accurate pitch over the utterance.

The effect of other types of real noises on the proposed pitch detection algorithm will be investigated as a future study.

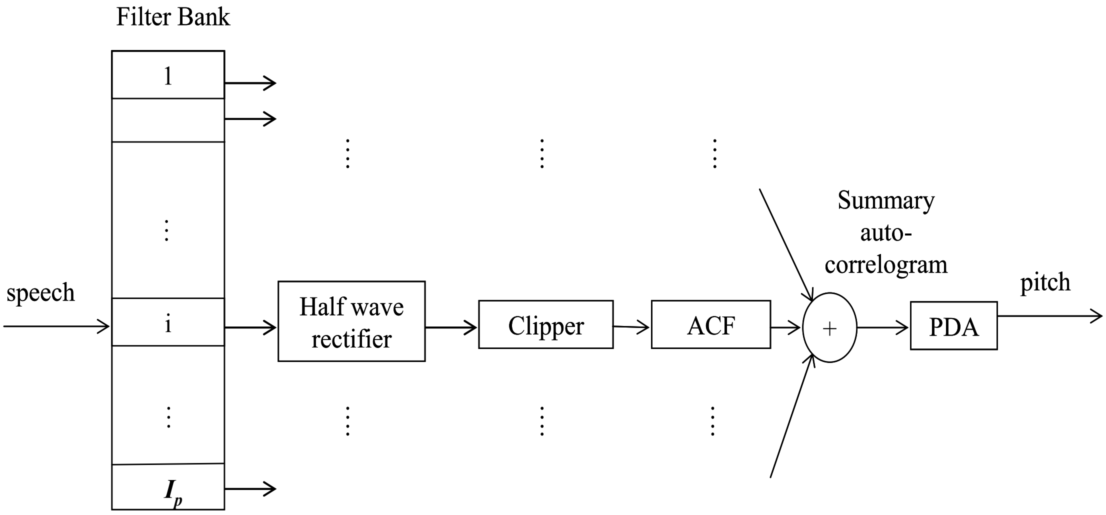
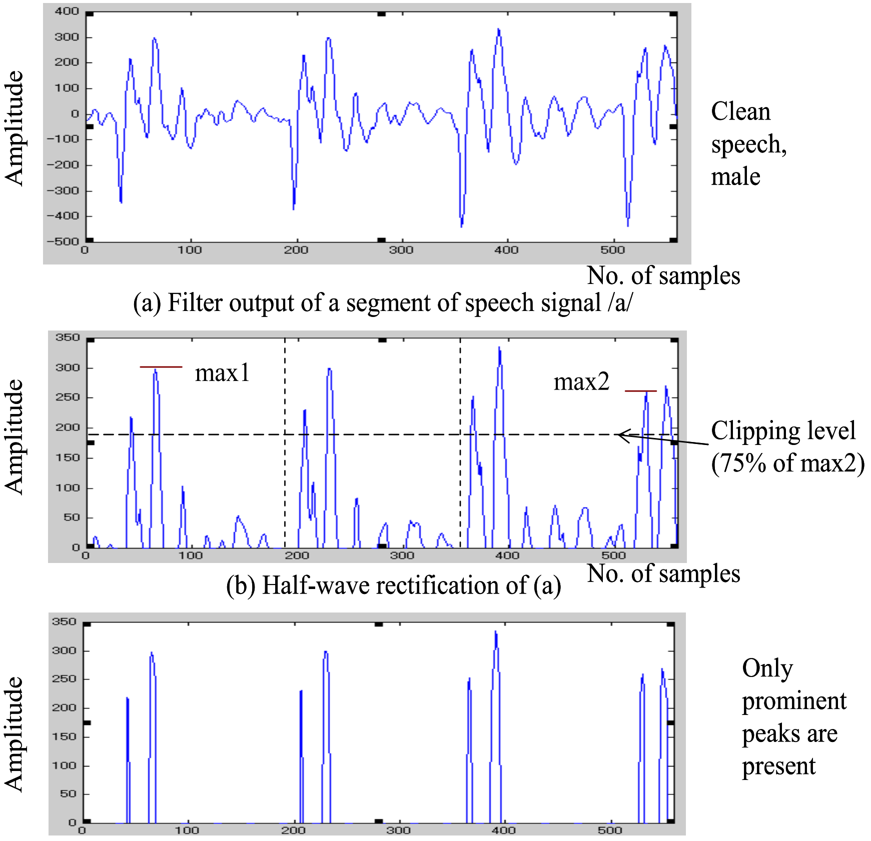


Fig. 1. Block diagram of the preprocessing steps of the proposed pitch detection algorithm (PDA).

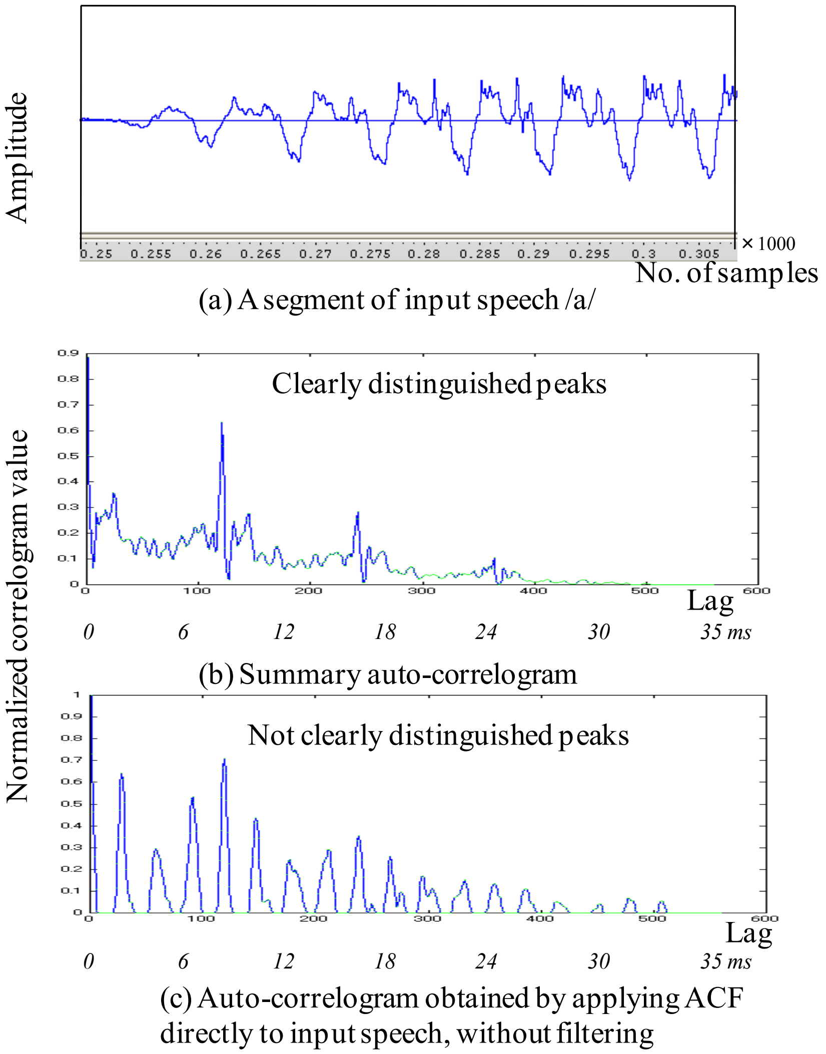


(a) Filter output of a segment of speech signal /a/

(b) Half-wave rectification of (a)

(c) Center clipping on (b)

Fig. 2. A demonstration of half-wave rectification and center clipping.



(a) A segment of input speech /a/

(b) Summary auto-correlogram

(c) Auto-correlogram obtained by applying ACF directly to input speech, without filtering

Fig. 3. Illustration of advantage of using summary auto-correlogram, spanned over several filter, rather than using ACF directly on input speech without filtering. The label of horizontal axis corresponds both to lag number (0 100 200 300 400 500) and time (in ms).

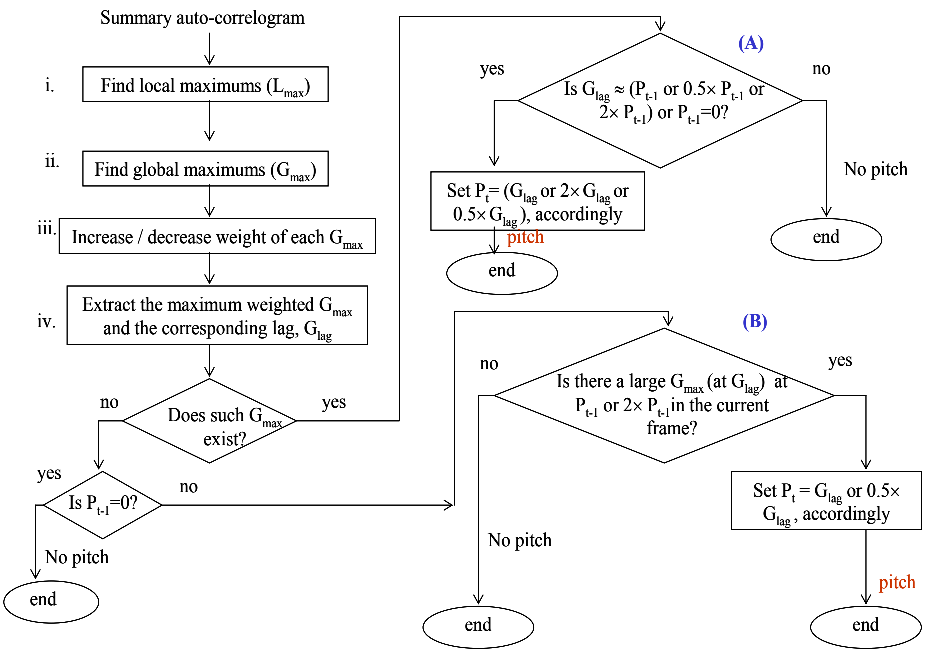

Summary auto-correlogram
i. Find local maximums (L_{max})
ii. Find global maximums (G_{max})
iii. Increase / decrease weight of each G_{max}
iv. Extract the maximum weighted G_{max} and the corresponding lag, G_{lag}
Does such G_{max} exist?
no
yes
Is P_{t-1}=0?
yes
no
No pitch
end
(A)
Is $G_{lag} \approx$ (P_{t-1} or 0.5× P_{t-1} or 2× P_{t-1}) or P_{t-1}=0?
yes
no
Set P_t= (G_{lag} or 2× G_{lag} or 0.5× G_{lag}), accordingly
pitch
end
No pitch
end
(B)
Is there a large G_{max} (at G_{lag}) at P_{t-1} or 2× P_{t-1} in the current frame?
no
yes
Set P_t = G_{lag} or 0.5× G_{lag}, accordingly
No pitch
end
pitch
end

Fig. 4. Flow chart of the proposed pitch detection algorithm.

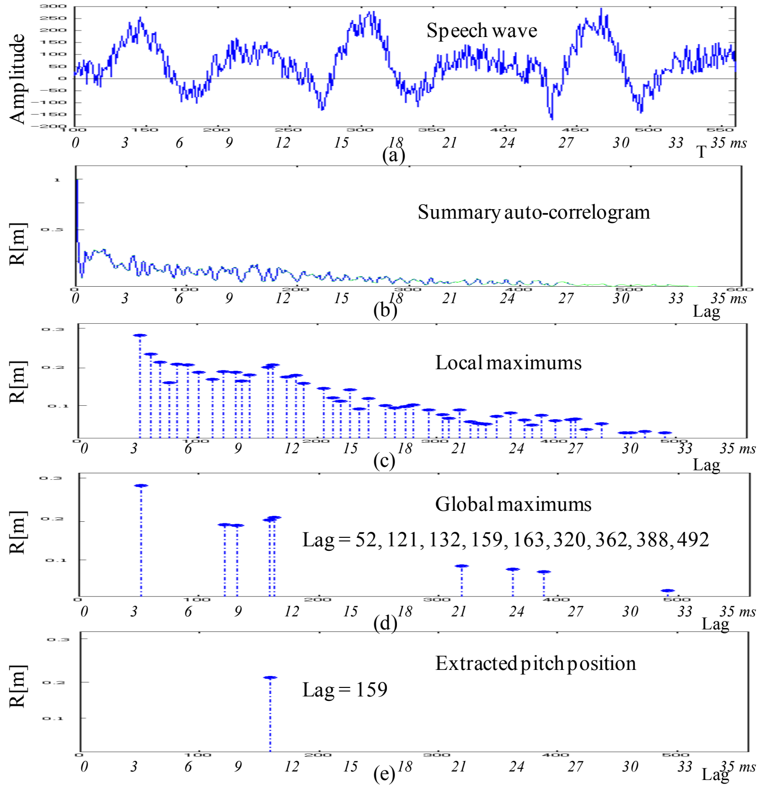
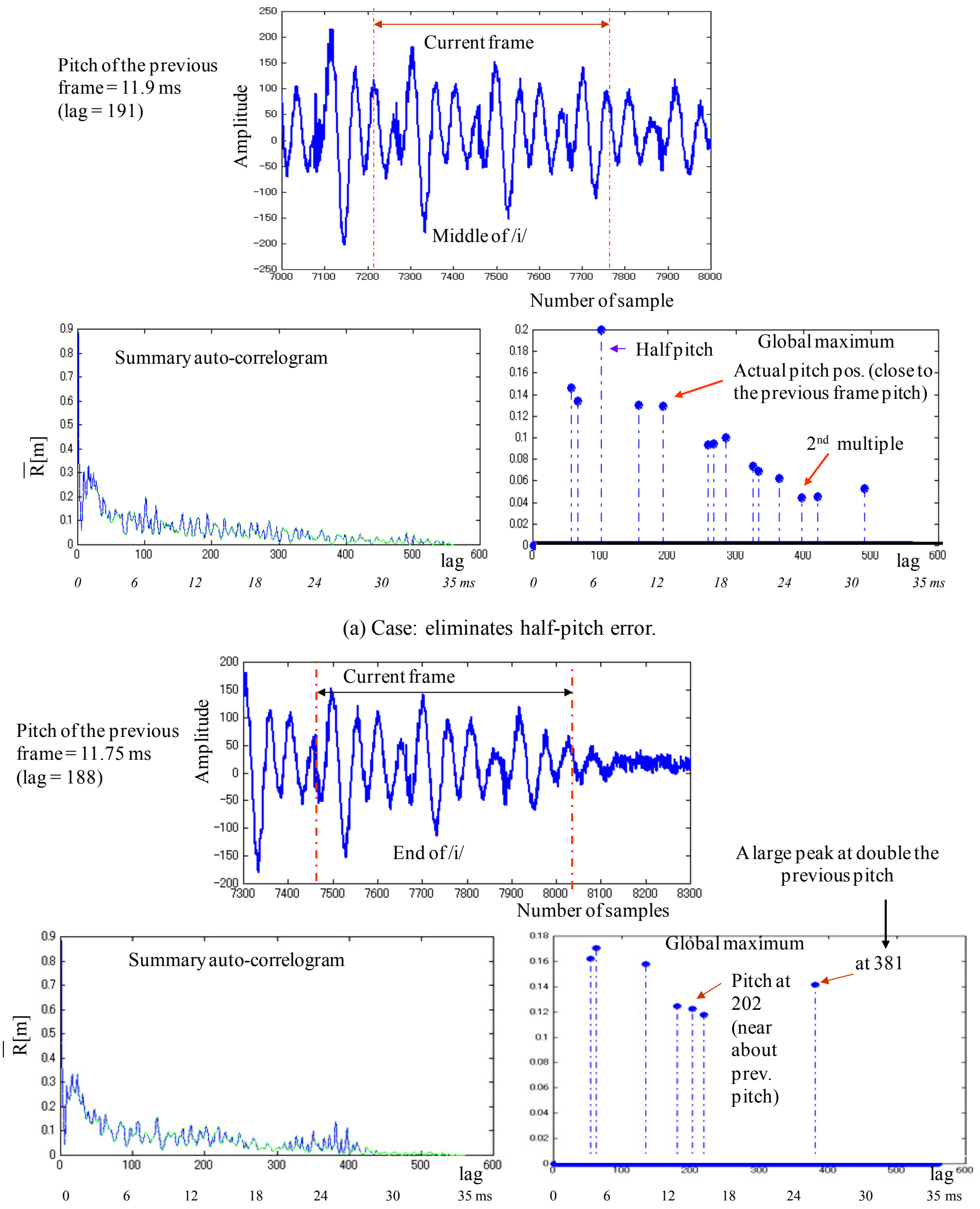


Fig. 5. Illustration of the basic steps of the proposed pitch detection algorithm. The label of horizontal axis corresponds both to lag number (0 100 200 300 400 500) and time (in ms).



(a) Case: eliminates half-pitch error.

(b) Case: eliminates 'no pitch' error.

Fig. 6. Examples for the enhancements of pitch detection. (a) The pitch candidate, located at lag 100, does not have G_{max} near its 3rd multiple lag 300, and hence its merit is decreased. On the other hand, the pitch candidate at lag 196 has a peak near its 2nd multiple lag 400, and 3rd multiple lag is out of frame. So its merit is increased. Also, pitch candidate at lag 196 locates closely to the pitch of previous frame (lag 191). The enhancement eliminates half-pitch error and detects the correct pitch. (b) A large G_{max} is located around double the previous pitch lag, and it helps to detect the current pitch at lag 202.

Table 1. Performance of different methods of pitch detection for male speech. Gross error, V to U/S (voiced to unvoiced/silent error), and U/S to V (unvoiced/silent to voiced error) are given in %.

Error (%)	Method	SNR (dB)			
		Clean	10	5	0
Gross error	(1)	0.13	0.78	2.08	2.65
	(2)	0.61	1.31	3.33	6.41
	(3)	0.72	1.68	4.21	8.34
	(4)	0.70	1.51	4.14	8.11
V to U/S	(1)	0.11	0.53	1.79	2.78
	(2)	0.52	1.25	3.08	5.97
	(3)	0.69	1.51	3.92	6.96
	(4)	0.63	1.41	3.73	6.65
U/S to V	(1)	0.51	1.15	1.60	2.37
	(2)	1.02	2.06	3.32	5.88
	(3)	1.06	2.42	4.03	6.96
	(4)	1.06	2.40	3.93	6.87

(1) The proposed PDA; (2) without blocks (A), (B);
(3) without blocks (A), (B), and ‘decrease merit’ at step iii. in Fig. 4 ;
(4) RAPT.

Table 2. Performance of different methods of pitch detection for female speech. Gross error, V to U/S (voiced to unvoiced/silent error), and U/S to V (unvoiced/silent to voiced error) are given in %.

Error (%)	Method	SNR (dB)			
		Clean	10	5	0
Gross error	(1)	0.09	0.65	2.01	2.54
	(2)	0.49	1.17	3.22	6.30
	(3)	0.60	1.59	4.13	8.11
	(4)	0.57	1.48	4.09	7.95
V to U/S	(1)	0.07	0.47	1.68	2.70
	(2)	0.41	1.13	2.94	5.81
	(3)	0.52	1.42	3.62	6.64
	(4)	0.49	1.37	3.51	6.45
U/S to V	(1)	0.42	0.97	1.43	2.01
	(2)	0.83	1.91	3.11	5.32
	(3)	0.95	2.13	3.32	6.11
	(4)	0.91	2.02	3.16	5.84

(1) The proposed PDA; (2) without blocks (A), (B);
(3) without blocks (A), (B), and ‘decrease merit’ at step iii. in Fig. 4 ;
(4) RAPT.

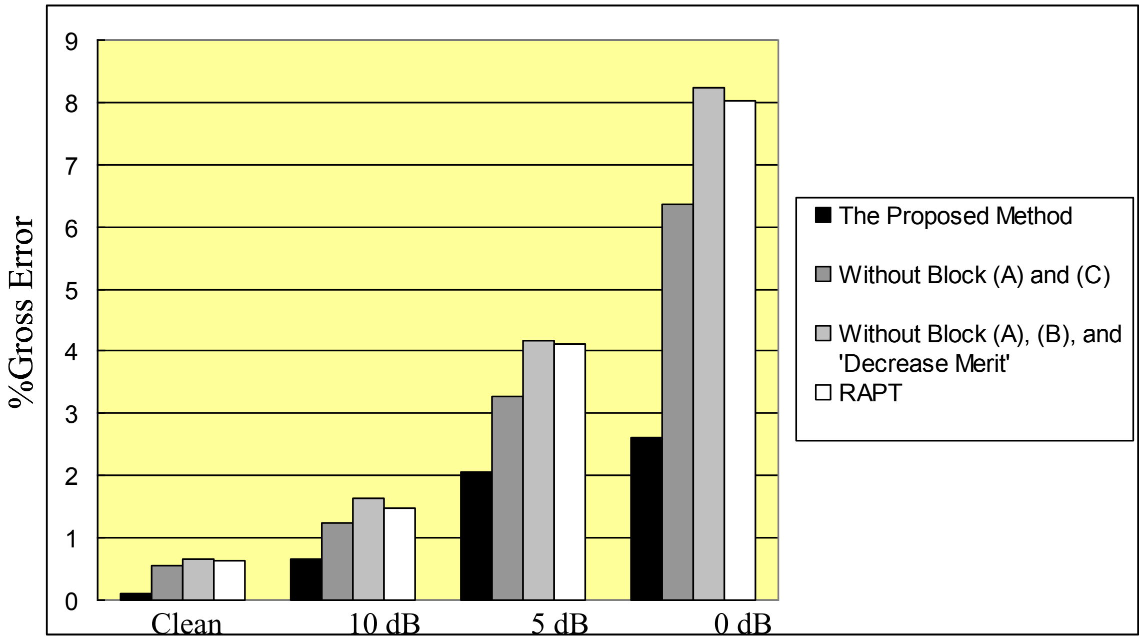


Fig. 7. Average %Gross Error of male and female speech for different methods.

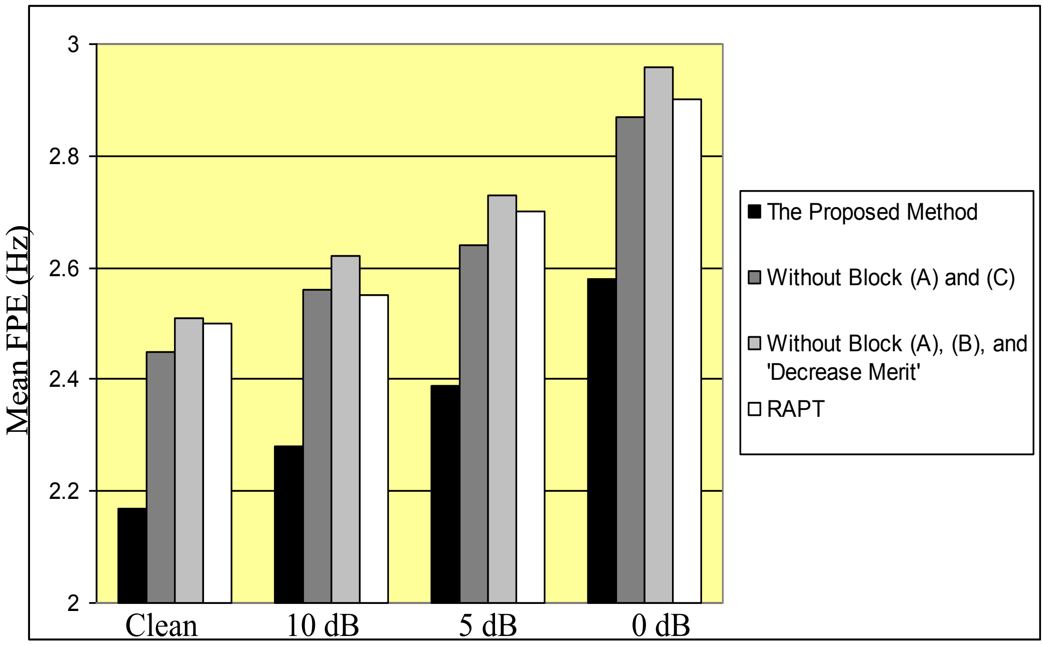


Fig. 8. Mean FPE (fine pitch error) of different methods for male speech.

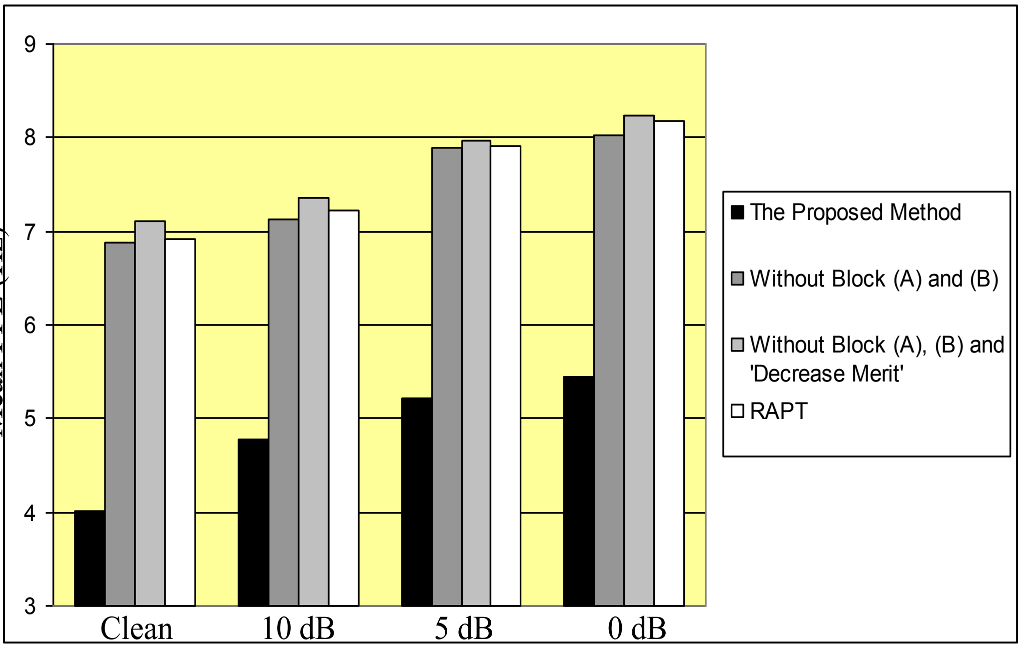


Fig. 9. Mean FPE (fine pitch error) of different methods for female speech.

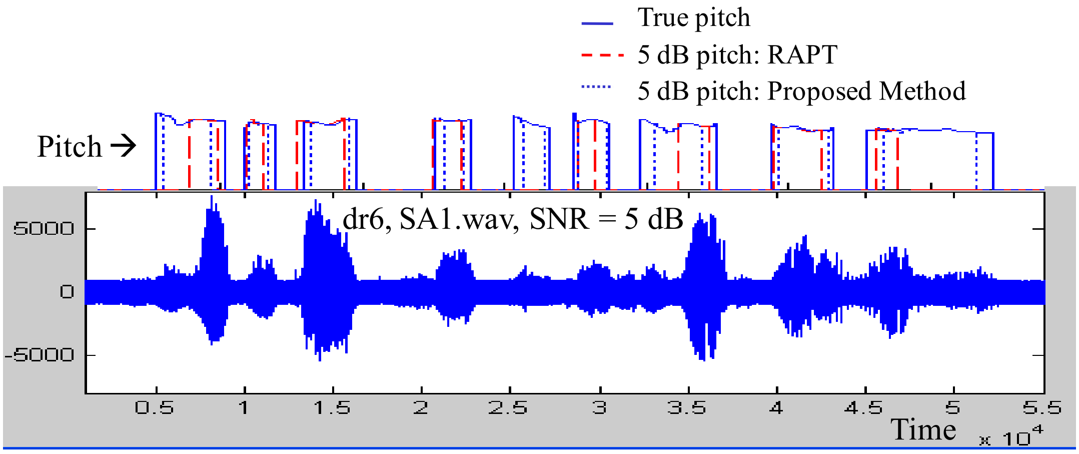


Fig. 10. Comparison of extracted pitch between RAPT and the proposed method. The utterance is contaminated with white noise at SNR = 5 dB.

References

[1] Gu L. and Liu R., "The government standard linear predictive coding algorithm", Speech Technology, pp. 40-49, 1982.

[2] Tamura M., Masuko T., Tokuda K., and Kobayashi T., "Adaptation of pitch and spectrum for HMM-based speech synthesis using MLLR", Proc. IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP), pp. 805-808, May 2001.

[3] Razak A. A., Abidin M.I.Z., and Komiya R., "Emotion pitch variation analysis in Malay and English voice samples", Proc. The 9th Asia-Pacific Conference on Communications (APCC),
Vol. 1, pp. 108 – 112, September 2003.

[4] Bagshaw P.C., Miller S.M., and Jack M.A., "Enhanced pitch tracking and the processing of the F0 contours for computer aided intonation teaching", Proc. Eurospeech'93, pp. 1003-1006, 1993.

[5] O'Shaughnessy D., "Speech communications: human and machine", IEEE Press, NY, second edition, 2000.

[6] Rabiner L.R., Cheng M.J., Rosenberg A.E., and McGonegal C.A., "A comparative performance study of several pitch detection algorithm", IEEE Trans. Acoustic, Speech, Signal Processing (ASSP), vol. 24, no. 5, pp. 399-417, 1976.

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

[8] Ross M.J., Shaffer H.L., Cohen A., Freudberg R., and Manley H.J., "Average magnitude difference function pitch extractor", IEEE Trans. ASSP, vol. 22, pp. 353-362, 1974.

[9] Abe T., Kobayashi T., and Imai S., "Robust pitch estimation with harmonics enhancement in noisy environment based on instantaneous frequency", Proc. International Conference on Spoken Language Processing (ICSLP)'96, vol. 2, pp. 1277-1280, 1996.

[10] Fangming W. and Yip P., "Cepstrum analysis using discrete trigonometric transforms", IEEE Trans. ASSP, vol. 39, no. 2, pp. 538-541, 1991.

[11] Huang H., and Pan J., "Speech pitch determination based on Huang-Hilbert transform", Signal Processing, vol. 86, no. 4, pp. 792-803, 2005.

[12] Hui L., Dai B., and Wei L., "A pitch detection algorithm based on AMDF and ACF", Proc. IEEE ICASSP'06, vol. 1, pp. 377-380, 2006.

[13] Talkin D., "A robust algorithm for pitch tracking", Speech Coding and Synthesis, Elsevier Science, pp. 495-518, 1995.

[14] Kasi K. and Zahorian S.A., "Yet another algorithm for pitch tracking", Proc. IEEE ICASSP'02, pp. 2294-2297, 2002.

[15] Zhang W., Xu G., and Wang Y., "Pitch estimation based on circular AMDF", Proc. IEEE ICASSP'02, vol. 1, pp. 341-344, 2002.

[16] Tabrikian J., Dubnov S., and Dickalov Y., "Speech enhancement by harmonic modeling via MAP pitch tracking", Proc. IEEE ICASSP'02, vol. 1, pp. 549-552, 2002.

[17] Wang C. and Seneff S., "Robust pitch tracking for prosodic modeling in telephone speech", Proc. IEEE ICASSP'00, pp. 1343-1346, 2000.

[18] Hermes D.J., "Measurement of pitch by sub-harmonic summation", J. Acoustic Soc. America, vol. 83(1), pp. 257-264, 1988.

[19] TIMIT Acoustic-Phonetic Continuous Speech Corpus, LDC Catalog No.: LDC93S1: *available at: http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC93S1*

[20] Ghulam M., Fukuda T., Horikawa J., and Nitta T., "A noise-robust feature extraction method based on pitch-synchronous ZCPA for ASR", Proc Interspeech'04, pp. 133-136, 2004

نظام محسن و ذاتي العلاقة للكشف عن تردد الرقيقتين الصوتيتين بدون التأثر بالضوضاء

غلام محمد

قسم هندسة الاسب، كلية علوم الحاسب والمعلومات

جامعة الملك سعود، ص.ب: ٥١١٧٨، الرياض ١١٥٤٣، المملكة العربية السعودية

Ghulam@ksu.edu.sa

(قدم للنشر في ٤/٥/٢٠٠٨م؛ وقبل للنشر في ٠٦/٠٥/٢٠٠٩م)

ملخص البحث. نقترح في هذه الورقة البحثية خوارزمية لإيجاد النغمة بحيث تكون فعالة و مقاومة للضوضاء. هذه الخوارزمية تستند على دالة التطابق الزمني(ACF). وقد أستعملت مجموعة مصفيات تنافسية لإنتقاء النغمة المرشحة. و تم تطبيق معيار ترجيح، يشمل كلا من الزيادة والنقصان في الجدارة، على النغمة المرشحة من أجل اكتشاف وجود أوعدم وجود توافقيات النغمة. وأخيرا أدمج تعزيز بسيط لتلطيف محيط النغمة. تم تقييم الخوارزمية المقترحة لاكتشاف النغمة و نشاط الصوت في قاعدة البيانات TIMIT مع أنواع ومستويات مختلفة من الضوضاء. وتظهر النتائج التجريبية تفوق الطريقة المقترحة على أساليب أخرى معروفة.