A High Resolution Pitch Detection Algorithm Based on AMDF and ACF

K. Abdullah-Al-Mamun[1], F. Sarker[2], and G. Muhammad[3]

[1]Institute of Sound and Vibration Research (ISVR), University of Southampton, Southampton, UK

[2]School of Electronics and Computer Science (ECS), University of Southampton, Southampton, UK

[3]College of Computer and Information Sciences, King Saud University, Riyadh, Saudi Arabia

Abstract

This paper proposes a noise robust high resolution pitch detection algorithm based on AMDF and ACF. The falling trend of AMDF is eliminated by an alignment technique, and AMDF and ACF are combined to take the advantage of their complementary nature. These two functions are combined by multiplication and addition over several band pass filters to enhance important candidates and suppress less important candidates. Then using a sophisticated weight assignment procedure, the candidate with the highest weight is selected as pitch. The proposed method is evaluated on different colored noisy speech at different intensity. Experimental result shows noise robustness of the proposed method in varied environments.

1. Introduction

The problem of pitch estimation is an area of research during all the evolution of digital signal processing. Pitch determination has numerous applications in speech processing. Accurate pitch extraction has been demonstrated to play a very important role in speech coding, speech compression, speech synthesis, speech recognition and speaker identification, as well as in musical world. A good estimation of pitch period is crucial to improve the performance of speech analysis and synthesis systems. There are many pitch detection algorithms such as the short-time average magnitude difference function (AMDF) [1], short-term autocorrelation function (ACF) [2], direct time domain fundamental frequency estimation (DFE) [3], weighted autocorrelation (WAC) [4], and zero-cross rate with autocorrelation [5] algorithms. Although many pitch detection algorithms have been discovered, few of them have been built in special purpose digital hardware able to work on noisy environment and real time [6].

---

[1] *Corresponding author*: km@isvr.soton.ac.uk

In general, the short-time AMDF and short-term ACF are often employed in pitch detection. AMDF based methods exhibit less computational complexity while ACF [2, 7] based methods perform better in case of noisy speech. The AMDF based methods are thus widely used in real-time systems because of its time efficiency. Using the AMDF based method; however, two types of estimation errors often happen. One is that the estimated pitch period is multiple of the actual, while the other is that the actual value is multiple of the estimated. We refer the first one as double pitch error, and the other one as half pitch error. The reason for these errors is mainly the complication of speech waveforms. Besides this, these errors occur mainly due to the falling trend of the AMDF peaks at higher lags. For the noisy speech signals, this tendency increases the occurrence of octave errors in a greater degree. Number of improvements developed based on basic AMDF method such as high resolution AMDF (HRAMDF) [8] and circular AMDF (CAMDF) [9] to overcome the errors. Both the methods are successful to eliminate the falling trend in most cases but causes the magnitude at pitch multiples or factors (i.e. dips) to be emphasized, which triggers new octave errors. Alignment AMDF [10] technique effectively eliminates the falling trend by aligning the AMDF peaks along a straight line. But the real challenge still exists to detect the correct pitch from highly corrupted speech signal.

In this paper, we propose an effective pitch estimation method based on AMDF and ACF for speech severely corrupted by noise with reliability and accuracy as the prime focus. In our proposed method, we eliminate the falling trends by modification to the original AMDF by adding threshold with two third of each frame, that aligns the peaks along a straight line. We combined the result of mirror AMDF and ACF and apply post processing that provides better estimation of pitch with Voice Activity Detection (VAD) for noisy digital speech signal.

The paper is organized as follows. First, a review of the AMDF and ACF pitch detection algorithm are given in Section 2. Section 3 describes the proposed pitch detection algorithm based on AMDF and ACF. Section 4 presents the experimental frame work. Finally the experimental result and discussions are presented in section 5. The paper is concluded in section 6.

2. Review of AMDF and ACF Pitch Detection Algorithm

2.1. *AMDF pitch detection algorithm*

The short-time AMDF is defined as

$$D_{AMDF}(m) = \frac{1}{N} \sum_{n=0}^{N-1} | x(n) - x(n+m) | \tag{1}$$

where $x(n)$ are the samples of speech. For a periodic signal with period T_0, this function is expected to have a strong minimum when the lag index m equals T_0.

The pitch period is, in general, estimated as follows:

$$T_0 = MIN(D_{AMDF}(m)), \text{ for } m = m_{min} \text{ to } m_{max} \qquad (2)$$

where the values of m_{min} and m_{max} are chosen to cover the expected pitch-range.

2.2. *ACF pitch detection algorithm*

A commonly used method to estimate pitch is based on detecting the highest value of the autocorrelation function in the region of interest. Given a discrete time signal *x(n)*, defined for all *n*, the auto-correlation function is generally defined in (3):

$$D_{ACF}(m) = \frac{1}{N} \sum_{n=0}^{N-1} x(n) \times x(n+m) \qquad (3)$$

The autocorrelation function of a signal is basically a (noninvertible) transformation of the signal that is useful for displaying structure in the waveform. Thus, for pitch detection, if we assume *x(n)* is exactly periodic with period *P*,
i.e., *x(n) = x(n + P)* for all *n*, then it is easily shown that:

$$R_x(m) = R_x(m+P) \qquad (4)$$

i.e., the autocorrelation is also periodic with the same period. Conversely, periodicity in the autocorrelation function indicates periodicity in the signal.

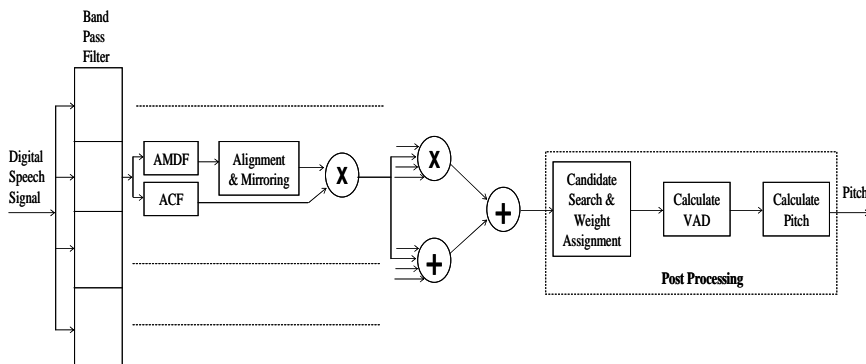## 3. Proposed Pitch Detection Algorithm Based on AMDF and ACF

Fig. 1. Block diagram of the proposed high resolution pitch detection based on AMDF and ACF.

Fig. 1 shows the block diagram of the proposed high resolution pitch detection based on AMDF and ACF and its working procedure are as follows.
(i) *Band pass filtering*: Speech signal is passed through four band pass filters (50 – 200 Hz, 150 – 300 Hz, 250 – 400 Hz, and 350 – 500 Hz). In the proposed method we apply AMDF and ACF to the output of each filter.

(ii) *Avoiding falling tendency of AMDF*: We avoid the falling trends of the original AMDF by adding threshold from two third to last sample of each frame to align the peaks along a straight line. The threshold is a cumulative addition of 0.01 to each sample in the last one-third length of a frame. We also trialed this threshold using different values, however, 0.01 gave the best result. This alignment operation leaves the relative magnitude (i.e. relative height of the peaks and dips) of the AMDF unchanged and conquers the falling trend at the same time.

(iii) *Combining AMDF and ACF*: AMDF provides notch output while ACF provides peak. Mirroring is applied to AMDF to covert notches into peak to make it similar as ACF output. In some cases, AMDF can estimate pitch near to the actual pitch while some other cases it cannot. ACF also provides complementary information of AMDF. We combine the result of mirrored AMDF and ACF by multiplication and get four multiplication results for each filter. This multiplication is applied to reduce the number of unwanted candidates. These results are combined for all filters through addition and multiplication and thereby getting two outputs. Finally, by adding these two results we get the final candidates that provide a better pitch estimation for noisy speech signal. We also applied multiplication instead of addition at the final combination but it did not provide good result. The multiplication at this stage filters out many potential candidates resulting in pitch error.

(iv) *Post processing*: Post processing is applied on the final candidates. Post processing consists of candidate search and weight assignment, VAD and pitch calculation. Candidate search and weight assignment are performed at each frame while the VAD and pitch calculation are performed at whole utterance. The post processing steps are shown in Fig. 2.

(iv.a) At first, candidate selection is performed by searching the peak (i.e. local maximum) $m_0, m_1, m_2, \ldots, m_{k-1}$ from each frame and weight is assigned to each peak in a frame based on the periodicity.

(iv.b) Candidate peaks ($m_0, m_1, m_2, \ldots, m_{k-1}$) are sorted based on maximum weight. A candidate who has the maximum weight but peak amplitude is less than a threshold is not selected as pitch. The threshold is set as 2/3 of average of the highest and the lowest peak amplitude.

(iv.c) VAD is calculated as follows: the average amplitude of selected pitch is calculated over the utterance. If the amplitude for a certain frame is greater than 33% from the average (i.e. 1.3 x Avg) then frame is considered as voiced.

(iv.d) VAD and Pitch smoothing: For VAD, if previous frame and next frame is voiced then the current frame is also considered as voiced. Again if previous frame and next frame is unvoiced then the current frame is considered as unvoiced. For the unvoiced frame pitch is set to zero.

Fig. 3 shows graphically the procedure of the proposed method. Speech signal is a female voiced frame contaminated with subway noise at SNR = 0 dB. Fig. 3 (c) shows alignment of AMDF. This alignment reduces the possibility of finding minimum at multiple positions of true pitch by eliminating the falling tendency of the original AMDF.

From Fig. 3 (f), (g) and (h), we can see the effect of using several band pass filters. Addition and multiplication over the channels reduce the height of peaks at multiple-pitch positions, and enhance the peak at actual pitch position. The result of candidate selection is shown in Fig. 3 (i), while the actual pitch, after selecting the highest weighted candidate is in Fig. 3 (j).

Post Processing	

Frame by frame	**1. Candidate selection:**
	    1.1 Search the peaks (i.e. local maximum) m_0, m_1, m_2, …, $m_{(k-1)}$ from the each frame.
	    1.2 Assign weight to each peaks in a frame based on periodicity.
	2. Best Candidate for Pitch:
	    2.1 Sort the peaks (m_0, m_1, m_2, …, $m_{(k-1)}$)based on maximum weight.
	    2.2 Calculate best peaks for pitch based on weight and pick amplitude.

Over the utterance	**3. VAD calculation:**
	    3.1 Calculate voiced and unvoiced frame over the utterance.
	4. VAD and Pitch smoothing:
	    4.1 Smoothing VAD over the utterance.
	    4.2 Smoothing Pitch over the utterance.

Fig. 2. Post processing steps for the proposed high resolution pitch detection based On AMDF and ACF.
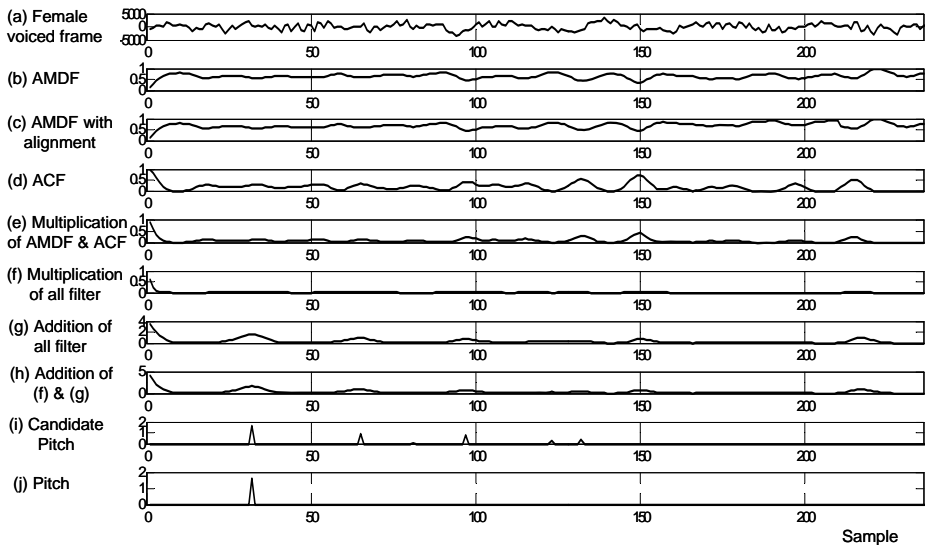

Fig. 3. Illustration of the proposed pitch detection based on AMDF and ACF. (a) is a female voiced frame at SNR = 0 dB; (c) shows alignment of (b) AMDF; (d) ACF, (e) multiplication of ACF and AMDF. (b) – (e) are the results from filter 4. (f) and (g) show multiplication and addition, respectively, over all the filters. (h) gives addition result of (f) and (g). Selected pitch candidates are shown in (i), while (j) shows the final detected pitch.

4. Experimental Framework

4.1. *Database*

Aurora-2J database [11] is used in the experiments. The utterances are connected Japanese digit strings and sampling rate is 8 kHz. Selections of 8 different real-world noises have been added to the speech over a range of signal-to-noise ratios (SNRs: -5 dB, 0 dB, 5 dB, 10 dB, 15 dB, 20 dB, clean). Eight different real noises are divided into two groups for testing. Data in Test A are added to by noises of Subway, Babble, Car, and Exhibition. Data in Test B are added to by noises of Restaurant, Street, Airport, and Station. In Test C, besides the additive noise, channel distortion is also included. In our experiments, we use 20 different utterances consisting of 10 males and 10 females with different SNRs (from 0 dB to 20 dB) from Data in Test A and Test B.

4.2. *Experimental setup*

The KTH's Wavesurfer implementation of a noise robust algorithm for pitch tracking [12] is used as baseline in the proposed improvement. This algorithm is based on normalized cross-correlation and dynamic programming. Reference pitch is extracted by applying Wavesurfer on corresponding clean speech (available in Aurora-2J). In our experiment, rectangular window of 32 ms is applied to the input speech at 10 ms frame rate.

5. Results and Discussions

Results are given in percentage gross pitch error (%GPE). If any estimated pitch is not within 1 ms of the reference pitch, then it is termed as gross error. %GPE is provided for both male and female speech.

Tables 1 and 2 show %GPE of baseline wavesurfer and the proposed pitch detection method for both female and male speech, respectively, at SNR = 20 dB, 15 dB, 10 dB, 5 dB and 0 dB. From the tables we can see that the proposed method outperforms Wavesurfer in terms %GPE for both male and female speech. For example, in SNR = 0 dB, the proposed improves %GPE from 43.75%, obtained by the wavesurfer, to 25.22% for female speech, and from 40.74% to 22.22% for male speech.

Table 1. Performance comparison of the proposed method and wavesurfer in terms of global pitch error (%GPE) for female speech.

SNR	20 dB	15 dB	10 db	5 dB	0 dB
Proposed method	6.25	12.51	16.22	18.75	25.22
Wavesurfer	12.50	25.32	31.25	38.01	43.75

Table 2. Performance comparison of the proposed method and wavesurfer
in terms of global pitch error (%GPE) for male speech.

SNR	20 dB	15 dB	10 db	5 dB	0 dB
Proposed method	7.40	12.01	14.81	18.51	22.22
Wavesurfer	11.11	14.81	18.51	25.92	40.74

Figs. 4 and 5 show pitch values as well as VAD result (voiced when pitch is other than zero) for an utterance /roku/ by female and male, respectively, at SNR = 0 dB (subway noise). The proposed method produces better VAD result (close to true VAD) while the Wavesurfer reduces voiced parts from both sides.
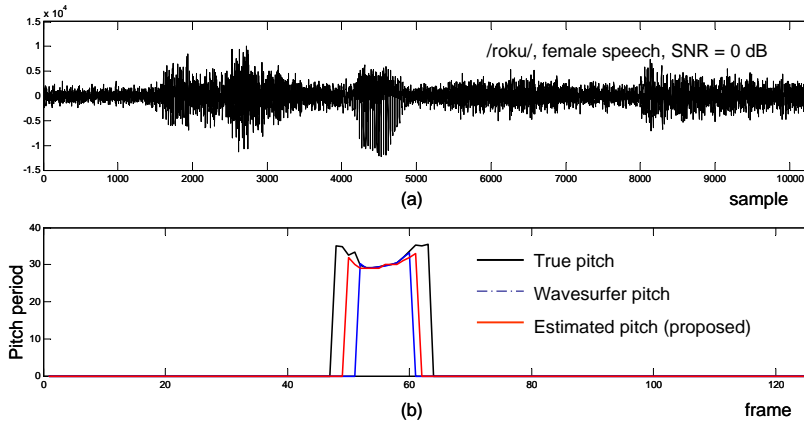

Fig. 4. Pitch and VAD comparison between the methods for a female speech with SNR = 0 dB. The proposed method provide better than the wavesurfer.
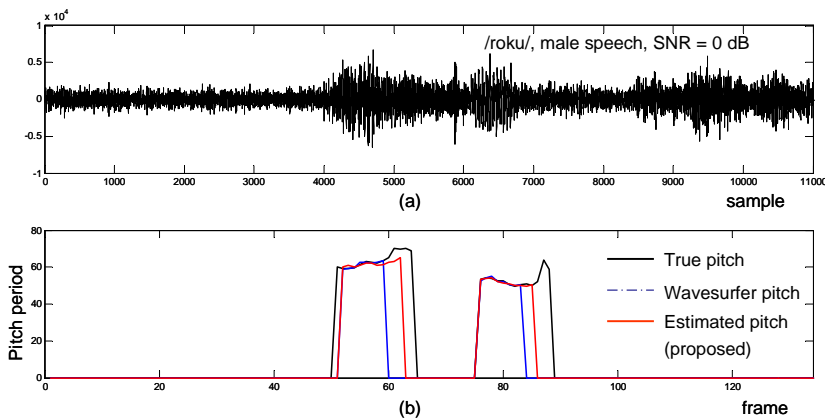

Fig. 5. Pitch and VAD comparison between the methods for a male speech with SNR = 0 dB. The proposed method provide better than the wavesurfer.

From these results, we conclude that the proposed AMDF and ACF based pitch detection performs well even in very noisy condition and at different environmental noises. It may be noted that some of the data consists of babble noise, which is a major source of pitch error.

6. Conclusions

A noise robust pitch detection algorithm based on AMDF and ACF was proposed. Complementary nature of AMDF and ACF was combined by multiplication and addition, and a weight assignment procedure was applied to extract correct pitch. The proposed method showed better performance compared to other method in both male and female speech corrupted with different colored noise. The proposed method will be evaluated on a larger dataset in future.

References

1. M. J. Ross, H. L. Shaffer, A. Cohen, R. Freudberg, and H. J. Manley, IEEE Trans. on Acoustics, Speech, Signal Processing **22**, 353 (1974). doi:10.1109/TASSP.1974.1162598
2. X-D. Mei, J. Pan and S-H. Sun, Efficient algorithms for speech pitch estimation, *Proc. Int. Symp. on Intelligent Multimedia, Video and Speech Processing*, (Hong Kong, 2001) pp. 421-424.
3. H. Boril and P. Pollak, Direct Time Domain Fundamental Frequency Estimation of Speech in Noisy Conditions, *Proc. European Signal Processing Conference*, vol. 1 (Vienna, Austria, 2004) pp. 1003-1006.
4. T. Shimamura and H. Kobayashi, IEEE Trans. on Speech and Audio Processing, **9** (7), 727 (2001). doi:10.1109/89.952490
5. R. G. Amado and J. V. Filho, Pitch detection algorithms based on zero-cross rate and autocorrelation function for musical notes, *Proc. Int. Conf. on Audio, Language and Image Processing*, (Shanghai, China, 2008) pp. 449-454.
6. T. Nakatani and T. Irino, J. Acoustic Society America **116** (6), 3690 (2004). doi:10.1121/1.1787522
7. G. Muhammad, Noise robust pitch detection based on extended AMDF, *Proc. 8th IEEE Int. Symp. on Signal Proessing and Information Technology*, (Sarajevo, Bosnia & Herzegovnia, 2008) pp. 133-138.
8. L. Gu and R. Liu, The Government Standard Linear Predictive Coding Algorithm, Speech Technology Magazine (1982) pp. 40-49.
9. W. Zhang, G. Xu and Y. Wang, Pitch estimation based on circular AMDF, Proc. of Int. Conf. on Acoustics, Speech, and Signal Processing (Florida, USA, 2002) pp. 341-344.
10. M. S. Rahman, H. Tanaka and T. Shimamura, Pitch determination using aligned AMDF, *Proc. INTERSPEECH 2006* (Pittsburgh, USA, 2006) pp. 1714-1717.
11. S. Nakamura, K. Yamamoto, K. Takeda, S. Kuroiwa, N. Kitaoka, T. Yamada, M. Mizumachi, T. Nishiura, M. Fujimoto, A. Sasou and T. Endo, Data Collection and Evaluation of AURORA-2 Japanese Corpus, *IEEE Workshop on Automatic Speech Recognition and Understanding* (2003) pp. 619-623.
12. D. Talkin, A robust algorithm for pitch tracking (RAPT), Speech Coding and Synthesis, Elsevier Science, (1995) pp. 495-518.