Pitch Determination Using Aligned AMDF

M. Shahidur Rahman, Hirobumi Tanaka, and Tetsuya Shimamura

Department of Information and Computer Sciences
Saitama University, Saitama 338-8570, Japan
rahmanms@sie.ics.saitama-u.ac.jp

Abstract

A pitch determination method based on AMDF (Average Magnitude Difference Function) is proposed in this paper. The AMDF is often used to determine the pitch parameter in real-time speech processing applications. Falling trend of AMDF at higher lags, however, makes the method vulnerable to octave errors (pitch doubling or halving). In this paper, we propose an alignment technique that effectively eliminates the falling trend by aligning the AMDF peaks along a straight line. Experimental results on speech signals spoken by male and female speakers show that the current method can reduce the occurrence of octave errors in greater numbers when compared with other AMDF based functions.

Index Terms: speech processing, pitch determination, AMDF method, autocorrelation method.

1. Introduction

Pitch determination has manifold applications in speech processing. Accurate pitch estimation has been demonstrated to play an important role in speech compression, speech synthesis, speech/ speaker recognition and in musical world as well. Numerous pitch determination algorithms have been reported in the literature [1, 2, 3, 4, 5, 6, 7].

Autocorrelation based methods [2], in general, exhibit better performance in case of noisy speech while average magnitude difference function (AMDF) [1] based methods are more time-efficient. The AMDF based methods are thus widely employed in real-time systems. Octave errors (both lower and upper) are, however, common in using AMDF based methods. These errors occur mainly due to the falling trend of the AMDF peaks at higher lags. For speech signals corrupted with noise, this tendency increases the occurrence of octave errors in a greater degree. The basic AMDF method is thus followed by a number of improvements such as high resolution AMDF (HRAMDF) [8] and circular AMDF (CAMDF) [9]. Both the methods are successful to eliminate the falling trend in most cases but causes the magnitude at pitch multiples/ factors (i.e. dips) to be emphasized, which triggers new octave errors. In this paper we propose a modification to the original AMDF that aligns the peaks along a straight line. The alignment operation leaves the relative magnitude (i.e. relative height of the peaks and dips) of the AMDF unchanged and conquer the falling trend at the same time. The experimental results show that the aligned AMDF (AAMDF) outperforms both HRAMDF and CAMDF and reduces the octave errors to a minimum.

The AAMDF can also be effectively used in combination with the autocorrelation function (ACF) as described in [6], where it is shown that the characteristics of ACF and the inversed AMDF are similar but the noise components of ACF and AMDF are uncorrelated and behave independently. In noisy environments, these two properties lead to an emphasis of the pitch candidate when the ACF is combined with the inversed AMDF. Experimental results show that the AAMDF results in a further improvement in pitch detection when its reciprocal is combined with the ACF.

2. Description of AMDF and Existing Improvements

The short-time AMDF is defined as [1]

$$D(\tau) = \frac{1}{N}\sum_{n=0}^{N-1} \mid x(n) - x(n+\tau) \mid \qquad (1)$$

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 τ equals T_0. The pitch period is, in general, estimated as follows:

$$T_0 = MIN(D(\tau)), for\ \tau = \tau_{min}\ to\ \tau_{max} \qquad (2)$$

where the values of τ_{min} and τ_{max} are chosen to cover the expected pitch-range. In this study, we limit the range for detecting fundamental frequencies from 60 Hz to 400 Hz, which is adequate for typical speech processing.

In (1), the samples past the length N is zero and at higher lags less data is involved in the computation of the function D which causes it to fall off. Furthermore, D is sensitive to noise and intensity variations which influence directly the magnitude of the principal minimum as well. The AMDF obtained from a typical speech waveform is shown in Fig. 1. An estimate of the pitch period is obtained from the location of principal minimum.

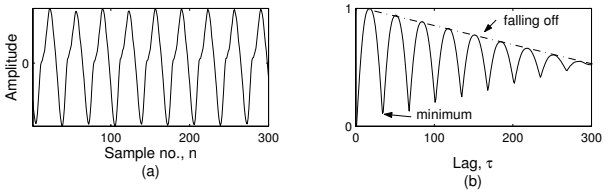

Figure 1: *a) Speech waveform; b) obtained AMDF.*

Due to the falling trend of the AMDF peaks as apparent in Fig. 1(b), magnitude at pitch multiples (e.g. $2T_0$) or at pitch factors (e.g. $T_0/2$) approaches the magnitude at true pitch location. Depending on the nature of speech segment, magnitude at pitch

multiples or at pitch factors can be more evident than that at the true pitch location, thus causing pitch doubling or halving. An example of pitch period doubling is shown in Fig. 2, where the minimum magnitude is observed at $2T_0$. To overcome this falling ten-

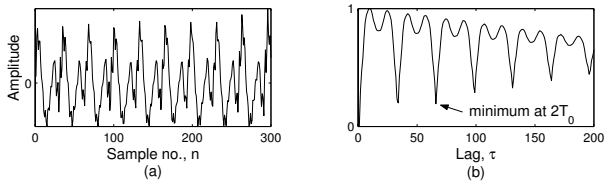

Figure 2: *Pitch period doubling using AMDF: a) Speech waveform; b) obtained AMDF.*

dency HRAMDF is employed in speech coding standard LPC.10 [8] which is described as

$$D_H(\tau) = \sum_{n=(N/2-\tau)/2+1}^{(N/2-\tau)/2+N/2} | \, x(n) - x(n+\tau) \, | \, . \qquad (3)$$

Two different speech segments are involved in the computation of D_H. Unlike (1), every lag in (3) is well averaged and the falling trend is almost alleviated. This avoids some situations of octave errors. Unfortunately, the modification makes the magnitude at the locations of other pitch multiples/ factors stronger as well which in turn introduces some additional octave errors. Fig. 3 shows such a situation.

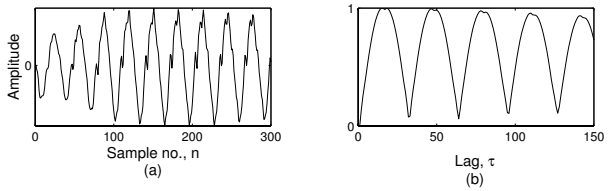

Figure 3: *Pitch period doubling using high resolution AMDF. a) Speech waveform; b) obtained HRAMDF.*

Recently, Zhang *et al*. described CAMDF in [9] which is defined as

$$D_C(\tau) = \sum_{n=0}^{N-1} | \, x(mod(n+\tau, N)) - x(n) \, | \qquad (4)$$

where $mod(n+\tau, N)$ represents the modulo operation, meaning that $n+\tau$ modulo N. The function D_C is symmetrical around $\tau = N/2$, i.e. $D_C(\tau) = D_C(N-\tau)$. Pitches are, therefore, calculated only from $\tau \in [0, N/2]$, which indicates that a double sized segment is required for pitch determination. This function has similar drawbacks as in D_H. Particularly, when the segment is positioned pitch-synchronously (i.e. when the beginning is a natural continuation from the end) magnitudes at all the pitch multiples are emphasized that introduces octave errors. This is illustrated in Fig. 4.

The speech segment used in Fig. 4 is the same as that in Fig. 1. The magnitudes at the pitch multiples in Fig. 4(b) are, however, over emphasized with respect to the same in Fig. 1(b), which causes pitch period doubling.

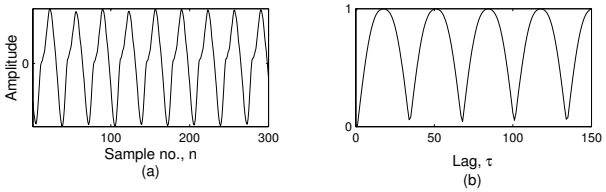

Figure 4: *Pitch period doubling using circular AMDF. a) Speech waveform; b) obtained CAMDF.*

3. The Proposed Algorithm

The improvements described in the previous section conquer the falling trend of AMDF by the cost of enhancing the magnitude at pitch multiples. However, a necessary condition for pitch determination using AMDF is that magnitude at the successive pitch multiples be in increasing order (since pitch period is estimated from the location of minimum amplitude) as commonly observed in the AMDF as seen in Fig. 1. A natural solution to the problem is thus alleviating the falling trend which will retain the increasing trend in magnitude at the pitch multiples. Though the AMDF obtained from speech with long pitch period is apparently suitable for pitch extraction, it possesses serious drawback for speech with short pitch period. In this section we propose an alignment technique that aligns the AMDF sequence horizontally while keeping the relative magnitude unchanged. The procedure is illustrated in Fig. 5.

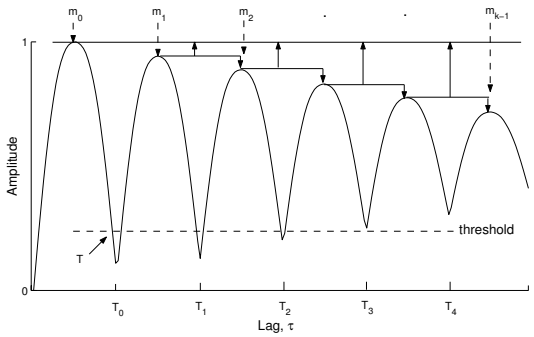

Figure 5: *The AMDF alignment process.*

The algorithm can be summarized as follows:
1) Calculate the AMDF, D, according to (1).
2) Search the peaks (i.e. local maxima), $m_0, m_1, m_2, .., m_{k-1}$ as seen in Fig. 5, within $[0, \tau_{max}]$.
3) Adjust the magnitude of samples from $D(m_1)$ to $D(m_2 - 1)$ by adding the difference $(D(m_0) - D(m_1))$, adjust the samples from $D(m_2)$ to $D(m_3 - 1)$ by adding $(D(m_0) - D(m_2))$, and so on.
4) Determine the pitch from the aligned AMDF using the principle of (2).

An assumption of the pitch period (as pointed as T in Fig. 5) is made by a simple threshold logic which can be computed, for example, as 80-90% of the minimum obtained at $\tau \in [21, \tau_{max}]$. A local maximum is determined at every interval T.

The AAMDF obtained from the AMDF of Figs. 1(b) and 2(b) are shown in Fig. 6(a) and 6(b), respectively. As evident in

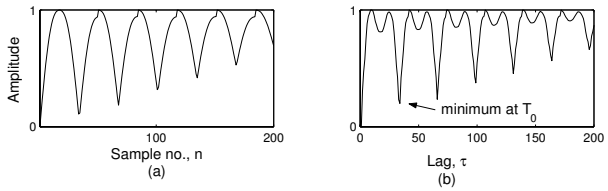

Figure 6: *Aligned AMDF. a) obtained from the AMDF in Fig.1(b); b) obtained from the AMDF in Fig. 2(b).*

Fig. 6(b), the alignment operation eliminates the pitch period doubling that occurred using AMDF in Fig. 2(b).

The detail experimental results are presented in Section 4.

4. Experimental Results

The performance of the proposed method is examined on natural speech spoken by a Japanese male and a female speaker. Speech materials are two 11 sec long sentences sampled at 10 kHz rate taken from the database developed by NTT [10].

The reference file is constructed by computing the fundamental frequencies every 10 ms using a semi-automatic method. Pitch estimation error is calculated as the difference between the reference and estimated fundamental frequency. Pitch estimation error obtained from speech spoken by a female speaker at SNR=∞, 10 dB, and 0 dB are shown in Figs. 7, 8, and 9, respectively. The same thing is repeated for a male speaker in Figs. 10, 11, and 12. Figures at the left, center, and right panels are obtained using AMDF, CAMDF, and AAMDF, respectively. Two error parameters GPE (Gross Pitch Error) and FPE (Fine Pitch Error) are commonly used as a measure of errors in estimating pitch period. The possible sources of GPE (usually greater than 10 samples) is pitch doubling, halving, inadequate suppression of formants as to affect the estimation, etc.. This is obvious in Figs. 7 through 12 that the number of GPE can be greatly reduced using the AAMDF method. In almost all the examples the proposed method performs better than the CAMDF method. The FPE (usually less than 10 samples), on the other hand, is attributed to measurement techniques. The current method is supposed to produce the same FPE as that of the AMDF method.

As mentioned earlier, weighted autocorrelation method in [6] has shown its effectiveness in noisy environments. Fig. 13 shows a further improvement of accuracy when ACF is weighted by the reciprocal of AAMDF instead of AMDF. The number of GPE in Fig. 13(b) is almost half the same in Fig. 13(a).

5. Conclusion

Accurate pitch estimation is a tough problem in speech analysis especially in high-pitched voices. A common problem is that the estimated pitch is one octave lower or upper than the actual pitch. The original AMDF method with lower computational complexity, however, fails to deal with these problems. The proposed method adheres some extra computations in return of significant reduction of octave errors even for severely corrupted noisy speech.

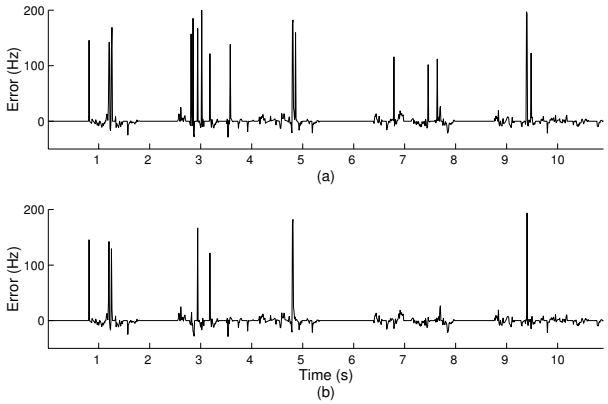

Figure 13: *Pitch estimation error using speech in Fig. 7 after being corrupted with white noise at SNR=0 dB. a) when ACF is weighted by inversed AMDF; b) when ACF is weighted by inversed AAMDF.*

6. References

[1] Ross M. J., Shaffer H. L., Cohen A., FreudBerg R., and Manley H. J., "Average magnitude difference function pitch extractor," IEEE Trans. Acoust, Speech, Signal Processing, vol.22, pp.353-362, 1974.

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

[3] W. J. Hess, Pitch Determination of Speech Signals. Berlin, Germany: Springer-Verlag, 1983.

[4] Krubsack D. A. and Niederjohn R. J., "An autocorrelation pitch detector and voicing decision with confidence measures developed for noise-corrupted speech," IEEE Trans. Acoust., Speech, Signal Processing, vol.39, pp.319-329, 1991.

[5] Liu D. and Lin C., "Fundamental frequency estimation based on the joint time-frequency analysis of harmonic spectral structure," IEEE Trans. Speech and Audio Processing, vol.9, no.6, pp.609-621, 2001.

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

[7] Nakatani T. and Irino T., "Robust and accurate fundamental frequency estimation based on dominant harmonic components," J. Acoustical Society of America, vol.116, no.6,pp.3690-3700, 2004.

[8] Gu L. and Liu R., "The Government Standard Linear Predictive Coding Algorithm," Speech Technology Magazine, pp.40-49, April 1982.

[9] Zhang W., Xu G., and Wang Y., "Pitch estimation based on circular AMDF," Proc. of ICASSP'2002, Florida, USA, pp.341-344, May 2002.

[10] Multilingual Speech Database for Telephometry, NTT Advance Technology Corp., Japan, 1994.

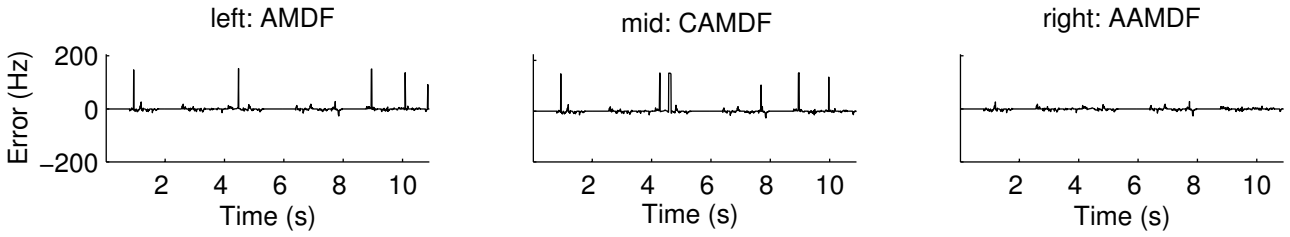

Figure 7: *Pitch estimation error using clean speech spoken by a female speaker.*

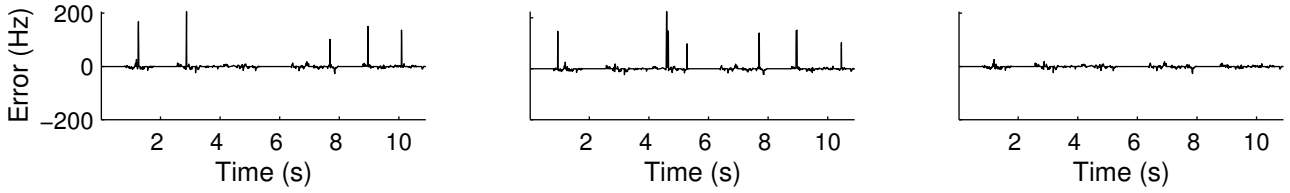

Figure 8: *Pitch estimation error obtained from speech used in Fig. 7 after being corrupted with white noise at SNR=10 dB.*

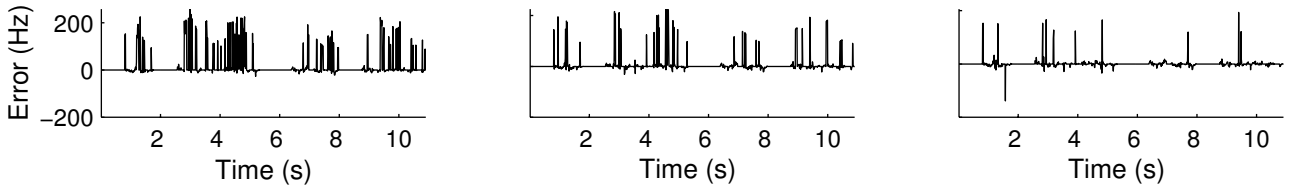

Figure 9: *Pitch estimation error obtained from speech used in Fig. 7 after being corrupted with white noise at SNR=0 dB.*

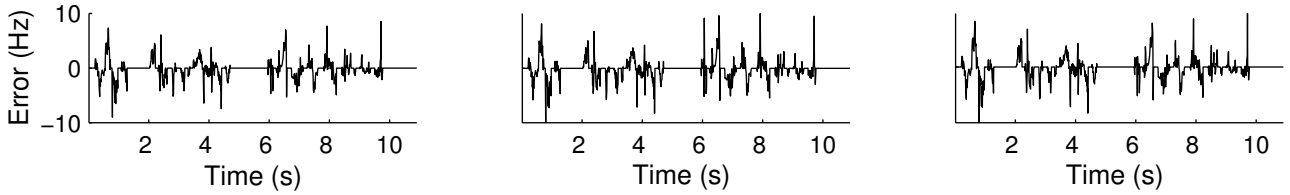

Figure 10: *Pitch estimation error using clean speech spoken by a male speaker.*

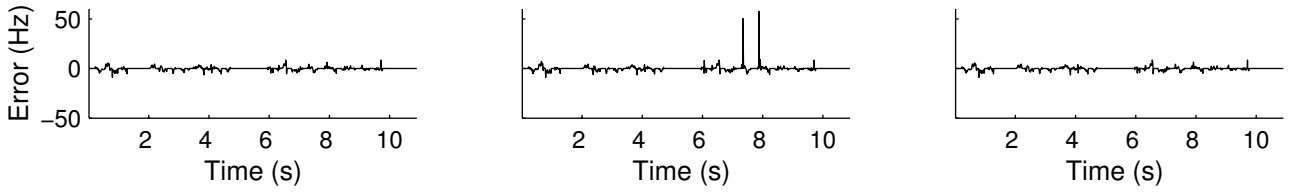

Figure 11: *Pitch estimation error obtained from speech used in Fig. 10 after being corrupted with white noise at SNR=10 dB.*

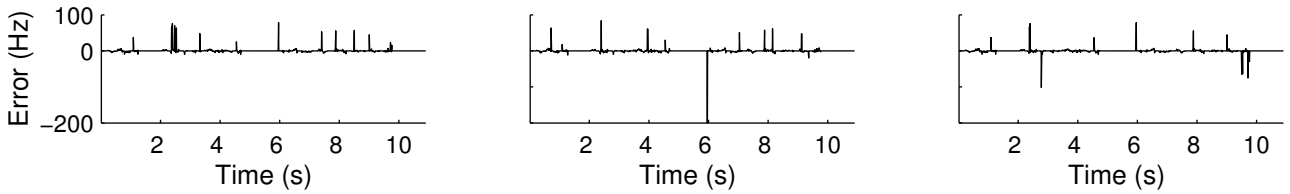

Figure 12: *Pitch estimation error obtained from speech used in Fig. 10 after being corrupted with white noise at SNR=0 dB.*