Fundamental Frequency Estimation Using Modified Higher Order Moments And Multiple Windows

Alipah Pawi Saeed Vaseghi Ben Milner[1], Seyed Ghorshi[2]

School of Engineering and Design, Brunel University, UK
[1]School of Computing Sciences, University of East Anglia, UK
[2] School of Science and Engineering, Sharif University of Technology, Kish Island, IRAN

{alipah.pawi, saeed.vaseghi}@brunel.ac.uk, B.milner@uea.ac.uk, ghorshi@sharif.edu

Abstract

This paper proposes a set of higher-order modified moments for estimation of the fundamental frequency of speech and explores the impact of the speech window length on pitch estimation error. The pitch extraction methods are evaluated in a range of noise types and SNRs. For calculation of errors, pitch reference values are calculated from manually-corrected estimates of the periods obtained from laryngograph signals. The results obtained for the 3rd and 4th order modified moment compare well with methods based on correlation and magnitude difference criteria and the YIN method; with improved pitch accuracy and less occurrence of large errors.

Index Terms — Speech, pitch, higher order moments.

1. Introduction

Pitch is the sensation of the fundamental frequency F_0 of a periodic audio signal; whereas the fundamental frequency may be accurately measured by an electronic instrument, pitch is the perception of the 'tone' of a signal by the human audio sensory system. The fundamental frequency of a periodic signal, $F_0 = 1/T_0$ expressed in units of Hz, is the inverse of the period T_0 of the signal. In this paper the terms pitch and fundamental frequency are used interchangeably.

Speech is composed of a combination of quasi-periodic and non-periodic signals. The term quasi-periodic implies that the signal is seemingly, but not strictly, periodic because the period varies over time. The pattern of time-variation of the pitch, known as intonation, conveys such information as pragmatics of speech, intent, style and accent. In English language pitch does not affect the word identity, however, in tonal languages, such as Chinese and some African languages, the word identity can change with the pitch intonation.

The harmonic plus noise character of speech is modelled as

$$x(m) = \sum_{k=1}^{N_h} a_k(m)cos(2\pi k F_0(m)m + \theta_k(m)) + v(m) \quad (1)$$

where $F_0(m)$ is the time-varying fundamental frequency at discrete-time m, $a_k(m)$ and $\theta_k(m)$ are the amplitude and phase of k^{th} harmonic of speech, N_h is the number of harmonics and $v(m)$ is the noise component of the signal [1]. Pitch extraction methods utilise the similarity of speech samples at time m, $x(m)$, with those T samples away; $x(m+T)$ or $x(m-T)$. For example, correlation-based pitch extraction methods estimate the period as the value of T for which the average of the product of $x(m)x(m-T)$ over a frame of speech samples, known as the short-time correlation, attains a maximum value [2-8]. Magnitude-difference-based pitch extraction methods estimate the period as the value of T for which the average magnitude difference $|x(m)-x(m-T)|$ over a frame of speech samples attains a minimum.

Pitch estimation [2-9] is complicated by several factors including:

(a) Half and double pitch estimation; a periodic signal with a period of T exhibits high correlation at integer multiples of T. This may lead to 'half pitch' estimation error, in cases where the similarity criterion is stronger at $2T$ than at T. For some speech segments periodicity is also exhibited at half period leading to 'double pitch' estimation in cases where the similarity criterion is a stronger at $T/2$ than at T.

(b) Voicing errors; for the purpose of pitch estimation, speech is broadly composed of two states; a voiced state with a harmonic structure and an unvoiced state with a noise-like structure. The error in detection of voice/unvoiced states effects the accuracy of pitch estimate.

(c) The time-varying nature of pitch; implies that the period, or pitch, estimated from a speech frame is at best the average period or pitch value within the frame. The period can vary substantially over a frame or it may oscillate within a frame depending on the emotional state of the voice.

(d) Indeterminate nature of some quasi-periodic speech; for some speech segments it is difficult, even for an expert, to visually determine the correct pitch value.

(e) Missing fundamental; sometimes the fundamental frequency coincides with a trough of the spectral envelop and hence the first observable harmonic is the second or a higher harmonic.

In this paper we propose novel pitch extraction methods using higher order moments criteria. The proposed methods compare well with established pitch extraction methods.

The structure of this paper is organized as follows: section 2 provides an overview of conventional pitch extraction methods. Section 3 introduces the proposed pitch extraction methods based on the higher order moments. In section 4, the effect of window length is discussed. Section 5 presents the evaluations and discussion on the performance of the proposed method. Section 6 concludes this paper.

2. Pitch Extraction Methods

In this section the main approaches to pitch extraction are reviewed.

2.1. Correlation-based Pitch Extraction

The conventional approach to pitch extraction is to estimate the period from the autocorrelation function (ACF) of speech signal [2-8]. The autocorrelation of N samples of a signal $x(m)$ for a lag T, is defined as

$$r_{xx}(T) = \frac{1}{N}\sum_{m=0}^{N-1} x(m)x(m-T) \quad (2)$$

The speech period T_0 may be estimated as the value of the lag T corresponding to maximum of ACF in the range T_{min} to T_{max}

$$T_0 = \arg\max_T r_{xx}(T) \qquad T_{min} < T < T_{max} \quad (3)$$

where T_{min} and T_{max} are the minimum and maximum values of the period. Since the ACF of a periodic signal is itself periodic, an energy maximising function that utilises the periodic energy peaks of the autocorrelation function is defined as

$$E(T) = \frac{1}{N_T}\sum_{k=1}^{N_T} r_{xx}(kT) \qquad T_{min} < T < T_{max} \quad (4)$$

where $N_T = fix(N/T)$ is the maximum number of periods that can be fitted in the N samples length of a speech frame. The estimate of the period T_0 is obtained as

$$T_0 = \arg\max_T E(T) \qquad T_{min} < T < T_{max} \quad (5)$$

2.2. Average Magnitude Difference Function

The general form of the average magnitude difference function (AMDF) is defined as [9]

$$d(T) = \frac{1}{N}\sum_{m=0}^{N-1} |x(m) - x(m-T)|^{\propto} \quad (6)$$

where for $\alpha = 1$ we have the AMDF function and for $\alpha = 2$ we have the squared magnitude difference function. The AMDF attains a minimum at the period T and its integer multiples when $x(m)$ has a value similar to $x(m-kT)$.

2.3. YIN Method

YIN method [8] uses an expansion of the squared magnitude difference criteria, Eq. (6) with $\alpha = 2$, as

$$d(T) = r_{xx}(0) + r_{xx,T}(0) - 2r_{xx}(T) \quad (7)$$

where $r_{xx}(0)$ and $r_{xx,T}(0)$ are time-varying autocorrelations at lag zero, calculated at time zero and T respectively and $r_{xx}(T)$ is autocorrelation at lag T. In [8] about 80% a decrease in pitch error is reported when the squared error function of Eq (7) is used instead of the conventional autocorrelation function of Eq (2). For more details the interested reader is referred to [8].

3. Modified Higher Order Moments

The theoretical basis for using higher order moments (HOMs) for pitch estimation may be explained as follows. Conventional pitch estimation uses the average of the product (i.e. correlation) of two samples spaced at a distance of T, $x(m)$ and $x(m-T)$, as a measure of similarity or periodicity. This idea can be extended to employ the average of the product of K samples, $x(m)x(m-T)\ldots x(m-KT)$, as the similarity criterion for a proposed value of period T. The advantage gained is a greater reinforcement of the product of K similar samples that may result in a sharper criterion.

The modified higher order moments, introduced in this paper, are obtained by splitting a signal $x(m)$ into a positive-amplitude $x_+(m)$ part and a negative-amplitude $x_-(m)$ part defined as

$$x_+(m) = \begin{cases} x(m) & if\ x(m) > 0 \\ 0 & otherwise \end{cases} \quad (8)$$

$$x_-(m) = \begin{cases} x(m) & if\ x(m) < 0 \\ 0 & otherwise \end{cases} \quad (9)$$

clearly $x(m) = x_+(m) + x_-(m)$. The reason for splitting the signal $x(m)$ into a positive-amplitude and negative-amplitude parts is to prevent cancellation, instead of desired reinforcement, of similar samples. For example the sum of the third order reinforcement of the positive-amplitude part of a signal with a pitch period of T, $x(m)x(m-T)x(m-2T)$ can cancel the third order reinforcement of the negative-amplitude part of the signal if the signals are not split as proposed here.

The K^{th} order modified moment is defined as

$$m_K(T) = \frac{1}{N_1}\left\{ \sum_{m=0}^{N_1} x_+(m)\cdots x_+(m-(K-1)T) + abs\left[\sum_{m=0}^{N_1} x_-(m)\cdots x_-(m-(K-1)T)\right]\right\} \quad (10)$$

where $N_1 = N - (K-1)T$. For estimation of period, based on the K^{th} order modified moment, an energy maximising function is defined as

$$E(T) = \frac{1}{N_T}\sum_{l=1}^{N_T} m_K(lT) \qquad T_{min} < T < T_{max} \quad (11)$$

where $N_T = fix(N/T)$ is the maximum number of periods that can be fitted in the function $E(\cdot)$. The estimate of the period T_0 is obtained as

$$T_0 = \arg\max_T E(T) \qquad T_{min} < T < T_{max} \quad (12)$$

4. Effect of Window Length on Pitch Error

The choice of speech window length has a significant impact on pitch estimation [7-9]. Two considerations that influence the choice of the speech window length are: (a) the time-variations of the speech signals and (b) the maximum delay allowed in voice communication. In voice communication systems usually the speech window length is chosen as 20 ms equivalent to 160 samples at a sampling rate of 8 kHz. For pitch estimation we suggest that a larger window length, spanning the current and several past frames, can be used to good advantage without compromising the delay constraint of the voice communication system as explained here.

From the estimation theory, specifically the Cramer-Rao lower bound, the variance of estimation error decreases with the increasing observation length [10]. Hence, as expected the choice of the speech window (or frame) length has a substantial influence on the variance of the pitch error. Generally pitch estimates improve with the increasing window length within a voiced segment of speech. However, if the window length is too large, the pitch estimate will not accurately follow the smooth variation of the pitch utterance curves and will give rise to a coarse estimate of the pitch curve that is step-wise.

To take advantage of the robust but coarse estimate offered by a larger window length and the finer time resolution offered by a shorter window length, a two stage approach is utilised in which:

(1) Initially a larger window length, based on concatenation of the current frame and a number of consecutive past speech frames, is used to obtain a robust estimate of F_0.
(2) The pitch estimation process is repeated over a shorter window length, composed of the current frame only, to obtain a locally optimised value of the pitch around the value of F_0 in the range $F_0 \pm aF_0$, where $0 < a < 1$. Typically $a = 0.1$.

The evaluation results are presented in the next section.

5. Evaluations and Discussion

Experiments are performed to evaluate and compare the performance of the proposed pitch extraction method with several established methods namely the correlation-based ACF, the AMDF method and the YIN method.

5.1 Speech Databases Used for Evaluation

Experimental evaluation for all pitch extraction methods have been performed on publicly available databases containing simultaneous recordings of speech and laryngograph signals[1]. The laryngograph signals provide a relatively clean recording of glottal vibrations from which the period and its inverse, the pitch, can be extracted with a high degree of accuracy. A zero-crossing method was used for the marking and extraction of the period information from the larnygograph signal. All the extracted pitch data were visually inspected and hand-corrected where necessary.

The speech databases used here, for the extraction of reference pitch and the evaluation of pitch extraction methods, contain phonetically-balanced utterances from ten male speakers with 304 utterances and ten female speakers with 215 utterances. All speech signals were resampled to 8 kHz which is the standard sampling rate for mobile phones. For fundamental frequency estimation, the maximum and minimum values of period and fundamental frequency are set as T_{min}=2.5 ms corresponding to $F_{0,\max}$=400 Hz and T_{max}=20 ms corresponding to $F_{0,\min}$=50 Hz.

5.2. Pitch Error Analysis Method

The percentage pitch error for speech frame m is defined as

$$E(m) = \frac{\left|\hat{F}_0(m) - F_0(m)\right|}{F_0(m)} \times 100\% \qquad (13)$$

where $F_0(m)$ and $\hat{F}_0(m)$ are the true value (obtained from laryngographs) and the estimate of pitch respectively. Note that the pitch error is calculated over voiced frames only. The voicing information is reliably obtained from the laryngeal signal.

For analysis of pitch accuracy three categories of mean percentage pitch error are considered: the overall error for all $E(m)$, small (fine) errors for $E(m) < 20\%$ and large (gross) errors for $E(m) \geq 20\%$. The choice of a threshold value of 20% is arbitrary, however, it is also used by other researchers to assess the tendency of a pitch extraction method to produce gross errors including half and double pitch estimates.

Figure 1 provides a comparative illustration of the shape of the pitch extraction criteria namely the correlation function ACF, the AMDF function and the higher order moments HOMs. Note that HOMs form a sharper curve around the period and its multiples.

5.3 Evaluation of the Effect of Window Length

Figure 2 illustrates the variation of pitch estimation error with a range of window lengths of 20ms (or 160 samples at 8 kHz sampling rate), 25ms (200 samples), 33ms (264 samples as in YIN method), 50ms (400 samples), 62.5 ms (500 samples), 75ms (600 samples), 87.5ms (700 samples) and 100 ms (800 samples). From Fig.2 a window length of 50 ms (400 samples) provides the least error among the selected window lengths. The increase in pitch error beyond the duration of 50 ms can be attributed to the time-varying characteristic of pitch signals.

Based on this result we employed a two-stage pitch estimation strategy whereby a larger window length of 400 samples (note for a delay sensitive communication system this larger speech window may include a number of stored past frames) is used to obtain an initial coarse but robust estimate which is subsequently fine-tuned in the locality of the current short frame of 160 samples long as explained in section 4.

5.4 The Effect of Varying Signal to Noise Ratio

The pitch extraction methods were evaluated in a range of signal-noise-ratio (SNR) from 30 dB down to 0 dB. The noisy speech samples were obtained by adding Gaussian white noise to clean speech signal. In these experiments a two stage strategy for pitch extraction was adopted whereby an initial estimate from a window of 400 samples was followed by a localized estimate from a window of 160 samples as described in section 4.

The evaluation results plotted in Figures 3(a-d) display the average overall pitch percentage error in the presence of four types of noise. The proposed methods based on the two-stage window length strategy outperform YIN in most cases.

Figure 4 (a-b) displays the average of fine and gross pitch errors for white noise. It is evident that for a range of SNRs

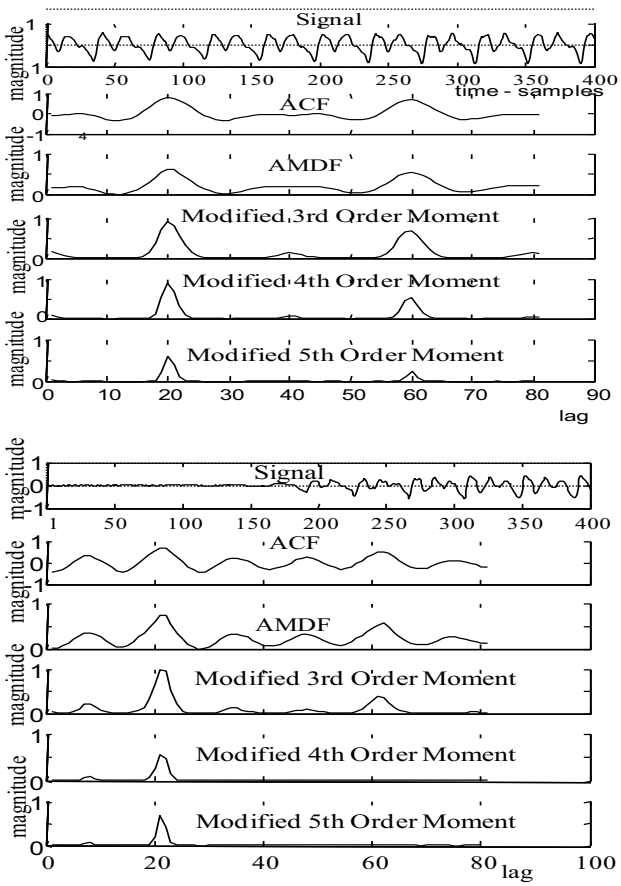

Figure 1: *Comparative illustration of the sharpness of ACF, AMDF, third-fifth order moments at a window length of 50 ms. Top panel; a segment of voiced speech. Bottom panel; a transitory segment, unvoiced speech followed by voiced speech.*

[1] http://www.festvox.org/dbs_kdt.html;
http://www.cstr.ed.ac.uk/research/projects/fda;
http://www.cstr.ed.ac.uk/research/projects/artic/mocha.html;
http://www.festvox.org/cmu_artic;
ftp://ftp.cs.keele.ac.uk/pub/pitch.

from 0 dB to 30 dB the proposed methods yield less large errors and are hence more robust compared to the YIN benchmark.

Figure 5(a-b) is the population of errors for errors less than 20% and greater than 20% respectively in the presence of white noise. This figure shows that the third order and fourth order moment methods results in a significantly smaller percentage of population of the pitch errors larger than 20%.

Furthermore, even the correlation method performs well compared with YIN when the strategy of estimating the pitch from a longer window followed by a localized estimate is used. This result underscores the importance of window length in pitch estimation.

6. Conclusions

This paper presented a modified form of higher order moments HOMs as the objective criteria for pitch extraction and explored the influence of window length on pitch error. The pitch extraction methods, based on the third order and fourth order moments compete favourably with the conventional methods and the benchmark yin method. Furthermore, the use of the two stage method, whereby a larger window length is initially used to obtain a coarse estimate of pitch followed by fine tuning in the locality of the current speech frame, yields improved and robust pitch estimates.

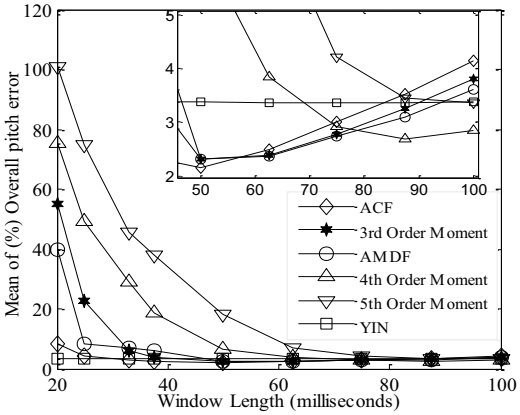

Figure 2: *The mean of (%) pitch error versus window lengths at clean signal (30dB SNR): from 20ms to 100ms windows length and inserted figure is zoomed in 50 ms to 100 ms windows length.*

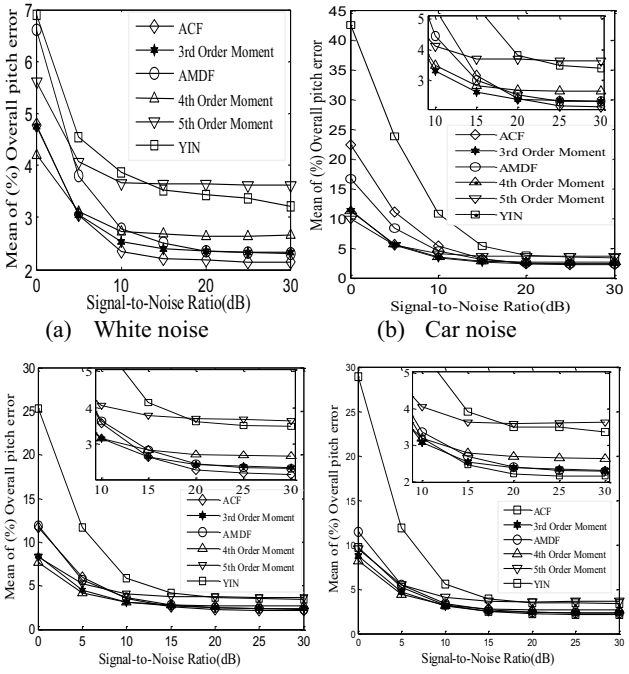

Figure 3: *The mean of overall (%) pitch errors as a function of SNR.*

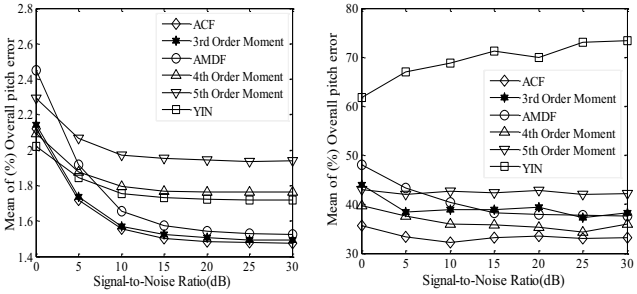

Figure 4: *(a) The mean of (%) fine pitch errors < 20%, (b) the mean of (%) gross pitch errors > 20% as a function of SNR.*

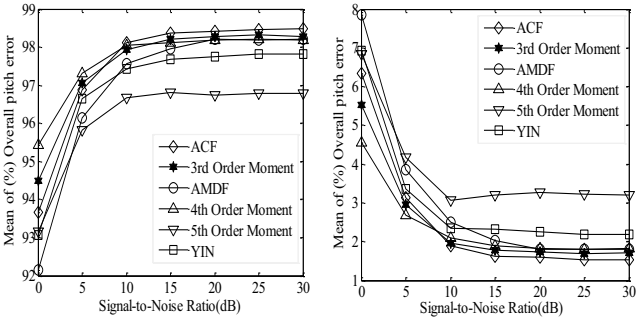

Figure 5: *The Mean (%) Population Pitch Errors (a) ≤ 20%, and (b) > 20%.*

7. References

[1] Y. Stylianou, "*Harmonic plus Noise Models for Speech, combined with Statistical Methods, for Speech and Speaker Modification,".* PhD thesis, Ecole Nationale Sup`erieure des T´el´ecommunications, Jan 1996.

[2] M. M. Sondhi,"*New Methods of Pitch Extraction*," IEEE Trans., Audio and Electro-acoustic, vol.16, pp. 262-266, 1968.

[3] L. R. Rabiner et al., "*A Comparative Performance Study of Several Pitch Detection Algorithms*," Acoustics, Speech and Signal Processing, IEEE Trans. on, vol. 24, pp. 399- 418, 1976.

[4] T. Shimamura, and H. Kobayashi, "*Weighted Autocorrelation for Pitch Extraction of Noisy Speech*," IEEE trans. on speech and audio processing, vol.9, pp.727-730, 2001.

[5] R. J. McAulay and T. F. Quatieri, "*Pitch Estimation and Voicing Detection Based on a Sinusoidal Speech Model,"* Acoustics, Speech, and Signal Processing, 1990. ICASSP-90., 1990 International Conference on, vol. 1, 1990.

[6] K.Hirose et al.," *A Scheme For Pitch Extraction of Speech Using Autocorrelation Function With Frame Length Proportional to The Time Lag*," Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on, vol. 1, pp. 149–152, 1992.

[7] T.Takagi et al., "*A method for Pitch Extraction of Speech Signals Using Autocorrelation Functions Through Multiple Window Lengths*", Electronics and Communications, part 3, 83(2), pp. 67-79, 2000.

[8] A.de Cheveigne and H. Kawahara, "*YIN, a Fundamental Frequency Estimator for Speech and Music*," J. Acoust. Soc. Am., vol. 111, pp. 1917-1930, 2002.

[9] M. J. Ross et al., "*Average Magnitude Difference Function Pitch Extractor,"* Acoustics, Speech and Signal Processing, IEEE Transactions on, vol. 22, pp. 353-362, 1974.

[10] S. V. Vaseghi, *Advanced Digital Signal Processing and Noise Reduction*, 4th Edition, Wiley 2009.