A TIME-FREQUENCY DOMAIN APPROACH FOR PITCH-PEAK PICKING OF NOISY SPEECH

Celia Shahnaz, Student Member, IEEE, Md. Rokanuzzaman, Md. M. Adnan, B. P. Shrestha, and S. A. Fattah, Student Member, IEEE

Department of Electrical and Electronic Engineering
Bangladesh University of Engineering and Technology, Dhaka-1000, Bangladesh
email: celia@eee.buet.ac.bd

ABSTRACT

This paper presents an improved autocorrelation based algorithm for pitch estimation of noisy speech. The autocorrelation function of the pre-processed noisy speech signal is weighted by an inverse of its variant, circular average magnitude difference function (CAMDF). The inversed CAMDF-weighted autocorrelation function (ICWAF) alone cannot guarantee the de-emphasization of the 'non-pitch-peaks'. However, as the dominant pitch-harmonic (DH) of the sinusoidal clean-speech model is expected to be an integer multiple of the pitch, this method concentrates in the effective estimation of the DH using a proposed cosine model of autocorrelation function. Hence, a variable period impulse train with its period being a function of the DH is optimally fitted with the rectified ICWAF to estimate the pitch. The simulation results using the Keele database show that the proposed method claims better accuracy even at a signal to noise ratio (SNR) as low as –10 dB.

1. INTRODUCTION

The pitch information of speech signal has received importance especially in speaker identification, speech recognition, speech synthesis and speech articulation training aids for the deaf. Several techniques have been focused in the literature for pitch extraction from clean speech [1-4]. Among them, the cepstral method [4] is a well-known pitch extraction approach but it yields unsatisfactory results for noisy speech. In noisy condition, the variant of the autocorrelation function (ACF) method [5] can estimate pitch with higher accuracy for female speakers but show deteriorating performance for low pitched male speakers. The correlation-based average magnitude difference function (AMDF) [3] method has the advantage of low computation but at low SNR, the generated pitch errors consist mostly of pitch doublings. Moreover, as there is a decreasing slope in the AMDF curve, the possibility of emphasization of pitch-peak cannot be guaranteed in strong noisy environment using AMDF-weighted autocorrelation (AWAC) method reported in [6]. Though the circular average magnitude difference function (CAMDF) method [7] conquers the defects of AMDF and introduces some improvement in pitch estimation, it exhibits poor performance in case of male speakers. Recently a sinusoidal autocorrelation model for pitch extraction is proposed in [8] assuming speech as the output of an AR system. This method claims better results at a low SNR only for short-duration sentences uttered by limited number of speakers.

In this paper, a time-frequency domain approach is proposed for pitch estimation. Here, in the time domain, autocorrelation function of the pre-filtered noisy speech (PFNS) signal is weighted by an inversed circular AMDF to emphasize the peak at the pitch lag. Yet to reduce the adverse effects of 'non-pitch-peaks' due to presence of noise, the dominant pitch-harmonic (DH) in the sine-wave clean-speech model is pre-estimated in the frequency domain using a cosine model of autocorrelation function. For pitch estimation, the periodicity of the inversed CAMDF-weighted autocorrelation function (ICWAF) is best matched with an impulse train by varying its period according to the integer multiple of the estimated DH. Using this proposed approach, the accuracy of pitch estimation can be significantly improved relative to the conventional autocorrelation based pitch estimators.

2. PROBLEM FORMULATION

Let $x(n)$ is the clean speech and $v(n)$ denotes the random white Gaussian noise with zero mean and variance σ_v^2. The observed noisy speech $y(n)$ is given by

$$y(n) = x(n) + v(n) \tag{1}$$

The autocorrelation function of $y(n)$, $R_{yy}(\tau)$, is obtained as

$$\begin{aligned} R_{yy}(\tau) &= R_{xx}(\tau) + \sigma_v^2, \quad \text{for } \tau = 0 \\ &= R_{xx}(\tau), \quad \text{for } \tau \neq 0 \end{aligned} \tag{2}$$

where $R_{xx}(\tau)$, the autocorrelation function of $x(n)$, is given by

$$R_{xx}(\tau) = \frac{1}{N} \sum_{n=0}^{N-1-|\tau|} x(n)x(n+|\tau|) \tag{3}$$

Here N is the length of the speech analysis frame and τ is the lag number. $R_{xx}(\tau)$ essentially exhibits peaks at the periodicity (T) of $x(n)$ (i.e., at $\tau = \rho T$, where ρ is an integer). The location of the second maximum peak (at $\tau = T$) relative to the largest peak (at $\tau = 0$) of $R_{xx}(\tau)$ is generally used to estimate the pitch period. Such method of pitch estimation using $R_{yy}(\tau)$ of equation (2) is noise robust

as long as speech and noise are truly uncorrelated, a requirement seldom met in practice. It is worth mentioning that at a very low SNR, $R_{yy}(\tau)$ includes significant error due to the existence of correlation between $x(n)$ and $v(n)$. To reduce the noise effect, the observed signal $y(n)$ is pre-filtered in the pitch domain [$2\pi(50\text{Hz}/f_s)$~$2\pi(500\text{Hz}/f_s)$] of most male and female speakers using discrete cosine transform (DCT) where f_s is the sampling frequency. The reconstructed pre-filtered noisy speech (PFNS) signal, $\varphi(n)$, contains less noise as the higher frequency components mainly due to noise are suppressed.

Similar to the autocorrelation function (ACF) [2], the AMDF [3] is periodic with speech periodicity and can be defined as

$$\xi(\tau) = \frac{1}{N}\sum_{n=0}^{N-1-|\tau|}\left|\varphi(n) - \varphi(n+|\tau|)\right| \qquad (4)$$

It is expected that the AMDF $\xi(\tau)$ exhibits deep notches where the ACF shows peaks (at $\tau = \rho T$). As in [6], the AMDF-weighted autocorrelation function can be used to estimate pitch. However under heavy noisy condition, the minimum notch in $\xi(\tau)$ may occur at $T < \tau \le 2T$ whereas the circular AMDF (CAMDF) [7] is supposed to provide a minimum notch at the true pitch period. Hence we have proposed an inversed CAMDF-operated autocorrelation function (ICWAF) to suppress the unnecessary erroneous peaks. Even with the aforementioned preprocessing, the pitch determination algorithm using the first and second largest peaks of ICWAF may give higher error rates at a very low SNR. As there is still room for improvement, we propose to determine the DH of the harmonic sine-wave clean-speech model to estimate pitch using a cosine model of autocorrelation function. Here the accuracy of the proposed method no longer depends on the position of the second largest peak of ICWAF, rather the novelty lies with the accurate determination of the DH.

3. PROPOSED METHOD

The main idea behind the concept of using AMDF in the autocorrelation domain was to divide the autocorrelation function by the AMDF to enhance peaks at the pitch periodicity of the autocorrelation function. However, using the conventional AMDF has a serious detrimental effect on the pitch-peak due to the falling trend of minima with increasing lag of AMDF. Instead of enhancing the pitch-peak, this in turn may make the other 'non-pitch-peaks' more pronounced and thereby contributes in deteriorating the performance of the AWAC [6] method.

We have adopted the property of horizontal developing minima of circular AMDF (CAMDF) [7] modifying the original definition of AMDF. The CAMDF is defined as

$$\xi'(\tau) = \sum_{n=0}^{N-1-|\tau|}\left|\varphi_W\left(\text{mod}(n+\tau, N)\right) - \varphi_W(n)\right| \qquad (5)$$

where mod ($n + \tau$, N) represents the modulo operation, meaning that ($n + \tau$) modulo N and $\varphi_w(n)$ is the windowed PFNS signal. The CAMDF shows a minimum (a deep valley point) at a lag equal to integer multiple of the period of the signal and the values of the front deep valley points (notches) are no less than that of the rear. Hence, to enhance the pitch-peak relative to the other peaks, we have proposed a modified autocorrelation function $\chi'(\tau)$ where the autocorrelation of the PFNS signal, $R_{\varphi\varphi}(\tau)$, is weighted by an inversed CAMDF and is given by

$$\chi'(\tau) = \frac{R_{\varphi\varphi}(\tau)}{\xi'(\tau) + \varepsilon} \qquad (6)$$

Here, ε is a small positive constant. As we are only concerned about the peaks of $\chi'(\tau)$, a half-wave rectified autocorrelation function $\psi(\tau)$ is defined as

$$\begin{aligned} \psi(\tau) &= \chi'(\tau) \quad \text{for} \quad \chi'(\tau) \ge 0 \\ &= 0 \quad \text{for} \quad \chi'(\tau) < 0 \end{aligned} \qquad (7)$$

However at a very low SNR, presence of spurious peaks obscure the desired pitch-peak of $\psi(\tau)$. Hence, a pre-estimation of pitch is obtained from the harmonic sine-wave speech model using a cosine model of autocorrelation function.

A frame of clean-speech waveform $x(n)$ can be represented by a set of sinusoidal waveforms for which all the frequencies are harmonically related as

$$x(n) = \sum_{k=1}^{r}\bar{b}(\omega_k)\exp\left[j(n\omega_k + \bar{\theta}_k)\right], \qquad \omega_k = k\omega_0 \qquad (8)$$

In this harmonic sine-wave speech model, r is the number of harmonics in the speech, $\bar{b}(\omega_k)$ represents the envelope of the vocal tract, $\bar{\theta}_k$ denotes the phase of the k-th harmonic, and $\omega_0 = 2\pi f_0/f_s$ is the fundamental angular frequency that corresponds to the true pitch period T. Strictly speaking, the above analysis is nearly exact for a strongly-voiced speech frame. However, it is a crude approximation for a frame of voiced/unvoiced mixture. Using the harmonic sine-wave speech model of x(n) given in equation (8), the autocorrelation of the noise-free speech signal can be derived as

$$R_{xx}(\tau) = \sum_{k=1}^{r} A_k \cos(\omega_k\tau), \tau \ge 0, \ \omega_k = k\omega_0 \qquad (9)$$

where, A_k is a constant ($A_k = b_k^2/2$). In deriving equation (9), the contributions from the cross-product terms of different harmonics are neglected. Unlike conventional approaches, we intend to estimate only the dominant harmonic (DH) of the harmonic sine-wave speech model rather than calculating all of the harmonics, and use it to determine the pitch. For our purpose, it is sufficient to estimate only the dominant component function, e.g., $R_k(\tau) = A_k\cos(\omega_k\tau)$ of equation (9). It is done by optimally fitting a finite sequence (e.g., τ = 1, 2, ……, M) of $R_k(\tau)$ with $R_{\varphi\varphi}(\tau)$. The fitted parameters will give an estimate of ω_k

and A_k. Mathematically ω_k and A_k of $R_k(\tau)$ are chosen iteratively such that the sum-squared error (SSE) $J^{(i)}(\omega_k)$ between $R_{\varphi\varphi}(\tau)$ and $R_k^{(i)}(\tau)$

$$J^{(i)}(\omega_k) = \sum_{\tau=1}^{M} \left| R_{\phi\phi}(\tau) - R_k^{(i)}(\tau) \right|^2 \qquad (10)$$

is minimized, where $R_k^{(i)}(\tau) = A_k^{(i)} \cos(\omega_k^{(i)} \tau)$ and superscript '(i)' represents the iteration number. The value of '(i)' depends on the range of $\{\omega_k\}$ within the pitch domain and the resolution of searching those $\{\omega_k\}$. For a given $\omega_k^{(i)}$, the corresponding $A_k^{(i)}$ can be uniquely determined by solving the equation $[dJ^{(i)}(\omega_k)/dA_k^{(i)}] = 0$. The optimum parameters are found as $A_k = A_k^{(i)}$, and $\omega_k = \omega_k^{(i)}$ for the value of '(i)' at which $J^{(i)}(\omega_k)$ is minimum in the least-squares sense. The frequency ($\omega_k/2\pi$) obtained from this best matched cosine component function $R_k(\tau)$, is the desired dominant harmonic f_d in the proposed method. Since the optimum ω_k is an integer multiple of ω_0, the obtained f_d is used to estimate an initial period T_d. To estimate true pitch period T from T_d, enhanced half-wave rectified autocorrelation function, $\psi(\tau)$, is weighted by an impulse train $d(n, \alpha)$ of length M containing a fixed number (λ) of unit impulses. The period of the impulse train (T_i) is taken to be equal to or some multiple of T_d ($T_i = \alpha T_d$, α is a nonzero positive integer). Thus T_i is varied iteratively for different values of α. For a particular value of α, the impulse train can be defined as

$$d(n,\alpha) = \sum_{\mu=0}^{\lambda-1} \delta(n - \mu\alpha T_d), \quad \text{for } n = 0, 1, \ldots, M-1 \qquad (11)$$

where, $\delta(n)$ is the Kronecker delta function. Finally, an objective function is defined by taking the inner product of $\mathbf{D}(\alpha) = [d(0, \alpha)\ d(1, \alpha) \ldots\ldots d(M-1, \alpha)]$ with $\boldsymbol{\psi}$ as

$$\eta(\alpha) = \mathbf{D}(\alpha)\, \boldsymbol{\psi}' \qquad (12)$$

where $\boldsymbol{\psi} = [\psi(0)\ \psi(1) \ldots\ldots \psi(M-1)]$ and $'$ denotes the transpose operation. The cost function $\eta(\alpha)$ is maximized by varying α. Since peaks of $\boldsymbol{\psi}$ are expected to occur at the points of integer multiple of T, $\eta(\alpha)$ would be maximum at the location where the impulse train period (T_i) matches with T. The value of α corresponding to the maximum of $\eta(\alpha)$, denoted by α_m, is used to estimate the desired pitch period as $T_{est} = \alpha_m T_d$.

4. EXPERIMENTS

4.1 Experimental Details

The performance of the proposed method is tested using the Keele pitch extraction reference database obtained from <ftp://ftp.cs.keele.ac.uk/pub/pitch/>. The core data consists of a phonetically balanced text named, "*The North Wind Story*". The speech signal is sampled at 20 kHz with 16-bit resolution. For many frames the original database has reference pitch values where the signal is hardly periodic. Excluding these frames manually after visual inspection a 'modified reference database' of 'clearly voiced' frames [9] has been developed. We have performed simulations for all the speakers and present results for three female (S_{F1}, S_{F2}, S_{F3}) and three male (S_{M1}, S_{M2}, S_{M3}) speakers using the modified database. The experiments were conducted by adding white Gaussian noise to the speech samples according to equation (1). Each 25.6-msec analysis frame was weighted by a 512-point rectangular window. The frame shift was set to be 10-msec as used to generate the reference pitch values given in the original database. Note that the value of the fixed number ε in equation (6), introduced to avoid the singularity constraint, was set to 1 as in [6]. We have used $R_{\varphi\varphi}(\tau)$ for $\tau = 1, 2 \ldots, M$ with M = 300 and M = 500 for female and male speakers, respectively to contain 2 to 3 complete pitch periods. The number of unit impulses ($\lambda \geq 2$) in the impulse train must be kept constant for a particular speaker. The parameter α in equation (11) is restricted by $(\lambda-1)\alpha T_d \leq (M-1)$. For each voiced frame of speech, the DH was searched from 60 to 500 Hz for female speakers and from 50 to 250 Hz for male speakers with a scanning resolution of 0.001. SNRs of the noisy speech signals used were ranging from –10 to 30 dB. Based on Rabiner's method [1], the error parameter $e(m)$ is defined as the difference between the true pitch period obtained from the 'modified reference database' and the pitch period determined using the proposed method. If $|e(m)| \geq 20$ samples (more than 1 msec error), we recognized the error as a gross pitch error (GPE), otherwise, the resulting error was termed as the fine pitch error (FPE). At a particular SNR, the Global %GPE is calculated from the total number of 'clearly voiced' frames of all three female (or male) speakers in which GPE occurs. The total number of 'clearly voiced' frames is 4839 for female and it is 4471 for male speakers.

4.2 Results and Performance Comparison

To show the superiority of the proposed method, we compare the results obtained with those of the AWAC method [6]. Figs. 1 and 2 focus on the trend of the performance evaluation index, Global %GPE, investigated at different SNRs for both female and male speakers. It is evident that for the proposed method, the Global %GPEs are significantly reduced in comparison to the AWAC method and performance is quite outstanding for all speakers at all SNRs ranging from –10 to 30 dB. The mean (m_{FPE}) and the standard deviation (σ_{FPE}) [the entries in the brackets] of FPEs using the proposed and the AWAC methods are summarized in Table 1 especially at SNR = –10 dB**.** For both female and male speakers it is observed that the m_{FPE} and the σ_{FPE} of the proposed method are lower relative to those of the AWAC method. From the simulations it is found that they are also within the acceptable limit and consistently satisfactory at other SNRs. Hence, the proposed method claims higher degree of accuracy both at high and low SNRs.

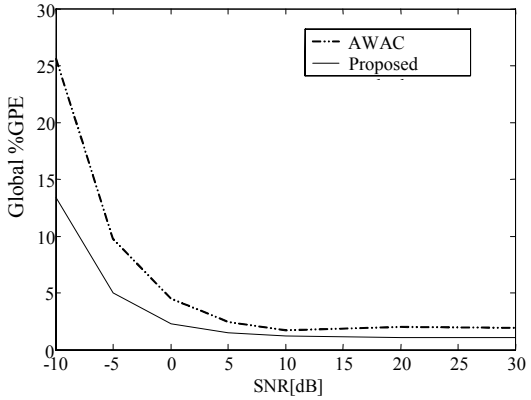

Fig. 1: Performance comparison in terms of Global %GPE of female speakers at different SNRs

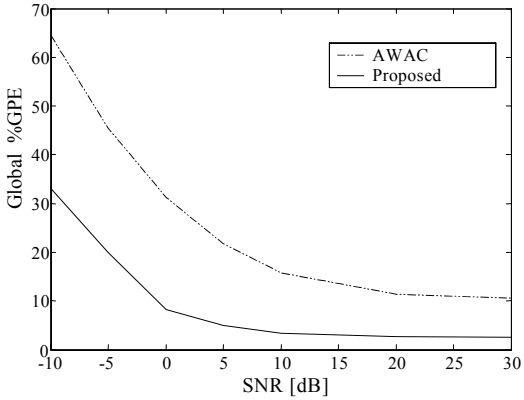

Fig. 2: Performance comparison in terms of Global %GPE of male speakers at different SNRs

Speakers	AWAC method	Proposed method
S_{F1}	7.0337 (8.2889)	6.2188 (6.6136)
S_{F2}	5.7255 (6.1490)	5.3657 (5.2701)
S_{F3}	7.3137 (8.0604)	6.3298 (6.3479)
S_{M1}	5.3908 (4.5422)	5.1688 (4.1510)
S_{M2}	5.8028 (4.3693)	5.1193 (3.9740)
S_{M3}	4.000 (3.300)	4.000 (3.100)

Table 1: Performance comparison of the proposed method with the AWAC method in terms of mean (m_{FPE}) and the standard deviation (σ_{FPE}) [in brackets] of the fine pitch errors at SNR = –10 dB.

5. CONCLUSION

This paper addresses a novel method of pitch extraction at a very low SNR using the inversed CAMDF-weighted autocorrelation function of the pre-filtered noisy speech (PFNS) signal. The use of the CAMDF conquers the limitations of the AMDF and can be used to enhance the autocorrelation function. Still the inversed CAMDF-weighted autocorrelation function cannot make the pitch-peak highly pronounced in practical heavy noisy condition. To overcome this constraint, we are motivated to utilize the dominant pitch-harmonic (DH) of sinusoidal speech model. The kernel of this method lies in the accurate estimation of the DH using the proposed cosine model of autocorrelation function. The proposed method outperforms the conventional and recent correlation-based algorithms in terms of the Global %GPE. Since the competitive values of mean and standard deviation of FPEs are praise-worthy, this method not only efficiently reduces the gross errors but also improves the precision of pitch estimation at all SNRs.

REFERENCES

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

[2] L. R. Rabiner, "On the use of autocorrelation analysis for pitch detection," *IEEE Trans. Acoust., Speech, Signal Processing,* vol. ASSP-25, no. 1, pp. 24-33, 1977.

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

[4] A. M. Noll, "Cepstrum pitch determination," *J. Acoust. Soc. Amer.*, vol. 41, pp. 293-309, 1967.

[5] D. A. Krubsack and R. J. Niederjohn, "An autocorrelation pitch detector and voicing decision with confidence measures developed for noise-corrupted speech," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-39, no. 2, pp. 319-329, 1991.

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

[7] W. Zhang, G. Xu and Y. Wang, "Pitch estimation based on circular AMDF," in *Porc. ICASSP2002*, Orlando, Florida, USA, May 2002, pp. 341-344.

[8] Md. K. Hasan, C. Shahnaz and S. A. Fattah, *"Determination of Pitch of Noisy Speech Using Dominant Harmonic Frequency",* in *Proc. ISCAS 2003*, Bangkok, Thailand, May 2003, pp. 556-559.

[9] J. Tabrikian, S. Dubnob, and Y. Dickalov, "Speech enhancement by harmonic modeling via MAP pitch tracking," in *Porc. ICASSP2002*, Orlando, Florida, USA, May 2002, pp. 549-552.