A ROBUST PITCH ESTIMATION ALGORITHM IN NOISE

C. Shahnaz, Student Member, IEEE, W. -P. Zhu, Senior Member, IEEE, and M. O. Ahmad, Fellow, IEEE

Centre for Signal Processing and Communications, Dept. of Electrical and Computer Engineering
Concordia University, Montreal, Quebec, Canada H3G 1M8

ABSTRACT

In this paper, we present a robust pitch estimation algorithm for noise-degraded speech. We propose a new circular average magnitude sum function (CAMSF) and a pseudo normalized correlation function (PNCF) both of which exhibit the periodicity at the pitch period of voiced speech. Exploiting the fact that CAMSF produces a peak while PNCF shows a notch, an integrated time-domain function (ITDF) is developed to enhance the pitch-harmonic-notches in presence of noise. Moreover, a frequency-frame relative smoothed noisy spectrum that acts as a harmonic spectral structure enhancer is utilized to accurately acquire a pitch-harmonic (PH) from noisy speech. We argued that employing the PH, pitch information can be effectively extracted through a variable-period impulse-train in conjunction with the proposed ITDF. It has been ascertained that the overall algorithm simulated using the *Keele* reference database is able to outperform some of the existing methods and well suited for a wide range of signal-to-noise ratios (SNRs) upto -10 dB.

Index terms — Pitch estimation, low SNR, circular average magnitude sum function, pseudo normalized correlation function, Impulse-train.

1. INTRODUCTION

Pitch period (or fundamental frequency) extraction plays an eminent role in different application areas [1] such as speech enhancement using the harmonic model, automatic speech recognition and understanding (ASRU), analysis and modeling of speech prosody, and low-bit rate speech coding. Pitch patterns are axiomatically important pre-requisite in speaker recognition and synthesis. Real time pitch displays can also give feedback to the deaf learning to speak.

The pitch detection algorithms (PDAs) face the real challenge in presence of noise. Among several PDAs, the autocorrelation function (ACF) based methods [2] exhibit a better performance in noise for females relative to males. The average magnitude difference function (AMDF) based approaches [2]-[3], have an advantage over ACF based methods with respect to speed but the magnitude of principal minimum of the AMDF is highly influenced by intensity variation and the background noise of speech causing half- and double pitch-errors. The recent PDA based on AMDF and ACF [4] is shown to be accurate in clean speech for a limited number of frames only, also it has not been tested for noisy speech. Although the circular average magnitude difference function (CAMDF) [5] is reported as advantageous over the AMDF to some extent, its pitch extraction performance for noisy speech is evaluated and found unsatisfactory in [6]. It is to be noted that due to the falling trend of the notches of the AMDF, the weighted autocorrelation (WAC) method [7] employing the AMDF emphasizes the non-pitch peaks and misguides pitch estimation especially at a very low SNR.

In this work, we propose an effective pitch estimation method for speech severely corrupted by noise with reliability and accuracy as the prime focus. Our contributions are two fold: (1) the notches only at the integer multiple of the pitch period are emphasized through an ITDF obtained by well exploiting the properties of the proposed CAMSF and PNCF , (2) a PH is accurately attained from a smoothed noisy spectrum with enhanced pitch-harmonic components. It should be stressed that the PH in collaboration with an impulse-train can optimally match the periodicity of the ITDF thus ensuring the pitch estimation with guaranteed accuracy and consistency in heavy noise.

2. PROPOSED METHOD

Assuming that the clean speech $x(n)$ is contaminated by a zero-mean additive white Gaussian noise (AWGN) $v(n)$, the observed noisy speech $y(n)$ can be written as,

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

The noisy speech $y(n)$ is segmented into frames with a frame-size N by the application of a window function $w(n)$, yielding a windowed noisy speech frame denoted as $y_w(n)$.

2.1. Pre-processing

Pre-processing of $y_w(n)$ is performed in the Fast Fourier Transform (FFT) domain. Since voiced speech is dominated by the energy in the first-formant range, the FFT coefficients corresponding up to the upper frequency limit of the first formant should be retained and the rest of the co-efficients should be set to zero. Then, a time-domain pre-processed noisy speech frame $y'_w(n)$ can be obtained through the

1-4244-0728-1/07/$20.00 ©2007 IEEE

inverse FFT. The proposed FFT pre-processing removes the influence of higher formants as well as that of the high frequency noise components while it preserves sufficient strong harmonics in $y'_w(n)$ to improve the accuracy of pitch estimation [1].

2.2. Enhancement in time-domain

In order to extract the pitch period of a frame of voiced speech, we propose a new short-time circular average magnitude sum function (CAMSF) of $y'_w(n)$ as given by,

$$\Gamma_y(\tau) = \sum_{n=0}^{Q-1} \left| y'_w(z) + y'_w(n) \right|, \tau \in \left[0 : \frac{Q}{2} \right] \tag{2}$$

where, Q is defined as,

$$Q = N + \tau_a \tag{3}$$

with τ_a being the number of additional speech samples considered from the next frame, τ represents the lag variable, and z the modulo of ($n + \tau$) with respect to Q, namely,

$$z = \mathrm{mod}(n + \tau, Q) \tag{4}$$

For a quasi-periodic frame of voiced speech with an approximate pitch period of T_0, $\Gamma_y(\tau)$ is found to exhibit local maxima at $\tau = \rho T_0$, ρ is an integer, ρ =0,1,2….. Through an analysis of many voiced frames, we have ascertained that it may be possible to estimate T_0 according to the location of the global maximum of $\Gamma_y(\tau)$ at $\tau = \rho T_0$, but strong noise can be attributed to erroneous results due to the presence of spurious peaks obscuring the pitch-peak. In order to overcome the shortcomings of pitch estimation using CAMSF alone, we introduce a pseudo normalized correlation function (PNCF), $\wp(\tau)$, given by,

$$\wp(\tau) = \nabla - \Pi(\tau) \ , \tau \in \left[0 : \frac{Q}{2} \right] \tag{5}$$

where, $\Pi(\tau)$ is the normalized correlation function (NCF) of $y'_w(n)$, expressed as,

$$\Pi(\tau) = \frac{1}{\sqrt{\varpi_0 \varpi_\tau}} \sum_{n=0}^{N-\tau} y'_w(n) y'_w(n+\tau) \tag{6}$$

$$\varpi_\tau = \sum_{n=\tau}^{\tau+N-(\frac{Q}{2}+1)} y'^2_w(n), \ \tau \in \left[0 : \frac{Q}{2} \right] \tag{7}$$

and ∇ represents a constant defined as,

$$\nabla = \frac{\Re_{\max}\left\{ \frac{Q}{2} + 1 \right\}}{\left\{ \frac{Q}{2} + 1 \right\} - \Theta_{\max}} \tag{8}$$

In (8), $\Re_{\max}$ and $\Theta_{\max}$ being constants representing the maximum value of $\Pi(\tau)$ and the argument of $\Re_{\max}$, respectively. The normalized correlation function $\Pi(\tau)$ in (6) is expected to show prominent peaks at $\tau = \rho T_0$. Since, we would like to exploit the notches instead of peaks for pitch estimation, the purpose of the linear transformation performed on $\Pi(\tau)$ using ∇ is to invert the amplitude characteristics of $\Pi(\tau)$. Since peaks of $\Pi(\tau)$ decrease as ρ increases, the resulting pseudo NCF (PNCF) is expected to exhibit a relatively deep front notches compared to the rear ones. Yet in heavy noise, compared to the notch at $\tau = \rho T_0$, there may exist even deeper notch of $\wp(\tau)$ at other lags causing pitch-errors. To this end, we are interested in utilizing the fact that in general, $\Gamma_y(\tau)$ in (2) reaches maxima while $\wp(\tau)$ in (5) approaches minima at integer multiple of T_0. Hence, to emphasize the true pitch-notch, we propose that $\wp(\tau)$ be weighted by the reciprocal of $\Gamma_y(\tau)$ yielding an integrated time-domain function (ITDF) given by,

$$\Im(\tau) = \frac{\wp(\tau)}{\Gamma_y(\tau) + \varepsilon}, \ \tau \in \left[0 : \frac{Q}{2} \right] \tag{9}$$

here, ε is a fixed positive number to prevent division overflow. Furthermore, the noise components included in $\Gamma_y(\tau)$ and $\wp(\tau)$ are uncorrelated and behave independently. Therefore, the proposed ITDF in (9) is able to quell the notches that are non-harmonic of pitch as well as the unwanted noise components. In contrast, it can enhance the desired pitch-harmonic-notches.

However, according to our extensive experimentation, it is found that the above approach of notch enhancement in time-domain can be well exploited to improve the performance of pitch detection to a considerable degree but it cannot completely avoid the pitch-errors, especially in weakly voiced sections of a noisy speech. In what follows, additional steps are taken to elevate the robustness of pitch estimation at a very low SNR.

2.3. Pitch estimation scheme

In this subsection, we intend to determine only one pitch-harmonic (PH) from the spectrum of $y'_w(n)$ and then use it for pitch detection. To begin with, we perform an N-point Discrete Fourier Transform (DFT) on the windowed noisy speech at frame p, namely, $y'_w(n, p)$. It is well known that the DFT of clean speech at the p-th frame exhibits peaks concentrated at or near individual harmonics of pitch frequency. While dealing with a noisy spectrum, it is worth incorporating frequency and frame relative smoothing, the overall effect of which is equivalent to an effective harmonic spectral structure enhancer. Hence, a linear smoother along the frequency axis is applied as,

$$S_f^{y_w}(k, p) = \frac{1}{2M+1} \sum_{m=-M}^{M} S^{y_w}(k+m, p) \tag{10}$$

In (10), M is the length of the smoother. Also, a first-order recursive averaging along the frame axis is employed as,

$$S_s^{y_w}(k, p) = (1 - \alpha_s) S_f^{y_w}(k, p) + \alpha_s S_s^{y_w}(k, p-1) \tag{11}$$

where, α_s ($0<\alpha_s<1$) is a smoothing factor. In (10), $S^{y_w}(k, p)$ represents the k-th component of the spectrum of $y'_w(n, p)$

given by,

$$S^{y_w}(k,p)=\sum_{n=0}^{N-1} y'_w(n,p)\exp(-\frac{j2\pi nk}{N})\,,\ 0\le k\le N-1 \quad (12)$$

As a straightforward and efficient approach, $S^{y_w}(k,p)$ is computed using the FFT command in MATLAB. Finally, $S_s^{y_w}(k,p)$ in (11) implies the *k*-th component of the smoothed noisy speech spectrum at frame *p*. According to (10)-(12), computing all the components of the smoothed noisy spectrum, $\{S_s^{y_w}(0,p), S_s^{y_w}(1,p), \ldots\ldots\ldots\ldots, S_s^{y_w}(N-1,p)\}$, it is ascertained that $S_s^{y_w}(k,p)$ is able to faithfully preserve the peak associated with the harmonic having the highest energy. The frequency point corresponding to the maximum-amplitude coefficient of the $S_s^{y_w}(k,p)$ is chosen as an estimate of a pitch-harmonic denoted by $\hat{\omega}_q$. According to the harmonic sinusoidal speech model described in [6], $\hat{\omega}_q$ denotes the *q*-th harmonic of ω_0 following a relationship,

$$\hat{\omega}_q = q\omega_0,\ \omega_0=\frac{2\pi}{T_0} \quad (13)$$

To estimate the pitch period T_0 using $\hat{\omega}_q$ obtained above, our idea is to determine a proper value of *q* by formulating an impulse-train with a variable-period to match the periodicity of the ITDF $\Im(\tau)$ which is proposed in the previous subsection. First, a submultiple of T_0 denoted as $\hat{T}_q$, is computed from $\hat{\omega}_q$ and then, an impulse-train $I(\tau,\ q)$ is originated as,

$$I(\tau,q)=\sum_{\partial=0}^{\gamma-1}\delta(\tau-\partial T_I),\ \tau\in\left[0:\frac{Q}{2}\right] \quad (14)$$

where, $\hat{T}_q$ is exploited to define the period of $I(\tau,\ q)$ as,

$$T_I = q\hat{T}_q \quad (15)$$

with γ being the number of unit impulses within $I(\tau,\ q)$ and $\delta(\tau)$ represents the Kronecker delta function. In order to find out the optimum value of *q* that leads to T_0, we propose an objective function as given by,

$$\eta(\tau,q)=\sum_{\tau=0}^{\frac{Q}{2}} I(\tau,q)\Im(\tau),\ \ \tau\in\left[0:\frac{Q}{2}\right] \quad (16)$$

It is well demonstrated that $\Im(\tau)$ in (9) is expected to show deep notches at $\tau=\rho T_0$. Hence, the objective function $\eta(\tau,\ q)$ is minimized when an appropriate value of *q* is chosen such that T_I is synchronized with T_0. The value of *q* corresponding to the minimum of $\eta(\tau,\ q)$, denoted as q_{opt}, is used to determine the desired estimate of the pitch period $\hat{T}_0$ and consequently the pitch frequency ($\hat{F}_0$) of a voiced frame is estimated as

$$\hat{F}_0=\frac{F_s}{\hat{T}_0}\ ,\quad \hat{T}_0 = q_{opt}\hat{T}_q \quad (17)$$

where, F_s is the sampling frequency in Hz.

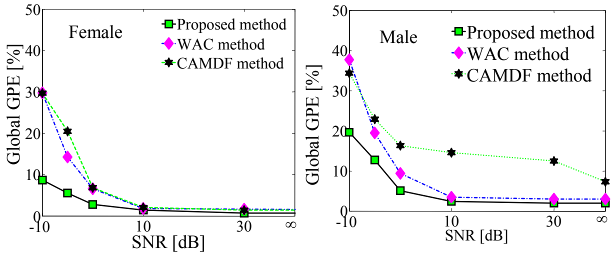

Fig. 1. Global GPE [%] as a function of SNR for female and male speakers.

3. RESULTS AND PERFORMANCE COMPARISON

The performance of the proposed algorithm is evaluated using the *Keele* pitch extraction reference database [8]. The core data consists of a phonetically balanced text, "*The North Wind Story"* of about 35 seconds, read by 5 mature male and 5 mature female speakers. The *Keele* database is studio quality, sampled at 20 kHz with 16-bit resolution. As a "ground" truth, this database provides a reference pitch obtained from a simultaneously recorded laryngograph trace. The pitch values are provided at a frame rate of 100 Hz with 25.6 *ms* window (*N*). In order to use this database for performance evaluation, the same analysis parameters (frame rate and window size) are chosen in the proposed algorithm. For windowing operation, we have used a normalized hamming window. In (3), $\tau_a<N$ is taken into account to cover the wide range of pitch for both male and female speakers. Note that ε in (9) is set to 1 as in [7]. The values of *M* and α_s in (10) and (11) are chosen as 1 and 0.6, respectively. Simulations are performed for both clean speech (indicated as ∞ dB SNR) and white noise-corrupted speech with an SNR varying from –10 dB to 30 dB.

We have evaluated the estimated pitch values only for clearly voiced frames based on the voiced/unvoiced labels included in the *Keele* database. According to Rabiner in [2], the gross pitch- error (GPE) is measured as the percentage of the pitch period estimation errors that are greater than 1 *ms* in their absolute values, otherwise, the error is termed as the fine pitch-error (FPE) measured by its mean (m_{FPE}) and the standard deviation (σ_{FPE}). Root-mean-square-error (RMSE) is also used in this paper to quantify the pitch detection accuracy. For a speaker group, the "global" error is calculated considering all five male (or all five female) speakers. From Fig. 1 and Fig. 2, it is evident that in comparison to CAMDF [5] and WAC [7] methods, the global GPE [%] and the global RMSE [Hz] of the proposed algorithm is significantly superior for both female and male speakers at almost all SNR levels ranging from –10 dB to 30 dB and the efficacy is also better for clean speech. It is noticeable from Table I. that for the female as well as the male speakers, not only at a high SNR but also at a very low SNR, the global m_{FPE} [Hz] and the global σ_{FPE} [Hz] (in brackets) of the proposed method are, within an acceptable limit and consistently competitive relative to the methods

Table I. Performance comparison of different methods in terms of global m_{FPE} [Hz] and global σ_{FPE} [Hz] (in brackets).

Method	30 dB		-5 dB	
	Male	Female	Male	Female
Proposed method	2.29 (2.15)	4.34 (4.56)	2.80 (2.83)	6.60 (6.47)
WAC method	2.30 (2.37)	5.24 (5.31)	2.86 (2.85)	6.82 (7.12)
CAMDF method	2.35 (2.38)	4.99 (5.03)	2.88 (2.99)	6.72 (6.82)

mentioned above. For a male speaker, Fig. 3 shows a reference pitch contour accompanied by the spectrogram of clean speech that corresponds to an excerpt of 3 s from the reference database. Also, pitch contours overlaid on the spectrograms of the noisy speech are portrayed for different methods. It is depicted that compared to the other methods, the pitch contour resulting from the proposed algorithm is comparatively smooth even at an SNR as low as –10 dB. This implies that if the precision is not a major concern, we can avoid the post-processing of pitch contour to reduce the delay for real-time applications. Research is in progress to evaluate the performance of the proposed algorithm in various colored noise environments.

4. CONCLUSION

In this paper, a robust frame work to estimate pitch in the presence of noise is addressed. The kernel of this approach lies in introducing two novel functions, namely, the CAMSF and the PNCF that are combined for enhancing pitch-harmonic notches in time-domain. To guarantee a robust pitch extraction from noisy speech, the algorithm further exploits a pitch-harmonic (PH) that is estimated accurately from the smoothed noisy speech spectrum thus assists in overcoming the double and half-pitch problem of the conventional methods. Through simulation results it has been exposed that the proposed algorithm significantly outperforms some of the reported pitch detection methods from high to very low SNR levels.

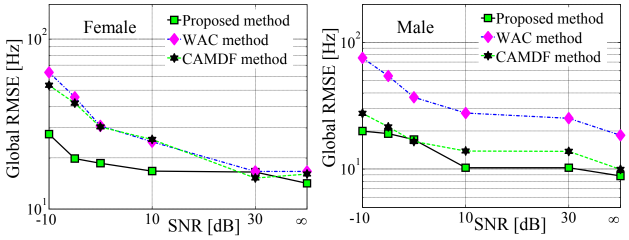

Fig. 2. Global RMSE [Hz] as a function of SNR for female and male speakers.

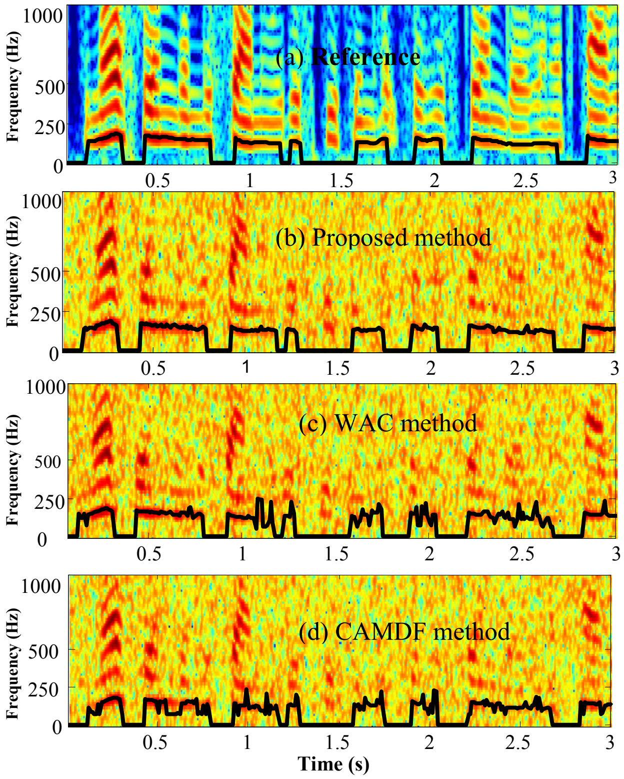

Fig. 3. Comparison of pitch contours at –10 dB

5. REFERENCES

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

[2] 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.

[3] O. Deshmukh, J. Singh, C. E.-Wilson, "Use of temporal information: detection of periodicity, aperiodicity, and pitch in speech," *IEEE Trans. Speech and Audio Processing*, vol. 13, pp. 776-786, 2005.

[4] L. Hui, B.-q. Dai, and L. Wei, "A pitch detection algorithm based on AMDF and ACF," in *Proc. ICASSP2006*, Toulouse, France, pp.377-380, May 2006.

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

[6] C. Shahnaz, W. -P Zhu, and M. O. Ahmad, "Robust pitch estimation at very low SNR exploting time and frequency domain cues," in *Proc. ICASSP2005*, Philadelphia, USA, pp. 389-392, March 2005.

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

[8] G. Meyer, F Plante and W. A. Ainsworth," A pitch extraction reference database," *EUROSPEECH'95*, Madrid, pp. 827-840, 1995.