ROBUST PITCH ESTIMATION AT VERY LOW SNR EXPLOITING TIME AND FREQUENCY DOMAIN CUES

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 joint time and frequency domain approach for pitch estimation of speech at a very low signal-to-noise ratio (SNR). The kernel of this approach lies in introducing a new function for detecting the time-domain cue by modifying the circular average magnitude difference function (CAMDF). By using the new function in conjunction with the half-wave rectified version of the autocorrelation function, the pitch-peak can be emphasized and the non-pitch peaks suppressed. The proposed method not only alleviates the effect of falling valleys in the conventional average magnitude difference function (AMDF) but also overcomes the difficulty of the basic CAMDF technique in estimating a pitch period of larger than one-half of the frame length. To guarantee a robust pitch detection in noisy speech, *a priori* frequency-domain estimate of the dominant pitch-harmonic (DH) is extracted as an additional cue and is utilized to optimally match the pitch-peak in time-domain. The proposed approach is simulated using the *Keele* reference database. It is shown that the proposed method using a joint time and frequency domain cues is able to give a superior accuracy relative to some of the existing methods even at a very low SNR of –10 dB.

1. INTRODUCTION

Pitch period or fundamental frequency of speech is the primary acoustic cue for sentence intonation and lexical stress and therefore, is crucial to phoneme identification in tonal languages. Pitch related prosodic features have axiomatic importance in automatic speech recognition and understanding (ASRU) systems. In most low-rate voice coders, accurate pitch estimation is a pre-requisite for reconstructing a good-quality speech. Some medium rate coders use pitch to reduce the transmission rate without degrading the speech quality. Pitch patterns also have extensive applications in speech articulation training for the deaf.

Reliability and accuracy have been the major focus of many pitch detection methods for noisy speech. The autocorrelation function (ACF) based methods [1] exhibit a better performance against noise for female's voice relative to male's. The average magnitude difference function (AMDF) based approaches [2] suffer from a serious pitch doubling problem in noisy environments. Since the minimum valleys of AMDF decrease, the weighted autocorrelation (WAC) method [3] using the inverse AMDF fails to suppress the erroneous peaks at a very low SNR. Even though the circular AMDF (CAMDF) [4] overcomes the defects of AMDF, it cannot be applied for speeches with pitch period larger than one-half of the frame length. Also, its performance for noisy speech is found unsatisfactory. The shortcomings of the AMDF technique are also inherent to the recently reported sinusoidal ACF method [5] as it maximizes a cost function involving AMDF to find the pitch of a noisy speech.

In this paper, we present an effective time and frequency domain pitch estimation method for noisy speech. The proposed method exploits both time and frequency-domain cues to estimate the pitch period, in which the time-domain cue is acquired by combining a new function obtained from the modified CAMDF with the half-wave rectified ACF while the frequency-domain cue is obtained by estimating the dominant pitch-harmonic from the cosine-modeled ACF. In attaining the time-domain cue, an adaptive weighting window is also employed to reduce the possibility of half- and double-pitch errors. The proposed strategy of integrating the time and frequency-domain information results in a very accurate pitch detection from heavily noise-corrupted speech signals.

2. PROBLEM FORMULATION

The ACF of noisy speech $y(n) = x(n) + v(n)$ can be written as

$$\phi_{yy}(\tau) = \begin{cases} \phi_{xx}(\tau) + \sigma_v^2, & \tau = 0 \\ \phi_{xx}(\tau), & \tau \neq 0 \end{cases} \quad (1)$$

where $\phi_{xx}(\tau)$ is the ACF of the clean speech $x(n)$ as given by

$$\phi_{xx}(\tau) = \frac{1}{N} \sum_{n=0}^{N-1-|\tau|} x(n)x(n+|\tau|) \quad (2)$$

and σ_v^2 represents the variance of the zero-mean white Gaussian noise $v(n)$. In (2), N is the speech frame size used. If $x(n)$ is periodic with period T, and $x(n)$ and $v(n)$ are truly uncorrelated, (1) will give a robust pitch estimation at the second maximum peak of $\phi_{yy}(\tau)$ corresponding to $\tau = T$. However, $x(n)$ and $v(n)$ may not be strictly uncorrelated in practice. Therefore, a variation of correlation analysis, AMDF of $y(n)$ given by

$$\psi(\tau) = \frac{1}{N} \sum_{n=0}^{N-1-|\tau|} |y(n) - y(n+|\tau|)| \quad (3)$$

has been proposed for pitch estimation in the presence of noise [2]. Note that the number of items summed in $\psi(\tau)$ reduces with the increase of the lag τ. The global minimum or valley of $\psi(\tau)$ often occurs at a higher multiple of T. Since $\psi(\tau)$ produces a valley while $\phi_{yy}(\tau)$ provides a peak, it has been suggested in [3] that $\phi_{yy}(\tau)$ be weighted by the reciprocal of $\psi(\tau)$ in order to emphasize the peaks at the speech periodicity T. Due to the falling trend of the valleys of $\psi(\tau)$, those peaks other than the pitch-peak become more prominent. This detrimental effect misguides pitch

estimation especially at a low SNR. As such, the circular AMDF (CAMDF) of $y(n)$ that is given by

$$\psi'(\tau) = \sum_{n=0}^{N-1} \left| y(\mathrm{mod}(n+\tau, N)) - y(n) \right| \quad (4)$$

has been proposed in [4] to improve the basic AMDF. Although the CAMDF is expected to provide a minimum valley at T, it gives a function that is unique up to one-half of the frame size. Hence, pitch period larger than $N/2$ samples cannot be estimated using an N-sample frame.

In the next section, we will propose a new pitch estimation approach that prevents the falling phenomenon of valleys as well as overcomes the aforementioned deficiency of CAMDF technique. The key to the proposed method is to develop a function for the detection of time-domain cue by using an improved CAMDF along with the half-wave rectified ACF. To minimize the pitch estimation error in noise, an additional frequency-domain cue obtained from the DH of the cosine model of ACF is combined with the time-domain cue. A pre-estimate of the DH is extracted from FFT domain.

3. PROPOSED METHOD

3.1 Pre-processing

The noisy speech, y(n), is first converted to frequency domain by using the discrete cosine transform (DCT). The DCT coefficients corresponding to the first formant range for most male and female speakers are then retained and converted back using the inverse DCT to time-domain signal. Thus, the pre-filtered noisy speech (PFNS), denoted as y′(n), is considered to be cleaner than y(n), since the influence of other formants on pitch estimation has been removed [6].

3.2 Signal-conditioning for time-domain pitch-cue

Let $y'_W(n)$ be the rectangular-windowed PFNS signal. Analyzing the characteristics of $\psi(\tau)$ given by (3), the basic window size (N) is added by $\tau_{\max}$ up to $N + \tau_{\max}$ to define a modified CAMDF (MCAMDF) where $\tau_{\max}$ is the preset maximum pitch-lag for female or male speakers. The MCAMDF is proposed as

$$\xi(\tau) = \sum_{n=0}^{\beta} \left| y'_W(m) - y'_W(n) \right|, \ \tau = 0,1,2,.....,\beta \quad (5)$$

here, m = mod $(n + \tau, N + \tau_{\max})$ representing the modulo operation of $(n + \tau)$ with respect to $(N + \tau_{\max})$, and $\beta = N + \tau_{\max} - 1$. $\xi(\tau)$ has the symmetry property around $\tau_s = (\beta + 1)/2$. Thus, we need to calculate $\xi(\tau)$ only with $\tau \in [0, \tau_s]$. Similar to $\psi(\tau)$, $\xi(\tau)$ is expected to maintain deep valleys at $\tau = \rho T$ with $0 \le \rho T \le \tau_s$ (ρ = 0, 1, 2….). In contrast with $\psi(\tau)$, $\xi(\tau)$ depends actually on two different speech frames. For each value of τ, all the speech samples in consecutive two frames are used twice to compute (5). Since the number of terms summed is a constant, i.e., $\beta+1$,at all lags, peaks of $\xi(\tau)$ remain in the same level preventing its valleys from falling. As $N > \tau_s > N/2$ and $\tau_s > \tau_{\max}$, $\xi(\tau)$ also overcomes the constraint of half-frame ($N/2$) symmetric property of $\psi(\tau)$ in (4). Since pitch usually changes slowly with time in a voiced section of the speech [6], using a larger window for $\xi(\tau)$ will not potentially increase the pitch estimation error.

We want to exploit the peaks instead of valleys for pitch estimation. Hence we define a new signal-conditioned function as

$$\chi(\tau) = \frac{\Im_{\max} . N}{N - \Omega_{\max}} - \xi(\tau), \quad (\tau = 0,1,2,\ldots,\tau_s) \quad (6)$$

where $\Im_{\max} = \max\{\xi(\tau)\}$ for $0 < \tau \le \tau_s$ and $\Omega_{\max} \le \tau_s$ is the index of $\Im_{\max}$. Note that the above defined reshaped MCAMDF, $\chi(\tau)$, is always positive but it reverses the peaks and valleys of $\xi(\tau)$. In the range of $0 \le \tau \le \tau_s$, the deepness of the valleys of $\xi(\tau)$ decreases with the increasing ρ. Hence, $\chi(\tau)$ has a relatively sharp front peaks compared to the rear ones.

We have found that $\chi(\tau)$ and the ACF of basic windowed PFNS, $\phi_{yy}(\tau)$, up to $\tau = \tau_s$ have very similar characteristics and share the same periodicity in noise–free case. In a noisy environment, we propose to combine $\chi(\tau)$ with a half-wave rectified $\phi_{yy}(\tau)$ denoted as $\Phi(\tau)$ in order to enhance the pitch-peak. To this end, $\Phi(\tau)$ and $\chi(\tau)$ are each first raised by a power of three, represented by $\Phi'(\tau)$ and $\chi'(\tau)$, respectively, which are then multiplied giving

$$\eta(\tau) = \Phi'(\tau)\chi'(\tau) \quad (7)$$

This power-of-three and product operation is able to reinforce the larger peaks reasonably. Since, the noise components included in $\chi'(\tau)$ and $\Phi'(\tau)$, are uncorrelated and behave independently, after the nonlinear computation, the desired pitch-harmonic-peaks can be enhanced and the non-harmonic peaks as well as the unwanted noise component suppressed. In order to reduce the possibility of half and double-pitch errors, we further modify $\eta(\tau)$ as

$$\eta'(\tau) = \eta(\tau) w(\tau) \quad (8)$$

where $w(\tau) = w_l(\tau) w_a(\tau)$ represents a two-wing window. The first part $w_l(\tau)$ is a small-delay weighting function that favors half-pitch lags for males, as given by

$$w_l(\tau) = \tau^{\log_2 \wp} \quad (9)$$

where $\wp = 0.85$ is a tuning parameter. The second part $w_a(\tau)$ of the two-wing is defined as

$$w_a(\tau) = (\left| \nabla_{old} - \tau \right| + p_L)^{\log_2 \wp_a} \quad (10)$$

where ∇_{old} represents the median-filtered pitch-lag of the previous voiced speech frames and p_L the lower pitch search limit. The weighting function $w_a(\tau)$ is centered around ∇_{old} that is updated only for strongly voiced frames. The weighting function depends also on the number (∂) of consecutive unvoiced speech frames before the current frame as determined by the value of $\wp_a$, namely, $\wp_a = 0.85$ if $\partial < 6$, $\wp_a = 0.95$ if $6 \le \partial \le 10$, or $\wp_a = 0.99$ if $\partial > 10$. Using this scheme, $w_a(\tau)$ is attenuated after unvoiced or silence periods and therefore, the estimated pitch is not biased towards the old-pitch.

The above-proposed time-domain method is able to improve the conventional pitch detection techniques to a considerable degree. However, it cannot completely avoid the double and half-pitch problem, especially in weakly voiced sections of a noisy speech. In what follows, we will make use of the frequency-domain information to enhance the robustness of pitch estimation.

3.3 A priori cue: for estimation of dominant pitch-harmonic

The harmonic sinusoidal speech model for a strongly-voiced frame of clean-speech, $x(n)$, can be represented by

$$x(n) = \sum_{k=1}^{r} b_k \exp\left[j(n\omega_k + \theta_k) \right], \quad \omega_k = k\omega_0 \quad (11)$$

where b_k is the envelope of the vocal tract and ω_k the k-th harmonic of the pitch frequency $\omega_0 = 2\pi f_0 / f_s$ in radians/sec with f_0 and f_s being the pitch frequency and the sampling frequency,

respectively. From (11), a cosine model (CM) for ACF of $x(n)$ can be derived as

$$R_{xx}(\tau) = \sum_{k=1}^{r} \gamma_k \cos(\omega_k \tau), \quad \tau \geq 0,\ \omega_k = k\omega_0 \tag{12}$$

where, γ_k is a constant ($\gamma_k = c/2$, $c = |b_k|^2$). The power spectrum of $x(n)$ is expressed as

$$P_{xx}(\omega) = \sum_{k=1}^{r} 2\pi(\frac{\gamma_k}{2})[\delta(\omega + k\omega_0) + \delta(\omega - k\omega_0)] \tag{13}$$

This discrete line spectrum can be viewed as sampling of a spectral envelope corresponding to the transfer function of the vocal tract, with the sampling points governed by the pitch of the speaker. Hence, unlike the conventional sinusoidal model-based approaches that calculate all the harmonics, we intend to determine the dominant harmonic (DH). Instead of using $\phi_{yy}'(\tau)$ directly, a close approximation of $R_{xx}(\tau)$ is computed through one constituent function of the CM model, $R_1(\tau)$. The total squared error $\Re^{(i)}(\omega_1)$ between $\phi_{yy}'(\tau)$ and $R_1^{(i)}(\tau)$ can be written as

$$\Re^{(i)}(\omega_1) = \sum_{\tau=1}^{\tau_s} \left| \phi_{yy}'(\tau) - R_1^{(i)}(\tau) \right|^2 \tag{14}$$

where, $R_1^{(i)}(\tau) = \gamma_1^{(i)} \cos(\omega_1^{(i)} \tau)$ and the iteration index '(i)' depends on the number of ω_1's used in (14) for optimization. The FFT coefficient (FC) of PFNS in the frame under analysis which corresponds to the largest energy provides the best probable cue for a pitch-harmonic. Thus, all the ω's in the entire pitch domain are not encouraged to use as ω_1. It is more appropriate to perform a search only in the neighborhood of *a priori* estimate of ω that corresponds to the FC having maximum amplitude in the FFT domain. Then the optimum $\omega_1 = \omega_1^{(i)}$ can be found at the i-th iteration for which $\Re^{(i)}(\omega_1)$ is minimized in the least-square sense. Due to this indirect noise compensation from all lags of $\phi_{yy}'(\tau)$, a quite accurate estimation of the DH ($f_d = \omega_1/2\pi$) can be obtained from the best matched $R_1(\tau)$.

3.4 Combination of time-frequency domain cues

Since $\omega_1 = k\omega_0$, DH (f_d) can be reasonably used to estimate the pitch period, T. According to the autoregressive model of speech production, the driving function to the vocal tract is a quasi-periodic impulse sequence for voiced sound. Hence we are motivated to use an impulse train $I(n, \alpha)$ [5] of length τ_s containing a fixed number (λ) of unit impulses to weight the function, $\eta'(\tau)$, proposed in (8). The period of the impulse train (T_i) is varied iteratively where $T_i = \alpha T_d$, $T_d = 1/f_d$, α is a nonzero positive scaling integer. For a particular value of α, $I(n, \alpha)$ is expressed as

$$I(n,\alpha) = \sum_{\mu=0}^{\lambda-1} \delta(n - \mu\alpha T_d), \quad n = 0,1,2,.....,\tau_s \tag{15}$$

where, $\delta(n)$ is the Kronecker delta function. Finally, the inner product of $\mathbf{I}(\alpha)$ and $\boldsymbol{\eta}'$ is taken as a target function,

$$\zeta(\alpha) = \mathbf{I}(\alpha)(\boldsymbol{\eta}')^{\mathrm{T}} \tag{16}$$

where $\boldsymbol{\eta}' = [\eta'(0)\ \ \eta'(1)\ \ \eta'(2)\ \ \ldots\ \ \eta'(\tau_s - 1)]$ and $\mathbf{I}(\alpha) = [\, I(0, \alpha)\ \ I(1, \alpha)\ \ \ldots\ \ I(\tau_s - 1, \alpha)]$. The target function $\zeta(\alpha)$ is maximized by varying α. Since, $\eta'(\tau)$ is expected to exhibit strong peaks at integer multiples of T, $\zeta(\alpha)$ will be maximized when T_i closely matches with T. The value of α corresponding to the maximum of $\zeta(\alpha)$, denoted by α_m, is used to estimate the desired pitch period as $T_{\text{est}} = \alpha_m T_d$. f_0 is calculated by inverse operation of T_{est}.

4. RESULTS AND PERFORMANCE COMPARISON

The performance of the proposed method is evaluated using the *Keele* reference database [7]. This database provides a reference pitch obtained from a simultaneously recorded laryngograph trace as "ground" truth. The pitch values are provided at a frame rate of 100 Hz with 25.6 ms window. Unvoiced frames are assigned with zero pitch values, and uncertain frames are filled with negative values. The data sequence 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. In order to use this database for performance evaluation, the same analysis parameters (frame rate and basic window size) are chosen in the proposed method. For each voiced frame, DH was searched with a scanning resolution of 0.001. The number of unit impulses ($\lambda \geq 2$) in the impulse train must be kept constant for a particular speaker depending on the value of $\tau_{\max}$. The parameter α in (15) is restricted by $(\lambda - 1)\alpha T_d \leq (\tau_s - 1)$. Simulations are performed for both clean speech (very large SNR) and noisy speech with an SNR varying from –10 dB to 30 dB at a 5 dB increase.

Similar to [8], only the "clearly voiced" reference frames in the *Keele* database are used for performance evaluation. According to Rabiner [1], the gross error rate (GER) 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) of the voiced region is also used in this paper to quantify the pitch detection accuracy. The "global" error is calculated considering all five male (or all five female) speakers.

As summarized in Table 1, the proposed method performs equivalently well for clean speech in comparison to the methods in [8] using the same *Keele* database. From Fig. 1, it is seen that the global GER [%] and RMSE [Hz] of the proposed algorithm is significantly superior for both female and male speakers at all SNR levels ranging from –10 dB to 30 dB and in clean speech (indicated as ∞ dB SNR) in comparison to original AMDF [2], WAC [3] as well as the basic CAMDF [4] methods. It is observed from Table 2 that for both the female and male speakers, the global m_{FPE} [Hz] and σ_{FPE} [Hz] of the proposed method are, within an acceptable limit, consistently competitive relative to the methods mentioned above both at –10 dB and 10 dB**.** Fig 2. shows a comparatively smooth pitch contour of a male utterance obtained from the proposed algorithm when SNR = –10 dB. The double and half-pitch errors have been significantly reduced as a result of using an adaptive weighting window in the proposed pitch detection method. This implies that if the precision is not a major concern, the post-processing of pitch contours using a 5-tap median filter as required in the conventional pitch estimation methods can be avoided so as to reduce the delay for real-time applications.

Table 1. Performance comparison of proposed method versus X_{waves} [8] and DLFT [8] in clean speech.

Method	GER [%]	m_{FPE} [Hz]	σ_{FPE} [Hz]	RMSE [Hz]
Xwaves	1.74	3.81	15.52	15.98
DLFT	3.24	4.61	15.58	16.25
Proposed (Female)	0.68	4.26	4.37	14.15
Proposed (Male)	1.99	2.29	2.58	8.85
Proposed (overall)	1.32	3.30	3.76	11.86

Table 2. Performance comparison of different methods in terms of global m_{FPE} [Hz] and global σ_{FPE} [Hz] (in brackets).

Method	–10 dB		10 dB	
	Male	Female	Male	Female
Proposed	3.07 (3.14)	6.23 (7.26)	2.16 (2.15)	4.34 (4.56)
WAC	3.31 (3.41)	8.32 (8.85)	2.22 (2.20)	5.01 (4.48)
CAMDF	3.37 (3.23)	7.71 (7.86)	2.33 (2.19)	4.94 (4.61)
AMDF	2.28 (1.93)	6.48 (5.94)	2.02 (1.87)	5.03 (4.25)

5. CONCLUSION

In this paper, we have proposed a new pitch detection algorithm using both time-domain and frequency-domain cues. A pitch detection function has been developed by combining the modified version of CAMDF and the half-wave rectified ACF in order to enhance the pitch- peak and suppress the non-pitch peaks. The new time-domain function is also able to estimate a larger pitch period compared to the existing CAMDF. A pre-estimate of the frequency-domain pitch cue has also been used to overcome the double and half-pitch problem of the conventional methods and to improve the robustness of the proposed pitch detection algorithm in noisy speech. Simulation results have shown that the proposed method significantly outperforms the recently reported correlation and AMDF based approaches at a very low SNR level.

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] 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.

[3] 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.

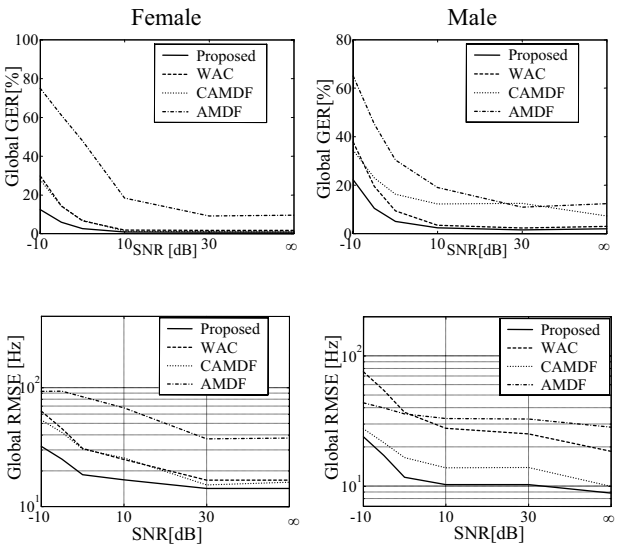

Fig. 1. Global GER [%] and global RMSE [Hz] as a function of SNR for female and male speakers.

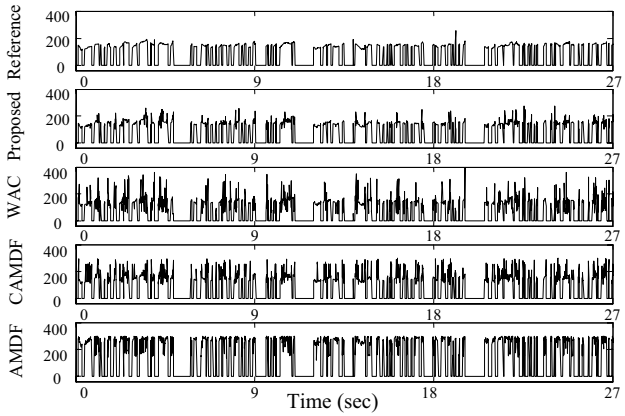

Fig. 2. Comparison of pitch contours at –10 dB

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

[5] M. K. Hasan, C. Shahnaz, and S. A. Fattah, "*Determination* of Pitch of Noisy Speech Using Dominant Harmonic Frequency," in *Proc. ISCAS2003*, Bangkok, Thailand pp. 556-559, May 2003.

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

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

[8] C. Wang and S. Seneff, "Robust pitch tracking for prosodic modeling in telephone speech," in *Porc. ICASSP2000*, Istanbul, Turkey, pp. 1343-1346, June 2000.