A PITCH EXTRACTION ALGORITHM IN NOISE BASED ON TEMPORAL AND SPECTRAL REPRESENTATIONS

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
(c_shahna, weiping, omair)@ece.concordia.ca

ABSTRACT

In this paper, a new algorithm for pitch extraction from noisy speech signals based on both temporal and spectral representations is presented. We derive a harmonic sinusoidal correlation (HSC) model of clean speech as a temporal representation. Given only a noisy speech frame, a noise-robust least-squares minimization technique is proposed to acquire the parameters of the HSC model which are directly employed for the accurate estimation of a pitch-harmonic (PH). Exploiting the extracted PH and based on a spectral representation which is an enhanced spectrum in the Discrete Cosine Transform domain, a two-fold criterion is developed in order to achieve the true consecutive number corresponding to PH that is finally adopted for pitch detection in the presence of noise. Simulation results using the *Keele* pitch extraction reference database manifest that combining the multi cues obtained from the temporal as well as spectral representations, the proposed algorithm is able to achieve a superior efficacy in comparison to some of the existing methods from high to very low signal-to-noise ratio (SNR) levels.

Index terms– Pitch extraction, harmonic sinusoidal correlation model, pitch-harmonic, Discrete Cosine Transform, low SNR.

1. INTRODUCTION

Pitch is an axiomatically important parameter for automatic speech recognition and understanding (ASRU), compression, and synthesis. It plays an eminent role in both the production and perception of speech. In application areas such as speech enhancement using the harmonic model, analysis and modeling of speech prosody, low-bit rate speech coding, and speaker recognition, the signal processing algorithms require accurate pitch for their reliable performance [1].

The pitch detection algorithms (PDAs) face complication in the presence of noise. Among several reported PDAs, the methods based on the autocorrelation function (ACF) [2] exhibit errors due to the sensitiveness of ACF to the spectral envelope. Accuracy has eluded the approaches [2]-[3] based on the average magnitude difference function (AMDF) owing to the falling trend of the minima of AMDF. Moreover, the magnitude of principal minimum of AMDF is highly influenced by intensity variation and the background noise of speech. The PDA based on AMDF and ACF [4] is shown to be accurate in clean speech and tested only for a limited number of frames. The circular average magnitude difference function (CAMDF) [5] is reported as advantageous over AMDF to some extent but it cannot be applied for speakers with pitch period larger than one-half of the frame size. The weighted autocorrelation (WAC) method [6] employing AMDF is found to emphasize the non-pitch peaks especially at a very low SNR. Recently, Celia *et al*. have proposed some PDAs [7]-[8] robust to white noise, whereas most of the PDAs are proposed only for clean speech.

In this work, we propose another robust frame work for pitch estimation of speech severely corrupted by a white noise. The kernel of this approach lies in introducing a harmonic sinusoidal correlation (HSC) model of clean speech from which a pitch-harmonic (PH) is estimated accurately in noise employing a least-squares (LS) minimization process. Then a dual criterion in the Discrete Cosine Transform (DCT) domain is proposed to resolve the harmonic number corresponding to PH which is required to extract pitch information. We argued that due to the collaboration of cues from the temporal and spectral representations, the proposed algorithm can ensure accurate pitch estimation even in an intricate noisy environment.

2. PROPOSED METHOD

Assuming that the clean speech $x(n)$ is contaminated by an additive noise $v(n)$, the noisy speech $y(n)$ can be written as,

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

$y(n)$ is segmented into frames with frame-size N by the application of a window function $w(n)$. As a preprocessing, the windowed noisy frame is filtered using a low-pass filter (LPF) to retain only the first formant (e.g., the 0-900 Hz range). The time domain pre-processed noisy speech frame is denoted as,

$$y_w(n) = x_w(n) + v_w(n) \tag{2}$$

This process removes the influence of higher formants which confound the pitch (F_0) estimation.

2.1 Pitch-harmonic extraction

A voiced frame of the preprocessed clean speech $x_w(n)$ can be assumed to be represented by a set of sinusoidal waveforms for which all the frequencies are harmonically related as,

$$x_w(n) = \sum_{p=1}^{K} b(\omega_p)\exp[j(n\omega_p + \phi_p)], \quad \omega_p = p\omega_0 \tag{3}$$

In this harmonic speech model (HSM), K is the number of harmonics in the speech bandwidth. The parameters $b(\omega_p)$, ω_p, ϕ_p stand for, respectively, the envelope of the vocal tract, the angular frequency and phase of the p-th harmonic of $\omega_0 = 2\pi F_0/F_s$ with F_0 being the pitch frequency and F_s the sampling frequency in Hz. Again, $x_w(n)$ in (3) can be considered as the output of an autoregressive (AR) system with order $q = 2K$ and given by,

$$\hat{x}_w(n) = -\sum_{j=1}^{q} a_j \hat{x}_w(n-j) + G\delta(n) \tag{4}$$

where, roots of the parameters a_j correspond to the poles of the AR system which is excited by a Kronecker delta function $\delta(n)$ with gain G. This marginal AR system has all of its poles located on the unit circle, namely, the p-th pole is given by,

$$p_p = |p_p|\exp(j\omega_p) = \exp(j\omega_p), \quad |p_p| = 1 \tag{5}$$

In (5), ω_p denotes the angular position of the p-th pole which in turn represents a pitch-harmonic (PH) of the HSM. It is to be mentioned that the periodicity of $x_w(n)$ in (3) is usually exhibited by its ACF which is regarded as a temporal representation and can be estimated as,

$$r_{x_w}(\tau) = \frac{1}{N}\sum_{n=0}^{N-1-|\tau|} x_w(n)x_w(n+|\tau|) \tag{6}$$

Note that, the response of the AR system $\hat{x}_w(n)$ closely resembles $x_w(n)$, though they are not equal. Nevertheless, the AR model for $\hat{x}_w(n)$ in (4) is adequate to represent the periodicity of $x_w(n)$. Hence, we like to use AR model based $\hat{x}_w(n)$ to determine a PH of the HSM to be utilized for pitch estimation. Since $\hat{x}_w(n)$ is real, considering the effect of complex poles, the ACF of $\hat{x}_w(n)$ can be derived and simplified as,

$$r_{\hat{x}\hat{x}}(\tau) = \sum_{p=1}^{\frac{q}{2}} \left[\alpha_p \cos(\omega_p\tau) + \beta_p \sin(\omega_p\tau)\right], \ \tau \in [0:M] \tag{7}$$

where, α_p and β_p are constants and (M +1) refers to the number of lags of the harmonic sinusoidal correlation (HSC) model of clean speech. Unlike conventional approaches, as we intend to determine only one ω_p, it can be directly obtained from the parameters of one component function of the HSC model as expressed by,

$$R_p(\tau) = \alpha_p \cos(\omega_p\tau) + \beta_p \sin(\omega_p\tau) \tag{8}$$

For this purpose, given the ACF for a voiced frame of the preprocessed noisy speech $y_w(n)$, which can be estimated as,

$$r_{y_w}(\tau) = \frac{1}{N}\sum_{n=0}^{N-1-|\tau|} y_w(n)y_w(n+|\tau|) \tag{9}$$

we can formulate a fitting problem that minimizes the total squares error between $r_{y_w}(\tau)$ and $R_p(\tau)$ as given by,

$$\Theta(\omega_p^{(i)}) = \sum_{\tau=1}^{M} \left| r_{y_w}(\tau) - R_p^{(i)}(\tau) \right|^2 \tag{10}$$

where, the superscript 'i' represents the index of the chosen ω_p. For each chosen $\omega_p^{(i)}$, the corresponding values of $\alpha_p^{(i)}$ and $\beta_p^{(i)}$ can be determined uniquely by minimizing $\Theta(\omega_p^{(i)})$ in the least-squares (LS) sense. Therefore, for a particular set of $\{\omega_p^{(i)}, \alpha_p^{(i)}, \beta_p^{(i)}\}$, when $R_p^{(i)}(\tau)$ best matches $r_{y_w}(\tau)$, the global minimum of $\Theta(\omega_p^{(i)})$ is reached and the optimum solution for ω_p is obtained as,

$$\omega_{popt} = \arg\min_{\omega_p^{(i)}} [\Theta(\omega_p^{(i)})] \tag{11}$$

Exploiting a PH, ω_{popt}, extracted accurately even at a very low SNR, pitch is estimated in the next subsection based on a spectral representation.

2.2 Pitch extraction

In this subsection, a smoothed noisy power spectrum in the Discrete Cosine Transform (DCT) domain is considered as a spectral representation to be used for pitch extraction in conjunction with ω_{popt}. For an input vector $\{y_w(1), y_w(2),\ldots, y_w(N)\}$, the DCT output vector $\{Y_w(1), Y_w(2),\ldots\ldots, Y_w(N)\}$ is given by the relation,

$$Y_w(k) = c(k)\sum_{n=1}^{N} y_w(n)\cos\left[\frac{\pi(2n-1)(k-1)}{2N}\right], k \in \{1,2,\ldots\ldots,N\} \tag{12}$$

It is known that DCT of $x_w(n)$ exhibits peaks concentrated at or near individual harmonics of pitch. While dealing with $y_w(n)$, in order to effectively enhance the harmonic spectral structure in noise, it is worth incorporating both frequency and frame relative smoothing [8]. Since, the estimated *p-th* harmonic ω_{popt} is related to ω_0 by,

$$\omega_{popt} = p\omega_0 \tag{13}$$

the key task is to resolve a true consecutive number p_{opt} corresponding to ω_{popt}. For this purpose, our idea is to develop a two-fold criterion, at first, by formulating an objective measure and then, computing a threshold for selecting a set of harmonic numbers to be used for maximizing the objective measure, wherein a smoothed and enhanced DCT power spectrum of $y_w(n)$ denoted as,

$$\Gamma(f) = \left|Y_w^s(k)\right|^2 \tag{14}$$

is employed as $|Y_w^s(k)|$ is capable of faithfully preserving the peak associated with the pitch harmonics. The spectral representation $\Gamma(f)$ is utilized to put forward the dual criterion in relation to $[f_{popt}/p]$, where, the optimum PH frequency f_{popt} is given by,

$$f_{popt} = \left\lfloor \frac{\omega_{popt}}{2\pi} \right\rfloor \tag{15}$$

Expressing the optimum PH frequency f_{popt}, the maximal and minimal F_0 values in number of points and denoting them as F_{popt}, $F_{0\max}$, and $F_{0\min}$, respectively, the range of harmonic

number p is defined as,

$$p \in [p_{\min}, p_{\max}] \tag{16}$$

$$p_{\min} = \left\lfloor \frac{F_{popt}}{F_{0\max}} \right\rfloor, \quad p_{\max} = \left\lceil \frac{F_{popt}}{F_{0\min}} \right\rceil \tag{17}$$

here, $\lceil . \rceil$ and $\lfloor . \rfloor$ are referred to as the ceil and floor operations, respectively. Considering σ number of harmonics which is dependent on p, we propose an objective measure Ω_p by computing the σ-*th* root of the multiplication of the coefficients (amplitudes) of $\Gamma(f)$. As a preliminary criterion, the objective measure Ω_p for each possible value of p can be formulated as,

$$\Omega_p = \sqrt[\sigma]{\prod_{\xi=1}^{\sigma} \Gamma(f_\xi)}, \quad \sigma = p + 2, \tag{18}$$

where, f_ξ are the ξ-*th* harmonic frequencies corresponding to Ψ which is given by,

$$\Psi = \frac{F_{popt}}{p}, \qquad f_\xi = \lceil \xi \Psi \rceil \tag{19}$$

To this end, as a second criterion for extracting an appropriate p_{opt}, on obtaining Ω_p as per (18), three highest maximums of Ω_p are taken into account satisfying the condition that each maximum exceeds an experimentally determined threshold *Th* as given by,

$$Th = \frac{\Omega_{p\max}}{3} \tag{20}$$

here, $\Omega_{p\max}$ refers to the highest maximum of Ω_p. Among the selected values of $\Omega^{nn}_{p\max}, nn \in [1:3]$ with associated p, the value of p for which the objective measure Ω_p is maximized is regarded as p_{opt},

$$p_{opt} = \arg\max_{p} [\Omega^{nn}_{p\max}, nn \in [1:3]] \tag{21}$$

Thus for a voiced frame, p_{opt} leads to the desired estimate of pitch in number of points as given by,

$$F_0^{opt} = \frac{F_{popt}}{p_{opt}} \tag{22}$$

and the pitch in Hz (F_0) can be determined as,

$$F_0 = \left(\frac{F_0^{opt}}{2N} \right) F_s \tag{23}$$

here, N is the number of points used for DCT computation.

3. RESULTS AND PERFORMANCE COMPARISON

The performance of the proposed algorithm is evaluated using the *Keele* pitch extraction reference database [9]. 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 "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. In order to use this database for performance evaluation, the same analysis parameters (frame rate and basic window size) are chosen in the proposed algorithm. For windowing operation, we have used a normalized hamming window. Simulations are performed for clean speech (indicated as ∞ dB SNR), and white noise-corrupted speech (–10 dB to 30 dB). For LS minimization, $\tau_{\max} < M < N$ number of lags of ACF are employed, where, $\tau_{\max}$ is the maximum possible pitch period in samples. We have evaluated the estimated pitch values only for voiced frames based on the voiced/unvoiced labels included in the *Keele* database. According to Rabiner [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 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 case of white noise, in comparison to CAMDF [5] and WAC [6] methods, the global GPE [%] and the global RMSE [Hz] of the proposed algorithm is significantly superior for female and male speakers at almost all SNR levels and the efficacy is also better for clean speech. It is noticeable from Table I that for a white noise-corrupted speech, not only at a high SNR but also at a low SNR value, the overall m_{FPE} [Hz] and the overall σ_{FPE} [Hz] of the proposed method is within an acceptable limit and consistently better than the other methods. For a female voice, Fig. 3 shows a reference pitch contour accompanied by the spectrogram of clean speech that corresponds to an excerpt of 4 s from the reference

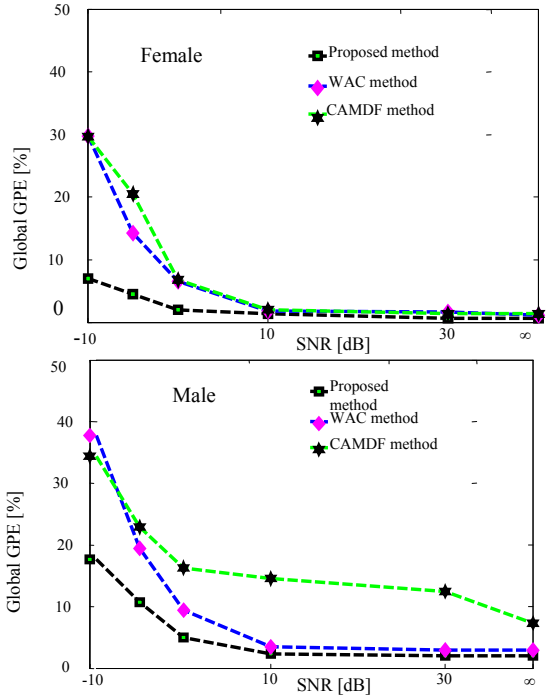

Fig. 1. Global GPE [%] as a function of SNR

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

Method	20 dB		-5 dB	
	Male	Female	Male	Female
Proposed	2.30 (2.16)	4.35 (4.57)	2.81 (2.84)	6.62 (6.49)
WAC	2.35 (2.38)	5.26 (5.33)	2.86 (2.85)	6.82 (7.12)
CAMDF	2.37 (2.39)	5.00 (4.99)	2.88 (2.99)	6.72 (6.82)

database. Also, pitch contours overlaid on the spectrograms of the white noise-corrupted speech at an SNR=-5 dB are portrayed for different methods. Through extensive analysis for white noise-corrupted speech signals, it is found that compared to other methods, the pitch contour resulting from the proposed pitch extraction algorithm is comparatively smoother even at a very low SNR.

4. CONCLUSION

In this paper, an algorithm to extract pitch from speech signals in the presence of a white noise is addressed. Our contributions are two-fold: at first, a harmonic sinusoidal correlation (HSC) model is proposed. Extraction of the parameters of this model via a least-squares minimization technique directly provides an accurate estimate of a pitch-harmonic (PH). Based on the PH, the characteristics of an enhanced power spectrum in the DCT domain are well utilized to develop a dual criterion that leads to an achievement of the desired harmonic number required for pitch detection. Through simulation results it has been ascertained that the proposed algorithm significantly outperforms some of the reported methods in a heavy noisy scenario.

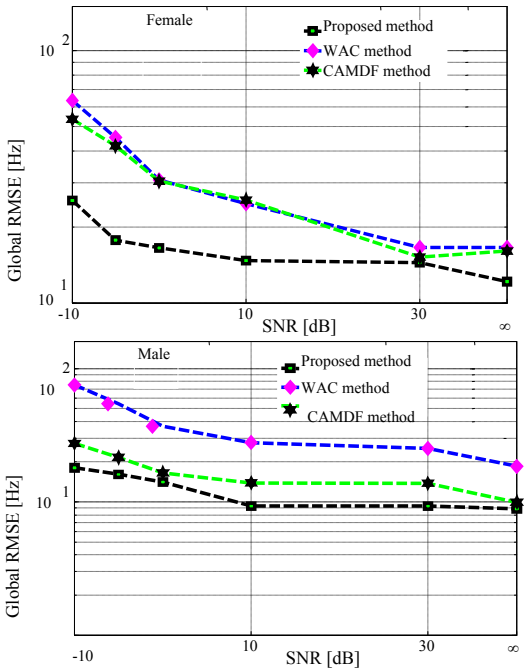

Fig. 2. Global RMSE [Hz] as a function of SNR

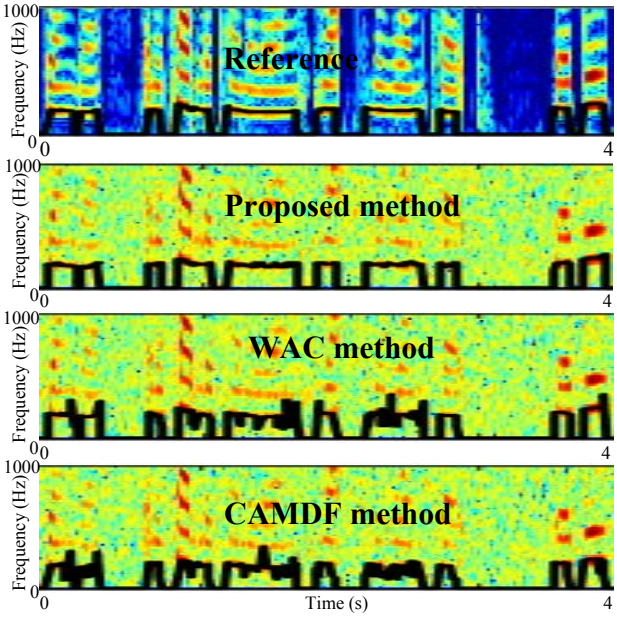

Fig. 3. Comparison of pitch contours at SNR=-5 dB in white noise

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] 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.

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

[8] C. Shahnaz, W. -P Zhu, and M. O. Ahmad, "A robust pitch estimation algorithm in noise," in *Proc. ICASSP2007*, Hawaii, USA, pp. 1073-1076, April, 2007.

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