A Residual-Cepstrum Method of Pitch Estimation from Noisy Speech

Celia Shahnaz, Wei-Ping Zhu, and M. Omair Ahmad
Centre for Signal Processing and Communications
Dept. of Electrical and Computer Engineering, Concordia University,
Montreal, Quebec, Canada H3G 1M8

1. INTRODUCTION

Pitch is an important speech parameter in speaker recognition, speech synthesis, coding, and articulation training for the deaf. The pitch estimation is to determine the fundamental frequency (F_0) or period (T_0) of a vocal cord vibration causing periodicity in the speech signal. This task becomes very difficult when the speech observations are heavily corrupted by noise. Most of the methods proposed in the literature are capable of estimating pitch from clean speech [1]. As noise obscures the periodic structure of speech, many existing methods fail to provide accurate pitch estimates under noisy conditions.

In this paper, residual and cepstral representations of speech are utilized for pitch estimation in a noisy environment. For a voiced speech, the major excitation of the vocal tract within a pitch period occurs at the instant of glottal closure (GC). It is possible to determine the pitch period by careful analysis of the speech signal with the help of GC instants. Some characteristics of the GC instants can be better observed in a residual signal (RS) of speech in comparison to the speech signal itself. However, it is difficult to use the RS directly for pitch estimation because of its bipolar fluctuations around the GC instants. In order to overcome this limitation, we derive a Hilbert envelope (HE) of the RS, which presents a unipolar nature at the GC instants. Under a severe noisy condition, the time difference of successive peaks of the HE of the RS may not provide an accurate estimate of the true pitch period. Hence, with a view to overcome the adverse effect of noise on the HE of the RS, we propose a discrete Fourier transform (DFT) based power cepstrum (DFTPC) of the HE that exhibits a more prominent pitch-peak even in a heavily degraded condition in comparison to that demonstrated by the conventional cepstrum of the noisy speech. Simulation results testify that the global maximization of the DFTPC yields an accurate pitch estimate compared to the state-of-the-art methods in an intricate noisy scenario for a wide range of speakers.

2. PROPOSED METHOD

2.1 Pre-processing

Each windowed noisy frame of the observed noisy speech is low-pass filtered to remove very high-frequency contents. Such a windowed filtered noisy speech frame is given by

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

where, $x(n)$ and $v(n)$ represent the windowed and low-pass filtered version of clean speech and uncorrelated additive noise, respectively. Such a pre-processing assumes to retain 4-5 formants, which facilitates the extraction of vocal-tract system parameters required for the RS generation

2.2 Pitch Estimation

One approach to derive the information about the GC instants for the extraction of pitch of speech signal is the Linear Prediction (LP) analysis. In order to remove the vocal-tract information from the process of pitch estimation, if we perform an inverse-filtering of the noise-corrupted speech $y(n)$ in a frame, where $y(n)$ is let to pass through an inverse vocal-tract system filter $\hat{A}(z)$ given by

$$\hat{A}(z) = 1 + \sum_{k=1}^{p} \hat{a}_k z^{-k} \tag{2}$$

the output of the inverse vocal-tract system filter is referred as the error or residual signal(RS):

$$e(n) = T^{-1}\left[\hat{A}(z)Y(z)\right] = y(n) + \sum_{k=1}^{p} \hat{a}_k y(n-k) \tag{3}$$

with T^{-1} representing the inverse operator of a z transform T, where $T[y(n)]=Y(z)$. In (2) and (3), p is the order of linear prediction and $\hat{a}_k$ are the vocal-tract system parameters to be identified prior to inverse filtering. They are obtained using the LP analysis based on the autocorrelation function (ACF) $\phi_y(m)$ of $y(n)$ as

$$\phi_y(m) = -\sum_{k=1}^{p} \hat{a}_k \phi_y(m-k), \ 0 < m \le p \tag{4}$$

where m is the discrete lag variable and note that $\phi_y(m)$ obeys a recursive relation that relates the $\phi_y(m)$ values to the $\hat{a}_k$ parameters. For an N-sample frame of $y(n)$, the ACF $\phi_y(m)$ in (4) can be estimated as,

$$\phi_y(m) = \frac{1}{N} \sum_{n=0}^{N-1-|m|} y(n)y(n+|m|), \ m = 0, \pm 1, \ldots.., \pm M, \ M < N \tag{5}$$

Since in the presence of noise, lower lags of $\phi_y(m)$ are generally become more corrupted than that of the higher lags, a few lower lags of $\phi_y(m)$ are avoided in the computation of the $\hat{a}_k$ parameters. Utilizing the ACF coefficients $\phi_y(p+1)$,…… $\phi_y(p+\mathrm{S})$ in (4) yields a set of linear equations, which can be represented in the following matrix form

$$\begin{bmatrix} \phi_y(p) & \phi_y(p-1) \ldots & \phi_y(1) \\ \phi_y(p+1) & \phi_y(p) & \ldots \phi_y(2) \\ \vdots & \vdots & \vdots \\ \phi_y(p+S-1) \ldots & & \ldots \phi_y(S) \end{bmatrix} \begin{bmatrix} \hat{a}_1 \\ \hat{a}_2 \\ \vdots \\ \hat{a}_p \end{bmatrix} = - \begin{bmatrix} \phi_y(p+1) \\ \phi_y(p+2) \\ \vdots \\ \phi_y(p+S) \end{bmatrix} \quad (6)$$

where S governs the number of equations to be used in (6). The $\hat{a}_k$ parameters can easily be obtained from the least-squares solution of (6) to generate the RS according to (3). It is found difficult to use the RS directly for the detection of the GC instants due to the occurrence of peaks of either polarity around the GC instants. Furthermore, in a noisy condition, the RS could be significantly different from the excitation signal due to the inaccurate estimates of $\hat{a}_k$ parameters. However, this ambiguity can be reduced by computing the Hilbert envelope (HE) of the RS as

$$E(n) = \sqrt{e^2(n) + e_h^2(n)} \quad (7)$$

where $e_h(n)$ is the Hilbert transform of the RS $e(n)$. The HE of $e(n)$ is a unipolar positive function. The correlation among the samples of the HE of $e(n)$ is high compared to the corresponding samples in the $e(n)$. However, in a severe noisy condition, by detecting the peaks at the GC instants in the HE of $e(n)$ and taking the time difference of its successive peaks, we may not obtain the true pitch period. Hence, with a view to overcome the undesirable effect of noise on the HE of the $e(n)$, we propose a discrete Fourier transform based power cepstrum (DFTPC) of the HE as given by

$$c(n) = \left(T^{-1}\left[\log\left|T[E(n)]^2\right]\right]\right)^2 \quad (8)$$

In (8), as usual natural logarithm is used and T^{-1} represents the inverse operator of the discrete Fourier transform T. The DFTPC of the HE is more effective in that it emphasizes the true pitch-peak even in a heavily degraded condition in comparison to that depicted by the conventional cepstrum of $y(n)$. If F_s is the sampling frequency (Hz), by searching for the global maximum of $c(n)$, the desired pitch (F_0) is obtained as

$$\hat{F}_0 = \frac{F_s}{\hat{T}_0}, \; \hat{T}_0 = \arg\max_m [c(n)] \quad (9)$$

3. RESULTS AND DISCUSSION

The performance of the proposed method is evaluated using the *Keele* reference database [2]. This database is of studio quality, sampled at 20 kHz with 16-bit resolution. It provides a reference pitch at a frame rate of 100 Hz with 25.6 ms window. The noisy speech with SNR varying from 5 dB to ∞ dB is considered for Simulations, where white noise from the *NOISEX'92* database is used. In order to use the *Keele* database, we have chosen the same analysis parameters (frame rate and basic window size). For windowing operation, we have used a normalized hamming window. In the estimation of $\hat{a}_k$ parameters by (6), S is chosen as 5p.

Table 1. Percentage gross pitch-error for white noise-corrupted speech at SNR = 5dB

Methods	Female	Male
Proposed Method	6.98	10.79
CEP Method	24.56	26.80
ACF Method	16.54	19.75
AMDF Method	19.40	28.75

For performance evaluation, we have used the voiced/ unvoiced labels included in the database as well as the true pitch value F_0. As our performance metric, we defined percentage gross pitch-error which is the ratio of the number of frames giving ''incorrect'' pitch values to the total number of frames multiplied by 100. As reported in [3], estimated $\hat{F}_0$ is considered as ''incorrect'' if it falls outside 20% of the true pitch value F_0.

We have compared the performance of the proposed pitch estimation method with the conventional cepstrum (CEP), autocorrelation function (ACF), and average magnitude difference function (AMDF) methods [1]. For a speaker group, the percentage gross pitch-error is calculated considering two male (or female) speakers. In Table 1, the percentage gross pitch-error for female and male speaker groups are summarized considering the white noise noise-corrupted speech signals at an SNR = 5 dB. It is evident that in comparison to the other methods, percentage gross pitch-errors of the proposed method are significantly reduced for both female and male speakers in the presence of a white noise with a low SNR value. The lower values of percentage gross pitch-errors obtained from the proposed method for all speaker groups in a noisy environment are the testimony of its accuracy against a background noise.

4. CONCLUSION

In this paper, a new method based on residual and cepstral features is presented for pitch estimation from speech corrupted by a white noise. We generated a Hilbert envelope (HE) of the residual signal (RS) and argue that the DFT based power cepstrum (DFTPC) of the HE is more capable of reducing the pitch-errors in a difficult noisy condition. Simulation results using naturally spoken sentences have shown that the proposed method can estimate pitch in a noisy environment with a superior efficacy for both female and male speakers compared to some of the existing methods.

REFERENCES

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

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

[3] Alain de Chevengne, and Hideki Kawahara, "YIN, a fundamental frequency estimator for speech and music," *J. Acoust. Soc. Amer.* , vol. 111, pp. 1917-1930, 2002.