PITCH ESTIMATION FROM NOISY SPEECH BASED ON RESIDUAL-TEMPORAL INFORMATION

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

In speaker recognition, speech synthesis, coding, and articulation training for the deaf, pitch is an important speech parameter. Determining the fundamental frequency (F_0) or period (T_0) of a vocal cord vibration causing periodicity in the speech signal is the aim of pitch estimation. 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, the task of pitch estimation becomes very difficult when the speech observations are heavily corrupted by noise. Hence, many existing methods fail to provide accurate pitch estimates under noisy conditions.

In this paper, residual and temporal representations of speech are utilized for pitch estimation in a noisy environment. For a voiced speech, at the instant of glottal closure (GC), the major excitation of the vocal tract within a pitch period occurs. By careful analysis of the speech signal with the help of GC instants, it is possible to determine the pitch period. In comparison to the speech signal itself, in a residual signal (RS) of speech, some characteristics of the GC instants can be better observed. However, because of the bipolar fluctuations of RS around the GC instants, it is difficult to use the RS directly for pitch estimation. 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, we propose a circular average magnitude sum function (CAMSF) of the HE that exhibits more prominent peaks even in a heavily degraded condition. Simulation results testify that the global maximization of the temporal function, CAMSF, 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

A windowed filtered noisy speech frame is given by

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

where, $s(n)$ and $v(n)$ represent the windowed and low-pass filtered version of clean speech and uncorrelated additive noise, respectively. Each windowed noisy frame of the observed noisy speech is low-pass filtered to remove very high-frequency contents. 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

Linear Prediction (LP) analysis is commonly used to derive the information about the GC instants for the extraction of pitch of speech signal. In LP analysis, in order to remove the vocal-tract information from the process of pitch estimation, an inverse-filtering operation is performed on the noise-corrupted speech $x(n)$ in a frame. The output of the inverse filter is referred as the error or residual signal (RS)

$$\Re(n) = T^{-1}\left[\hat{G}(z)X(z)\right] = x(n) + \sum_{k=1}^{p} \hat{g}_k x(n-k), \tag{2}$$

where T^{-1} represents the inverse operator of a z transform T with $T[x(n)]=X(z)$, p is the order of linear prediction, and $\hat{g}_k$ are the vocal-tract system parameters to be identified prior to inverse filtering. Here, $x(n)$ is passed through an inverse vocal-tract system filter $\hat{G}(z)$ given by

$$\hat{G}(z) = 1 + \sum_{k=1}^{p} \hat{g}_k z^{-k}. \tag{3}$$

For an N-sample frame of $x(n)$, the autocorrelation function (ACF) $\chi(m)$ of $x(n)$ can be estimated as,

$$\chi(m) = \frac{1}{N} \sum_{n=0}^{N-1-|m|} x(n)x(n+|m|),\ m = 0, \pm 1, \ldots\ldots, \pm M,\ M < N, \tag{4}$$

where m is the discrete lag variable, and $\chi(m)$ obeys a recursive relation that relates the $\chi(m)$ values to the $\hat{g}_k$ parameters as

$$\chi(m) = -\sum_{k=1}^{p} \hat{g}_k \chi(m-k),\ 0 < m \leq p. \tag{5}$$

The vocal-tract system parameters are obtained using the LP analysis based on the ACF $\chi(m)$ of $x(n)$. Since in the presence of noise, lower lags of $\chi(m)$ are generally become more corrupted than that of the higher lags, a few lower lags of $\chi(m)$ are avoided in the computation of the $\hat{g}_k$ parameters as

$$\begin{bmatrix} \chi(p) & \chi(p-1) \ldots & \chi(1) \\ \chi(p+1) & \chi(p) \ \ldots & \chi(2) \\ \vdots & \vdots & \vdots \\ \chi(p+\lambda-1) \ldots & \ldots & \chi(\lambda) \end{bmatrix} \begin{bmatrix} \hat{g}_1 \\ \hat{g}_2 \\ \vdots \\ \hat{g}_p \end{bmatrix} = -\begin{bmatrix} \chi(p+1) \\ \chi(p+2) \\ \vdots \\ \chi(p+\lambda) \end{bmatrix}. \tag{6}$$

A set of linear equations, which can be represented in the above matrix form is obtained utilizing the ACF coefficients $\chi(p+1)$,...... ,$\chi(p+\lambda)$ in (5). The number of equations to be used in (6) is governed by λ. The $\hat{g}_k$ parameters can easily be obtained from the least-squares solution of (6) to generate the RS according to (2). 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 $\hat{g}_k$ parameters. However, this ambiguity can be reduced by computing the Hilbert envelope (HE) of the RS as

$$H(n) = \sqrt{\Re^2(n) + \Re_h^2(n)} \ , \qquad (7)$$

where $\Re_h(n)$ is the Hilbert transform of the RS $\Re(n)$. The HE of $\Re(n)$ is a unipolar positive function. The correlation among the samples of the HE of $\Re(n)$ is high compared to the corresponding samples in the $\Re(n)$. However, in a severe noisy condition, by detecting the peaks at the GC instants in the HE of $\Re(n)$ 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 $\Re(n)$, we propose a circular average magnitude sum function (CAMSF) of the HE as given by

$$\eta(m) = \sum_{n=0}^{E-1} \left| H(\mathrm{mod}(n+m, E)) + H(n) \right|, \ m = 0,1,....., E-1. \qquad (8)$$

In (8), E is the number of speech samples employed to compute CAMSF for every lag m. The CAMSF of the HE is more effective in that it emphasizes the true pitch-peak even in a heavily degraded condition. By searching for the global maximum of the temporal function CAMSF, the desired pitch (F_0) is obtained as

$$\hat{F}_0 = \frac{F_s}{\hat{T}_0}, \ \hat{T}_0 = \arg\max_m [\eta(m)], \qquad (9)$$

where F_s is the sampling frequency (Hz).

3. RESULTS AND DISCUSSION

By using the *Keele* reference database [2], the performance of the proposed method is evaluated. 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. In order to use the *Keele* database, we have chosen the same analysis parameters (frame rate and basic window size). 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. For windowing operation, we have used a normalized hamming window. In the estimation of $\hat{g}_k$ parameters by (6), λ is chosen as $5p$.

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

Methods	Female	Male
Proposed Method	5.79	9.98
ACF Method	16.54	19.75
AMDF Method	19.40	28.75

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. For performance evaluation, we have used the voiced/ unvoiced labels included in the database as well as the true pitch value F_0.

For a speaker group, the percentage gross pitch-error is calculated considering two male (or female) speakers. We have compared the performance of the proposed pitch estimation method with the conventional autocorrelation function (ACF), and average magnitude difference function (AMDF) methods [1]. 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 is 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 temporal features is presented for pitch estimation from speech corrupted by a white noise. We computed a Hilbert envelope (HE) of the residual signal (RS) and found that the CAMSF 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.