A CEPSTRAL-DOMAIN ALGORITHM FOR PITCH ESTIMATION FROM NOISE-CORRUPTED 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

Reliable pitch estimation is one of the most significant problems in speech processing applications since any error in pitch detection has a deleterious effect on system performance. Many applications also need robust pitch estimation from noise-corrupted speech. Since noise obscures the periodic structure of speech, the pitch estimation in noise is an intricate task. Even though a large number of pitch estimation algorithms have been disclosed in literature for clean speech, it is rather surprising that only a few algorithms have been proposed for the pitch estimation in the presence of noise [1]-[2].

The main purpose of the present work was to develop an accurate algorithm for pitch estimation from noisy speech observations with an aim to substantially reduce the pitch-errors for a wide range of speakers. We propose to employ a Discrete Cosine Transform (DCT) based power spectral subtraction scheme for enhancing noisy speech prior to pitch estimation. Then, in order to remove the detrimental effect of formants, the de-noised speech is inverse filtered to yield an output referred to as the Linear Prediction (LP) residual. The kernel of the proposed method lies in the introduction of a DCT power cepstrum (DPC) of the LP residual that exhibits a more prominent peak at the true pitch relative to that demonstrated by the conventional cepstrum of noisy speech. Consequently, global maximization of the DPC results in a momentous improvement in the pitch estimation accuracy. Extensive simulation results confirm that for both low and high-pitched speakers, our algorithm consistently outperforms the state-of-the-art pitch estimation methods in the white or multi-talker babble environmental noise.

2. PROPOSED METHOD

2.1 Pre-processing

Assuming $x(n)$ and $d(n)$ as clean speech and additive noise signals, respectively, the observed noisy signal $y(n)$ is given by,

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

$y(n)$ is segmented into frames with a frame size N by the application of a window function $w(n)$. As a pre-processing, one-dimensional DCT is performed on the windowed noisy frame $y'(n)$ which is given by the relation,

$$Y'(k) = c(k)\sum_{n=1}^{N} y'(n)\cos\left[\frac{\pi(2n-1)(k-1)}{2N}\right] \tag{2}$$

In (2), the co-efficients c are given as,

$$c(k) = \sqrt{\frac{1}{N}} \quad for\, k = 1,\ c(k) = \sqrt{\frac{2}{N}} \quad for \quad 2 \le k \le N \tag{3}$$

The DCT co-efficients corresponding up to the upper frequency limit of the first formant is retained and the rest of the co-efficients is set to zero. From the resulting DCT sequence denoted as, $Y_w(k)$, a time domain pre-processed noisy speech frame, $y_w(n)= x_w(n)+d_w(n)$, can be obtained through the inverse DCT operation. Since noise is additive both in signal and DCT domain, $Y_w(k)$ can be written as,

$$Y_w(k) = X_w(k) + D_w(k) \tag{4}$$

where, $X_w(k)$ and $D_w(k)$ are the DCTs of $x_w(n)$ and $d_w(n)$, respectively.

2.2 Noise reduction

The instantaneous power spectrum of $y_w(n)$ is approximated as follows,

$$|Y_w(k)|^2 \approx |X_w(k)|^2 + |D_w(k)|^2 \tag{5}$$

Prior to pitch estimation, in order to enhance speech, a DCT based modified power spectral subtraction scheme is derived from (5) as,

$$|\hat{X}_w(k)|^2 = \begin{cases} |Y_w(k)|^2 - \alpha|\hat{D}_w(k)|^2 & if\ |\hat{X}_w(k)|^2 > 0 \\ \beta|\hat{D}_w(k)|^2 , otherwise \end{cases} \tag{6}$$

where, β is the spectral floor parameter, and α refers to the over-subtraction factor. The DCT noise power spectrum is estimated from the beginning silence frames and updated during the immediate past silence frames before the speech frame using the following averaging rule,

$$|\hat{D}_w^m(k)|^2 = \lambda|\hat{D}_w^{m-1}(k)|^2 + (1-\lambda)|Y_w^m(k)|^2 \tag{7}$$

where, m represents the frame index, and λ the forgetting factor. In order to compensate for the noise spectrum errors, the value of α is adequately adapted from frame to frame as a function of segmental noisy signal to noise ratio ($NSNR$) of the frame where, $NSNR$ and α are formulated as,

$$NSNR = \frac{\sum_k |Y_w(k)|^2}{\sum_k |\hat{D}_w(k)|^2}, \quad NSNR_{dB} = 10\log_{10} NSNR \tag{8}$$

$$\alpha = \alpha_0 - \frac{NSNR_{dB}}{s}, \quad -5 \le NSNR_{dB} \le 20 \tag{9}$$

where, $\alpha = 1$ for $NSNR_{dB} > 20$, $\alpha = \alpha_{\max}$ for $NSNR_{dB} < -5$, with α_0, the desired over-subtraction factor at $NSNR_{dB} = 0$, $\alpha_{\max}$, the maximum allowable value of α, the constant s are chosen experimentally as 4, 5 and 20/3, respectively. Once the subtraction is performed in the DCT domain based on (6), an enhanced speech frame is obtained using the following relationship,

$$x_w^{en}(n) = IDCT\left\{|\hat{X}_w(k)|e^{j\arg(Y_w(k))}\right\} \tag{10}$$

where, $IDCT$ stands for the inverse DCT.

2.3 Pitch Estimation

Our idea is to translate the pitch estimation problem from $y_w(n)$ into a problem of estimating the pitch from $x_w^{en}(n)$ based on the knowledge of the instants of Glottal Closure (GCIs) derived from the Linear Prediction (LP) analysis. Representing τ as the lag variable and estimating the ACF of $x_w^{en}(n)$ as,

$$\phi_x(\tau) = \frac{1}{N}\sum_{n=0}^{N-1-\tau} x_w^{en}(n)x_w^{en}(n+\tau), \quad \tau \geq 0 \tag{11}$$

least squares solution of the Linear Prediction Co-efficients (LPCs) denoted as, $\{a_k, k=1,\ldots.,p\}$ can be computed using the following set of equations,

$$\begin{bmatrix} \phi_x(p) & \phi_x(p-1) \ldots. & \phi_x(1) \\ \phi_x(p+1) & \phi_x(p) & \ldots. \phi_x(2) \\ \vdots & \vdots & \vdots \\ \phi_x(p+S-1) \ldots. & & \ldots. \phi_x(S) \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{bmatrix} = -\begin{bmatrix} \phi_x(p+1) \\ \phi_x(p+2) \\ \vdots \\ \phi_x(p+S) \end{bmatrix} \tag{12}$$

In (12), p is the order of prediction and $\tau = p+1, p+2, \ldots, p+S$ are considered, where, S governs the number of equations. In order to remove harmful vocal-tract (formants) effect from the extraction procedure of pitch, $x_w^{en}(n)$ is inverse filtered with the LP parameters $\{a_k\}$. The output of the inverse filter is referred to as the linear prediction (LP) residual given by,

$$\Lambda(n) = x_w^{en}(n) + \sum_{k=1}^{p} a_k x_w^{en}(n-k) \tag{13}$$

The LP residual $\Lambda(n)$ corresponds to an estimate of the excitation source of $x_w^{en}(n)$. In order to handle the heavy noise, a DCT power cepstrum (DPC) of the LP residual $\Lambda(n)$ is introduced as,

$$\Omega(n) = \left(IDCT(\log\left|DCT(\Lambda(n))\right|^2)\right)^2 \tag{14}$$

The DPC of the LP residual is more convenient in that it emphasizes the true pitch-peak compared to that revealed by the conventional cepstrum. If F_s is the sampling frequency (Hz), by searching for the global maximum of $\Omega(n)$, the desired pitch (F_0) is obtained as,

$$F_0 = \frac{F_s}{T_0}, T_0 = \arg\max_n[\Omega(n)] \tag{15}$$

3. SIMULATION RESULTS

3.1 Simulation conditions

The performance of the proposed method is evaluated using the *Keele* reference database [3]. This database provides a reference pitch at a frame rate of 100 Hz with 25.6 ms window. The *Keele* database has studio quality, sampled at 20 kHz with 16-bit resolution. In order to use this database, we have chosen the same analysis parameters (frame rate and basic window size). For simulation, white and multi-talker babble noises from the *NOISEX'92* database are used. The noisy speech with SNR varying from 5 dB to ∞ dB is considered for Simulations. For windowing operation, we have used a normalized hamming window. The parameter β was set to 0.002 and in the estimation of $\{a_k\}$ parameters, S is chosen as $5p$ with p=10.

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

Speaker	Proposed Method	CEP Method	WAC Method
Male	5.05	26.80	9.38
Female	3.71	24.56	6.61

Table 2. Percentage gross pitch-error for the multi-talker babble-noise corrupted speech at SNR = 5dB

Speaker	Proposed Method	CEP Method	WAC Method
Male	10.59	40.2	19.35
Female	5.40	31.63	14.53

3.2 Performance comparison

We have evaluated and compared the performance of the proposed pitch estimation algorithm with the cepstrum (CEP) [2] and weighted autocorrelation (WAC) methods [4]. For a speaker group, the percentage gross pitch-error GPE (%) is calculated considering two male (or female) speakers. We have defined GPE (%) which is the ratio of the number of frames giving ''incorrect'' pitch values to the total number of frames. The estimated pitch is considered as ''incorrect'' if it falls outside 20% of the true pitch value. In Table 1. and Table 2., GPE(%) for male and female speaker group are summarized considering the white noise and babble noise-corrupted speech signals, respectively, at an SNR=5 dB. It is evident that in comparison to the CEP and WAC methods, a significant reduction in the GPEs(%) is achieved by the proposed algorithm for both female and male speakers in the presence of a white or a babble noise.

4. DISCUSSION

In this paper, a new pitch estimation algorithm for noisy speech preceded by a DCT domain noise reduction scheme is presented. In order to indicate accurately the approximate location of the GCIs to be used for pitch estimation, a DCT power cepstrum (DPC) of the LP residual is proposed that is capable of reducing the pitch-errors in a difficult noisy condition. We argue through simulation results that the proposed method is suitable for a wide range of speakers and significantly outperforms some of the reported methods in the present of a heavy noise in terms of percentage gross pitch-errors.

REFERENCES

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

[2] W. J. Hess, (1993). Pitch Determination of Speech Signals. *New York: Springer.*

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

[4] T. Shimamura and H. Kobayashi, (2001). Weighted autocorrelation for pitch extraction of noisy speech. *IEEE Trans. Speech Audio Processing*, 9, 727-730.