A TIME-DOMAIN PITCH EXTRACTION SCHEME FOR NOISY SPEECH SIGNALS

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
{c_shahna, weiping, omair}@ece.concordia.ca

1. INTRODUCTION

Pitch is the primary acoustic cue in several applications like voiced/unvoiced classification, speaker recognition, speech enhancement, synthesis, coding, and articulation training for the deaf learning to speak. During the production of speech, the vibration of vocal folds appears to be periodic and the estimation of pitch involves determination of the fundamental frequency (F_0) or fundamental period (T_0) of this vibration. The extraction of pitch is the object of research over the past several decades [1]-[3]. All the algorithms proposed in the literature for clean speech may be broadly classified into three categories, namely, algorithms using time domain properties, algorithm using frequency domain properties and algorithms using both time and frequency domain properties of the speech signals. However, performance of these methods degrades under noisy conditions as noise obscures the periodic structure of speech.

In this paper, a new time-domain pitch extraction scheme for speech signals subject to noise degradation is presented. In order to remove the deleterious vocal-tract information, we propose to pass the pre-processed noisy speech through an inverse filter whose parameters are derived from the Linear Prediction (LP) analysis, yielding an output referred to as the LP residual. The novelty of the proposed scheme lies in the introduction of a new average magnitude sum function (AMSF) of the LP residual which is expected to reveal more prominent peaks at integral multiples of the pitch period compared to that revealed by the LP residual. With a view to quell the pitch-errors considerably in a severe noisy scenario, the peaks of AMSF at different pitch-harmonic locations are added and weighted by a periodicity dependent weighting factor for every possible pitch period. The resulting weighted and harmonically summed AMSF of the LP residual is globally maximized to extract the desired pitch period. The proposed method is able to reflect its efficacy to a significant extent for extracting pitch of both low and high-pitched speakers in the white or car environmental noise.

2. PROPOSED METHOD

Let $x(n)$ and $v(n)$ denote clean speech and uncorrelated additive noise, respectively. The observed noisy signal $y(n)$, given by, $y(n) = x(n) + v(n)$, is divided into overlapping frames with frame size N by applying a window function $w(n)$. The windowed noisy frame denoted as $y_w(n)$ 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 is denoted as $y_{PPS}(n)$.

The instant at which the closure of vocal folds occurs within a pitch period is defined as the GC event. One approach to derive information about the GC events for the extraction of pitch of speech signal is the Linear Prediction (LP) analysis. In the LP analysis, assuming $y_{PPS}(n)$ as the output of an autoregressive (AR) spectral shaping filter [1], the predicted sample of $y_{PPS}(n)$ can be modeled by a linear combination of the p previous output samples, $\tilde{y}_{PPS}(n) = -\sum_{k=1}^{p} a_k y_{PPS}(n-k)$,

here, p is the order of prediction and $\{a_k\}$ are the LP Co-efficients (LPCs) computed using the following equations,

$$R_y(l) = -\sum_{k=1}^{p} a_k R_y(l-k), \quad l = 1,2,3,\ldots\ldots,p \tag{1}$$

here, $R_y(l)$, the ACF of $y_{PPS}(n)$, can be estimated as,

$$R_y(l) = \frac{1}{N}\sum_{n=0}^{N-1-\tau} y_{PPS}(n)\, y_{PPS}(n+l), \quad l \geq 0 \tag{2}$$

In order to handle the noisy environment, $l = p + 1, p + 2,\ldots.,$ $p + S$ are considered in (1) and thus yielding the following modified least-squares Yule–Walker (MLSYW) equations,

$$\begin{bmatrix} R_y(p) & R_y(p-1) & \ldots. & R_y(1) \\ R_y(p+1) & R_y(p) & \ldots. & R_y(2) \\ \vdots & \vdots & & \vdots \\ R_y(p+S-1) & \ldots. & \ldots. & R_y(S) \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{bmatrix} = - \begin{bmatrix} R_y(p+1) \\ R_y(p+2) \\ \vdots \\ R_y(p+S) \end{bmatrix} \tag{3}$$

where, S governs the number of equations in (3). In order to reduce the estimation variance of Yule–Walker method, a combination of more than p equations are used. Least squares solution of (3) provides AR parameters $\{a_k\}$. To remove the information of vocal-tract (formants) from the extraction procedure of pitch, $y_{PPS}(n)$ is passed through an inverse or prediction filter which is given by,

$$A(z) = 1 + \sum_{k=1}^{p} a_k z^{-k} \tag{4}$$

where, $\{a_k\}$ are already derived from the LP analysis.

The output of the inverse filter is the error between the actual sample $y_{PPS}(n)$ and the predicted sample $\tilde{y}_{PPS}(n)$, which is referred to as the LP residual signal given by,

$$e(n) = y_{PPS}(n) + \sum_{k=1}^{p} a_k y_{PPS}(n-k) \quad (5)$$

As GC event is the place of significant excitation, large error is associated with GC event in the LP residual. One approach for the extraction of pitch is to detect the peaks at the GC events in the LP residual. It is possible to obtain the information about the pitch period (T_0) from the time difference of successive peaks of the LP residual. However, as a more convenient approach for the extraction of pitch from a noisy speech, an average magnitude sum function (AMSF) of the LP residual is proposed as,

$$\Re(l) = \frac{1}{N} \sum_{n=0}^{N-1} |e(n) + e(n+l)| \quad (6)$$

where, the property of high correlation among the samples of the LP residual around the GC events is exploited. For a quasi-periodic frame of voiced speech, $\Re(l)$ exhibits local maxima at $l = \rho T_0$, where ρ is an integer, ρ=0,1,2……and T_0 is the pitch period. Exploiting this feature, a weighted and harmonically summed AMSF of the LP residual is formulated as,

$$\chi(\tau) = \tau \sum_{\gamma=1}^{\partial} \Re(\gamma\tau), \qquad \tau = \tau_{\min}, \ldots, \tau_{\max} \quad (7)$$

where, τ represents the possible pitch-period that ranges from $\tau_{\min}$ to $\tau_{\max}$. In general, for most male and female speakers, $\tau_{\min}$ and $\tau_{\max}$ are found, respectively, as $[F_s/(500\text{Hz})]$ and $[F_s/(50\text{Hz})]$ where, F_s is the sampling frequency. In (7), ∂ is the number of pitch-harmonics and it should be chosen such that $\partial\tau \leq M$. It is important to note that instead of considering only the global maximum of $\Re(l)$, since we sum up the peaks of $\Re(l)$ at different harmonics of every possible pitch-period with a periodicity dependent weighting factor, $\chi(\tau)$ exhibits clear and quite sharper peaks for voiced frames. Therefore, by searching for the global maximum of $\chi(\tau)$, the desired pitch frequency $\hat{F}_0$ is obtained as, $\hat{F}_0 = F_s/\hat{T}_0$ and $\hat{T}_0 = \arg\max_{\tau}[\chi(\tau)]$ is the estimated pitch period.

3. SIMULATION RESULTS

We have defined percentage gross pitch-error which is the ratio of the number of frames giving ''incorrect'' pitch values to the total number of frames. As reported in [3], estimated $\hat{F}_0$ is considered as ''incorrect'' if it falls outside 20% of the true pitch value F_0. The performance of the proposed method is evaluated using the *Keele* reference database [4]. 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).

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

Speaker	Proposed Method	ACF Method	AMDF Method
Male	11.22	19.75	28.75
Female	6.12	16.54	19.4

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

Speaker	Proposed Method	ACF Method	AMDF Method
Male	20.24	27.75	32.29
Female	10.26	17.31	21.25

For simulation, white and car noises from the *NOISEX'92* database are used as the additive background noises. 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. In the estimation of $\{a_k\}$ parameters by (3), S is chosen as $5p$ with p=10 .We have evaluated and compared the performance of the proposed pitch detection scheme with the conventional average magnitude difference function (AMDF) and autocorrelation function (ACF) methods [2]. For a speaker group, the percentage gross pitch-error (GPE(%)) is calculated considering two male (or female) speakers. In Table 1. and Table 2., GPE(%) for male and female speaker group are summarized considering the white noise and car noise-corrupted speech signals, respectively, at an SNR=5 dB. It is evident that in comparison to the ACF and AMDF methods, GPEs(%) of the proposed algorithm are significantly reduced for both female and male speakers in the presence of a stationary white or a car noise.

4. CONCLUSION

In this paper, a new pitch detection algorithm for speech corrupted by a white or a car noise is presented considering the LP residual of the pre-processed speech as a representation for the GC events. A weighted and harmonically summed AMSF of the LP residual is proposed that is able to effectively quell the pitch-errors in the presence of a noise. Simulation results have shown that the proposed method significantly outperforms some of the reported pitch detection algorithms implemented in the same noisy environment.

REFERENCES

[1] O'Shaughnessy, D. (2000). *Speech Communications: Human and Machine* (2nd ed.). IEEE Press, NY.

[2] Rabiner, L., Cheng, M., Rosenberg, A., and McGonegal, C. (1976). A comparative performance study of several pitch detection algorithms. *IEEE Trans. Acoust., Speech, Signal Process.*, 24, 399–418.

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

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