EXTRACTION OF PITCH IN ADVERSE CONDITIONS

S. R. Mahadeva Prasanna and B. Yegnanarayana

Speech and Vision Laboratory
Department of Computer Science and Engineering
Indian Institute of Technology Madras, Chennai - 600 036, INDIA
Email: {prasanna,yegna}@cs.iitm.ernet.in

ABSTRACT

This paper proposes a method for the extraction of pitch in adverse conditions. Real environment in which the degradation is due to several unpredictable sources like, additive noise, reverberation and channel noise is treated as adverse condition in this study. The proposed method is based on the knowledge of Glottal Closure (GC) events. GC event is the instant at which closure of vocal folds takes place within a pitch period. The Hilbert envelope of the Linear Prediction (LP) residual gives information about the location of GC events. Autocorrelation analysis is performed on the Hilbert envelope of the LP residual. The properties of the Hilbert envelope of the LP residual are exploited for the extraction of pitch from the autocorrelation sequence. The results of the proposed method are compared with the Simple Inverse Filtering Technique (SIFT) algorithm. The performance of the proposed algorithm is found to be superior, even in adverse conditions.

1. INTRODUCTION

During the production of speech vibration of vocal folds is the major excitation, which results in the production of voiced 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. Even though finding the period of a perfectly periodic signal is straight forward, measuring the period of speech is difficult due to the variation in the signal value from one glottal cycle to the other. The speech signal may also be corrupted by the presence of adverse environmental conditions. Due to these factors the extraction of pitch is the object of research over the past several decades [1–4]. Apart from this, information about pitch is important in several applications like voiced/unvoiced classification, speaker recognition, speech enhancement and prosodic manipulation [5, 6].

All the algorithms proposed in the literature may be broadly classified into three categories, namely, algorithms using time domain properties, algorithms using frequency domain properties and algorithms using both time and frequency domain properties of the speech signals. The algorithms based on the time domain properties operate directly on the speech signal and the measurements made most often are peak and valley, zero-crossings and autocorrelation. The basic assumption is that if a quasiperiodic signal has been suitably processed to minimize the effects of the formant structure, then simple time domain measurement will provide good estimate of the pitch period. The group of algorithms based on the frequency domain properties of the speech signal assume that if the signal is periodic in the time domain, then the frequency spectrum of the signal will consists of a series of impulses at the fundamental frequency and its harmonics. Thus simple measurements can be made on the frequency spectrum of the signal or a nonlinearly transformed version of it, as in the cepstral method [1], to estimate the pitch period of the signal. In the third category, the algorithms use the properties of both time and frequency domains for pitch estimation. For instance, as in the case of Simple Inverse Filtering Technique (SIFT) algorithm [2], frequency domain approach may be used to spectrally flatten the time domain signal and then an autocorrelation measurement is used for the estimation of pitch.

Even though several algorithms have been proposed in the literature for the extraction of pitch, cepstrum and SIFT algorithms are simple and efficient methods for the estimation of pitch. However, the performance of these methods deteriorate significantly under degraded conditions. In this paper a method based on the GC events information is proposed for the extraction of pitch in adverse conditions. The paper is organized as follows: In Section 2 extraction of GC events from the speech signal is discussed. A method for the extraction of pitch using the GC event information is proposed in Section 3. Section 4 discusses about the Extraction of pitch in adverse conditions. Section 5 gives the performance evaluation of the proposed method. The paper is concluded and scope for future work is discussed in Section 6.

2. EXTRACTION OF GC EVENT INFORMATION

One approach to derive information about the GC events from the speech signal is by the Linear Prediction (LP) analysis [7]. In the LP analysis each sample is predicted as a linear combination of past p samples, where p is the order of prediction [7]. The predicted sample of $s(n)$ is given by

$$\hat{s}(n) = -\sum_{k=1}^{p} a_k s(n-k) \quad (1)$$

where $\{a_k\}$ are the Linear Prediction Coefficients (LPCs) computed by minimizing the squared error between the actual and the predicted sample.

The error $r(n)$ between the actual sample and the predicted sample is the LP residual and is given by

$$r(n) \quad = \quad s(n) + \sum_{k=1}^{p} a_k s(n-k) \quad (2)$$

A segment of voiced speech and its LP residual are shown in Fig.1(a) and (b), respectively. The instant at which closure of vocal folds occurs within a pitch period is defined as the GC event. As GC event is the place of significant excitation, large error is associated with GC event in the LP residual. However it is difficult to directly use the LP residual for the detection of GC events due to the occurrence of peaks of either polarity around the GC events [8]. Hence a feature Hilbert envelope of the LP residual is used [8], which is defined as

$$h(n) = \sqrt{r^2(n) + r_h^2(n)} \quad (3)$$

where $r_h(n)$ is the Hilbert transform of $r(n)$ and is computed as

$$r_h(n) = IDFT[jDFT[r(n)]] \quad (4)$$

The DFT and $IDFT$ are discrete and inverse discrete Fourier transforms, respectively. The Hilbert envelope for the LP residual in Fig.1(b) is shown in Fig.1(c). The peaks in the Hilbert envelope of the LP residual indicate the approximate location of the GC events.

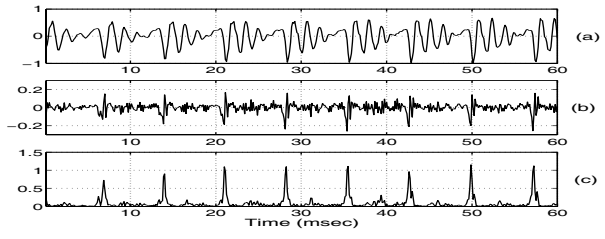

Fig. 1. (a) A segment of voiced speech. (b) LP residual. (c) Hilbert envelope of the LP residual.

3. PITCH EXTRACTION USING GC EVENT INFORMATION

In this study Hilbert envelope of the LP residual is used as the representation of GC events for the extraction of pitch. One approach for the extraction of pitch is to detect the peaks at the GC events in the Hilbert envelope of the LP residual and take the time difference of successive peaks, which gives information about the pitch period (T_0). However peak picking in degraded conditions is a difficult task. Alternatively a more convenient approach is to use autocorrelation of the Hilbert envelope of the LP residual.

Although autocorrelation of voiced speech segment generally displays a peak at the pitch period, autocorrelation peaks due to the detailed formant structure of the signal are also often present. One way to minimize the effects of formants is to low pass filter the speech signal using a cut off frequency of 900 Hz. However, low pass filtering may reduce the resolution of the extracted pitch values. Alternatively, the information related to formants may be removed by passing the speech signal through an inverse filter, whose parameters are derived from the LP analysis. The output of the inverse filter is the LP residual and autocorrelation of a segment of LP residual shows peak at the pitch period and there will not be any peaks corresponding to the formants. The peak corresponding to the pitch period may be detected using some heuristics based on the nature of speech signal. The ease with which this peak can be detected depends on the prominence of the peak, which in turn depends on the phase values around the GC events. However this ambiguity may be minimized by using the Hilbert envelope of the LP residual as the feature for autocorrelation. This is because correlation among the samples of the Hilbert envelope of the LP residual around the GC events is high compared to the corresponding samples in the LP residual.

A segment of degraded speech, its LP residual, Hilbert envelope of the LP residual and the corresponding autocorrelation sequences are shown in Fig.2. Since Hilbert envelope is unipolar in nature mean subtraction is performed before the autocorrelation. It can be observed from this figure that it is easy to process the peak corresponding to the pitch in the autocorrelation sequence of the Hilbert envelope of the LP residual compared to the peaks in the autocorrelation of the LP residual. This property of the Hilbert envelope of the LP residual, that is, high correlation among the samples around the GC events is exploited for extraction of pitch in this study.

The LP residual is computed from the differenced speech (sampled at 8 kHz) by the LP analysis using frame size of 20 msec, frame shift of 5 msec and LP order of 12. The Hilbert envelope of the LP residual is computed and is considered in frames of 20 msec with shift of 5 msec for the extraction of pitch. For each frame of the Hilbert envelope of the

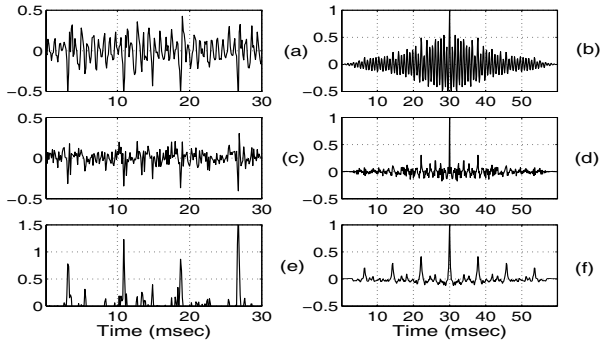

Fig. 2. Segment of (a) degraded speech, (c) LP residual, (e) Hilbert envelope of the LP residual and their respective autocorrelation sequences ((b), (d), (f)).

LP residual its mean is subtracted and the autocorrelation is computed. In the autocorrelation sequence, the first major peak after the center peak in the range of 2.5 to 12.5 msec is detected. The distance of the first major peak from the center peak is noted as the pitch period. Similarly pitch from the previous and next frames are computed. If the present frame pitch is same as either the previous or next frame with a tolerance of $\pm$ 0.25 msec (2 samples at 8 kHz), then the pitch value is retained for the next stage validation, else it is reset to zero.

Another property of the Hilbert envelope of the LP residual is that in case of voiced speech the behavior of the samples around the first major peak of the autocorrelation sequence will be similar, especially in adjacent frames. This similarity can be measured by comparing samples in a region of 2 msec on either side of the first major peak of the present frame, with the samples from the previous/next frame. This is measured using the correlation coefficient (c), which is given by

$$c = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2}\sqrt{\sum(y-\bar{y})^2}} \qquad (5)$$

where x and y represents samples around the first major peak in the current frame and previous/next frame, respectively and $\bar{x}$ and $\bar{y}$ represents their mean. Ideally the correlation coefficient will be 1 for the same frame. If it is more than 0.7 when it is computed with respect to the previous/next frame, then the pitch value of the present frame is accepted, else it is reset to zero.

In isolated utterances of speech there will not be much variation in the pitch values. However, in continuous speech pitch may vary by a large amount depending on the context. Further in continuous speech for deciding the extracted peak from the autocorrelation sequence as pitch or not, mostly energy is used as the feature and a decision is made using a threshold on the energy level. However, energy is a poor feature in case of low voiced regions and also in degraded conditions. Hence the pitch extraction algorithm should be able to handle all these factors in case of continuous speech. A segment of continuous speech and the extracted pitch contours by the proposed method and the SIFT algorithm are shown in Fig.3. The pitch contour by the proposed algorithm preserves the variations in the pitch values better compared to the SIFT algorithm.

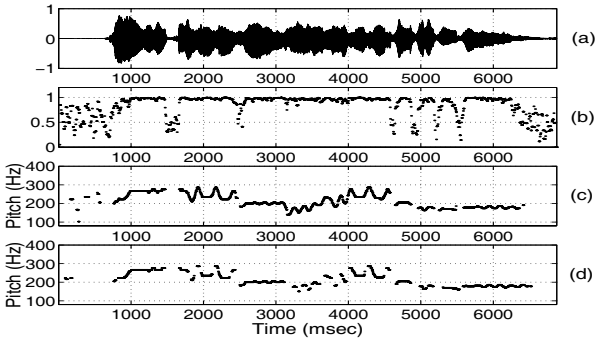

Fig. 3. (a) Segment of continuous speech. (b) Correlation coefficient values. Pitch values from the (c) proposed algorithm and (d) SIFT algorithm.

4. PITCH IN ADVERSE CONDITIONS

Practically there will be situations in which the speech signal may be degraded by the presence of additive noise, reverberation and channel noise. In such conditions humans are still able to perceive speech. This shows that pitch information is present in the degraded signal. Therefore processing speech to extract pitch in such conditions is a challenging task. Even in degraded conditions since the GC event information is available in the Hilbert envelope of the LP residual, the autocorrelation analysis of the same will give better information about the pitch. This is evident in the autocorrelation sequences shown in Fig.2.

A segment of continuous speech degraded by additive noise and reverberation along with the pitch contours extracted by the proposed method and the SIFT algorithm are shown in Fig.4. The pitch contour extracted by the proposed method appears to be more smoother compared to the SIFT algorithm. The main effect of degradation is on the phase values. This may affect the prominence of peak in the autocorrelation sequence. However, in the proposed method as the effect of phase is minimized, the pitch extraction appears to be better. A segment of speech signal collected over severely degraded channel condition and the pitch contours extracted by the proposed method and the SIFT algorithm are shown in Fig.5. The pitch contour shape is smoother in case of the proposed method. This example illustrates the robustness of the proposed method.

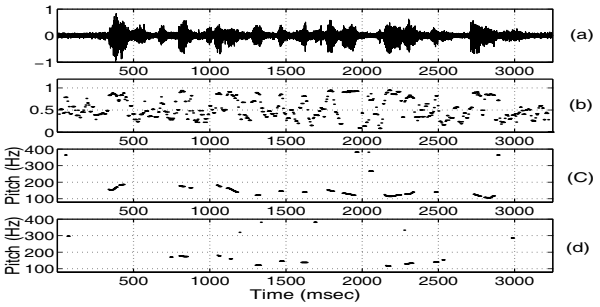

Fig. 4. (a) Segment of degraded speech affected by background noise and reverberation. (b) Correlation coefficient values. Pitch values from the (c) proposed algorithm and (d) SIFT algorithm.

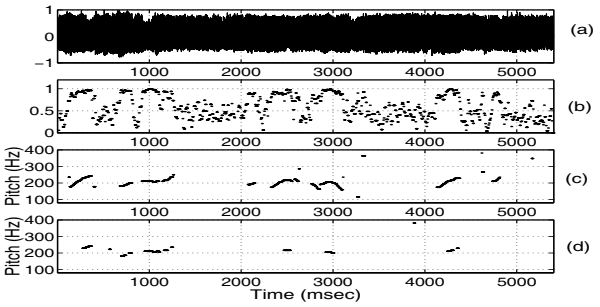

Fig. 5. (a) Segment of speech collected over a severe degraded channel. (b) Correlation coefficient values. Pitch values from the (c) proposed algorithm and (d) SIFT algorithm.

5. PERFORMANCE EVALUATION

The performance of the proposed method is evaluated by considering a speech utterance from the TIMIT database, spoken by a female speaker. Additive noise was added at different levels (3 dB and 0 dB) and the pitch contours extracted by the proposed method and the SIFT algorithm are shown in Fig.6. The proposed method seems to be providing better performance even at low SNR value like 3 dB and shows a graceful degradation at 0 dB.

6. CONCLUSIONS

In this paper a method for extraction of pitch in adverse conditions was proposed using the information about the GC events. Hilbert envelope of the LP residual was used as a representation for the GC events. The pitch was extracted by performing autocorrelation analysis on the mean subtracted Hilbert envelope frames. The performance of the proposed algorithm was tested on the data collected over real and severely degraded channels. The proposed method shows better performance compared to the SIFT algorithm.

The GC events in the Hilbert envelope of LP residual is only an approximate information. Better methods may be developed for the accurate detection of GC events and use them for the extraction of pitch, which may improve the performance of the proposed method.

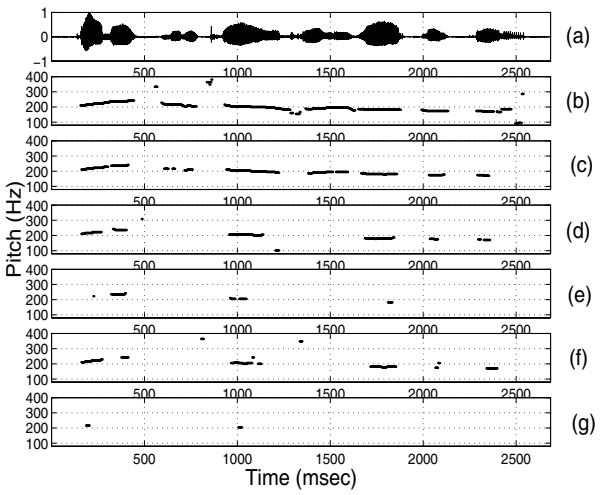

Fig. 6. (a) Speech utterance of female speaker taken from TIMIT database. Pitch contours by the proposed method for (b) clean speech, (d) speech at 3 dB SNR and (f) speech at 0 dB SNR. Pitch contours by the SIFT algorithm for (c) clean speech, (e) speech at 3 dB SNR and (g) speech at 0 dB SNR.

7. REFERENCES

[1] A. M. Noll, "Cepstrum pitch determination," *J. Acoust. Soc. Amer.*, vol. 41, pp. 293–309, 1967.

[2] J. D. Markel, "The SIFT algorithm for fundamental frequency estimation," *IEEE Trans. Audio Electrocoust.*, vol. AU-20, pp. 367–377, Dec. 1972.

[3] M. J. Ross, H. L. Shaffer, A. Cohen, R. Freudberg, and H. J. Manley, "Average magnitude difference function pitch extractor," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-22, pp. 353–362, Oct. 1974.

[4] H. Quast, O. Schreiner, and M. R. Schroeder, "Robust pitch tracking in the car envrionment," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. I, (Orlando, FL, USA), pp. 353–356, May 2002.

[5] B. Yegnanarayana, S. R. M. Prasanna, and K. S. Rao, "Speech enhancement using excitation source information," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. I, (Orlando, FL, USA), pp. 541–544, May 2002.

[6] K. S. Rao and B. Yegnanarayana, "Prosodic manipulation using instants of significant excitation," in *Proc. IEEE Int. Conf. Mul., Expo*, vol. I, (Baltimore, MD, USA), pp. 389–392, July 2003.

[7] J. Makhoul, "Linear prediction: A tutorial review," *Proc. IEEE*, vol. 63, pp. 561–580, Apr. 1975.

[8] T. V. Ananthapadmanabha and B. Yegnanarayana, "Epoch extraction from linear prediction residual for identification of closed glottis interval," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-27, pp. 309–319, Aug. 1979.