STUDY OF ROBUSTNESS OF ZERO FREQUENCY RESONATOR METHOD FOR EXTRACTION OF FUNDAMENTAL FREQUENCY

B. Yegnanarayana[1], *S. R. M. Prasanna*[2] *and S. Guruprasad*[1]

[1]International Institute of Information Technology Hyderabad, India
[2]Indian Institute of Technology Guwahati, India
yegna@iiit.ac.in, prasanna@iitg.ernet.in, guru@research.iiit.ac.in

ABSTRACT

The objective of this work is to develop and study the robustness of the zero frequency resonator (ZFR) based method for extraction of the fundamental frequency (F_0) of speech signals. The proposed ZFR method for estimating F_0 consists of zero frequency filtering of the Hilbert envelope (HE) of the linear prediction (LP) residual of speech signal, followed by short-term spectrum analysis of the filtered output. The robustness of the proposed method is tested using speech signals collected in practical environments like distant, reverberant, telephone, mobile and multispeaker. Experimental results show that the proposed ZFR method estimates F_0 in majority of the cases.

Index Terms— ZFR, F_0, Hilbert envelope, short-term spectrum analysis, robustness

1. INTRODUCTION

The objective of this study is to examine the robustness of the recently proposed [1] zero frequency resonator (ZFR) method for extracting the fundamental frequency (F_0) from speech signals collected in practical environments. The environments include the effect of distance, reverberation, telephone channel, mobile channel and presence of multiple speakers on the collected speech signal. These cases are considered to determine the robustness of the method for speech collected in real environments, and not in simulated conditions. Since these are the mostly used practical environments in daily life for speech communication, a method for estimating F_0 is essential. Such a solution provides a good tool for processing speech signals from practical environments.

The classical approaches for estimating F_0 include cepstrum and autocorrelation methods [2], [3]. These algorithms work very well for clean speech collected in a controlled environment. Several refinements have been proposed for F_0 estimation from degraded speech collected in practical environments. These include autocorrelation pitch detector and voicing decision with confidence measures for noise corrupted speech [4], weighted autocorrelation for pitch extraction from noisy speech [5], extraction of pitch in adverse conditions using autocorrelation of the Hilbert envelope (HE) of linear prediction (LP) residual [6], evaluation of different pitch detection algorithms in adverse conditions [7], evaluation of pitch detection algorithms under real conditions [8], and derivation of the instantaneous F_0 from two-speaker, two-microphone mixed signals using zero frequency filtering [9]. The objective of this work is to further study the robustness of the zero frequency filtering approach under practical environments.

In [1], it was shown that extraction of the instantaneous F_0 is robust to different types of additive noises. Performance of the method may degrade if spurious impulse-like noise occurs in addition to epochs, which may be caused by degradation such as reverberation. Some errors were corrected using the knowledge of epochs derived from the HE of the speech signal, as the HE has strong peaks highlighting the regions around the glottal closure instant (GCI) compared to other parts of the glottal cycle. The HE of the linear prediction (LP) residual was also used to separate the epochs due to individual speakers from two-speaker mixed signals collected from two channels [9]. In this case the HE due to one of the speakers is emphasized by combining the HEs of the LP residuals of the two mixed signals after compensating for the time delays. In fact autocorrelation of the HE of the LP residual was shown to be effective to estimate the pitch period from single channel speech data, even for degraded speech signals [6].

Robustness of estimation of F_0 using zero frequency filtering can be improved further by processing the zero frequency filtered signal. The Fourier transform of short segment (2-3 pitch periods) of the zero frequency filtered signal shows a strong peak at F_0 most of the time, even though the instants of zero crossings of the filtered signal may have been shifted from the true epoch locations due to impulse-like degradation [10]. The frequency corresponding to the peak is used to filter the zero frequency filtered signal for deriving accurate locations of epochs from distant speech signal, which is a low signal to noise ratio (SNR) signal. This paper develops F_0 estimation methods given in [6] and [10], and examines the robustness of the this method in the estimation of F_0 from speech degraded in several practical environments. The method uses a procedure similar to the one described in [10], but exploits the sequential constraint also. It is demonstrated that the proposed method is robust for different types of degradations.

The rest of the paper is organized as follows: Section 2 describes a method which uses the properties of the HE of LP residual, zero frequency filtering of the HE, the presence of strong spectral peak at F_0 in the Fourier transform of the zero frequency filtered signal, and continuity of F_0 in voiced regions of speech. Section 3 discusses the performance of the method for different types of degraded signals. Section 4 summarizes the ideas presented in the paper.

2. PROPOSED ZFR BASED METHOD FOR ROBUST ESTIMATION OF F_0

A method for extracting the instantaneous F_0 was proposed recently using the output of an ideal zero frequency digital resonator for input speech [1]. The method exploits the impulse-like excitation characteristics during the production of voiced speech [11]. Within each glottal cycle, the excitation of the vocal tract system is impulse-like around the GCI, due to rapid closure at the glottis. The region around

the GCI corresponds to the high SNR region due to strong excitation, and also due to decay of the resonances of vocal tract system within each cycle. An impulse has energy spread over all frequencies uniformly, whereas the frequency response of the vocal tract system has significant energy only over the formant regions. Hence it is possible to highlight the characteristics of the impulse-like excitation by passing the speech signal through an ideal digital resonator located at zero frequency. The output of the zero frequency resonator has approximately a polynomial growth, but the fluctuations in the output capture the characteristics of the sequence of impulse-like excitation. These fluctuations can be highlighted by removing the trend in the output signal. The trend removal is accomplished by subtracting the mean over an interval of about $1.5T_0$, where T_0 is the average pitch period. Note that the size of the window for trend removal is not critical. The trend removed signal is called the ZFR signal or filtered signal. The locations of positive-to-negative zero crossings in the filtered signal are hypothesized as epochs or GCIs during voiced speech, and the reciprocal of the interval (T_0) between successive epochs is hypothesized as the instantaneous fundamental frequency ($F_0 = 1/T_0$).

In [9], extraction of the instantaneous F_0 from two-speaker and two-microphone speech is accomplished by separating the impulse-like excitation characteristics of each speaker. By estimating the delay of each speaker at the two microphones, the delay-compensated signal is derived, which enhances the signal of one speaker relative to that of the other. Epoch extraction directly from the delay-compensated speech signals for each speaker does not produce the epochs at the desired locations in the zero frequency filtered signal, as the impulse-like excitations of the other speaker are still present in these compensated signals. This problem is overcome using the HE of the LP residual. The effect of impulse-like excitations due to the other speaker is reduced significantly using the delay compensated HEs of the LP residuals. The pitch period of each speaker is then estimated by the zero frequency filtering of the speaker-specific delay-compensated HE of the LP residual. The results show good agreement with the reference. The robustness of this method can be further improved to make it suitable for many degradations in practical environments by processing the zero frequency filtered signal in the frequency domain as described next.

In [10], the short-time spectrum of the zero frequency filtered signal of speech is computed using a frame size of 25 ms, and a frame shift of 5 ms. The frequency corresponding to the maximum value in the magnitude of the short-time Fourier transform (STFT) of the filtered signal corresponds to the fundamental frequency (F_0). The values of F_0 derived from successive segments are filtered using a 7-point median filter, to eliminate spurious values. The F_0 is found to be robust compared to direct estimation from the zero frequency filtered output.

The robustness of the zero frequency filtering can be further improved by combining the merits of the HE of LP residual and short-term spectrum analysis of zero frequency filtered output. The proposed method is therefore based on the short-term spectrum analysis of the zero frequency filtered signal of the Hilbert envelope of LP residual of the speech signal. The proposed method involves the following steps:

- Difference the speech signal (sampled at 8 kHz), $x[n] = s[n] - s[n-1]$
- Compute 12^{th} order LP residual $e[n]$ of $s[n]$ using a frame size of 25 ms and a frame shift of 5 ms.
- Compute the Hilbert envelope (HE) of the LP residual.

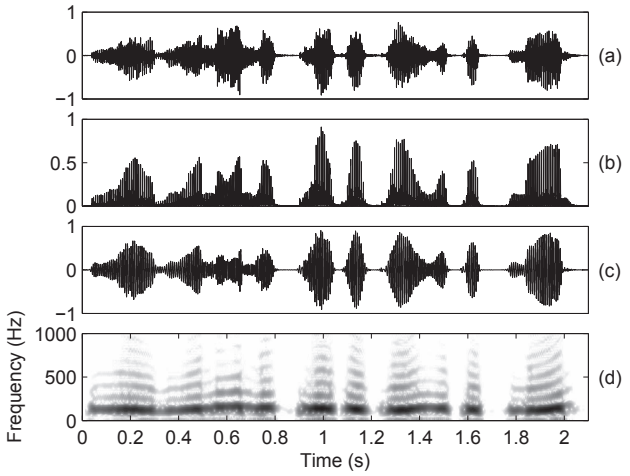

Fig. 1. (a) Speech signal. (b) HE of LP residual. (c) Zero frequency filtered signal derived from HE. (d) Spectrogram of the filtered signal.

- Derive the zero frequency filtered output of the HE of the LP residual.
- Compute the short-time spectrum of zero frequency filtered signal (frame size = 25 ms, frame shift = 1 ms), and determine the frequency location (F_0) of the strongest peak.
- Smooth the F_0 contour using neighborhood constraints.

The presence of pitch information in the zero frequency filtered signal derived from HE is illustrated in Fig. 1. The spectrogram of the filtered signal, shown in Fig. 1(d)), indicates that F_0 is the most dominant component in the short-time spectrum of the filtered signal.

3. ESTIMATION OF F_0 CONTOURS FOR SPEECH COLLECTED IN DIFFERENT ENVIRONMENTS

In this section, we illustrate the robustness of the proposed method for speech signals collected in different environments. Figure 2(a) shows the speech signal collected by close-speaking microphone and Fig. 2(b) shows the speech signal collected at a distance of 6 ft from the speaker. Speech signals are collected in a live room where the SNR is low due to distance, but the reverberation is not significant. Figure 2(c) is the instantaneous F_0 of the clean signal derived using the method proposed in [1], and which can be used as a reference. Figure 2(d) shows the HE of the LP residual of the distant speech signal in Fig. 2(b), and Fig. 2(e) shows the zero frequency filtered signal from the HE of the LP residual. The F_0 contour of Fig. 2(b) derived using the proposed method is shown in Fig. 2(f). Figure 2(g) shows the F_0 contour obtained from the filtered signal derived directly from the distant speech signal (Fig. 2(b)). The results show that the proposed method yields the desired F_0 even from the speech degraded due to distance.

Figure 3 shows the results for reverberant speech. Figure 3(a) is the clean signal and Fig. 3(b) is the reverberant signal. Figure 3(c) is the instantaneous F_0 of the clean signal derived using the method proposed in [1]. Figure 3(d) shows the F_0 contour of the reverberant speech, obtained using the proposed method. Figure 3(e) shows the F_0 contour of the reverberant speech, obtained from the filtered signal which derived directly from the reverberant signal.

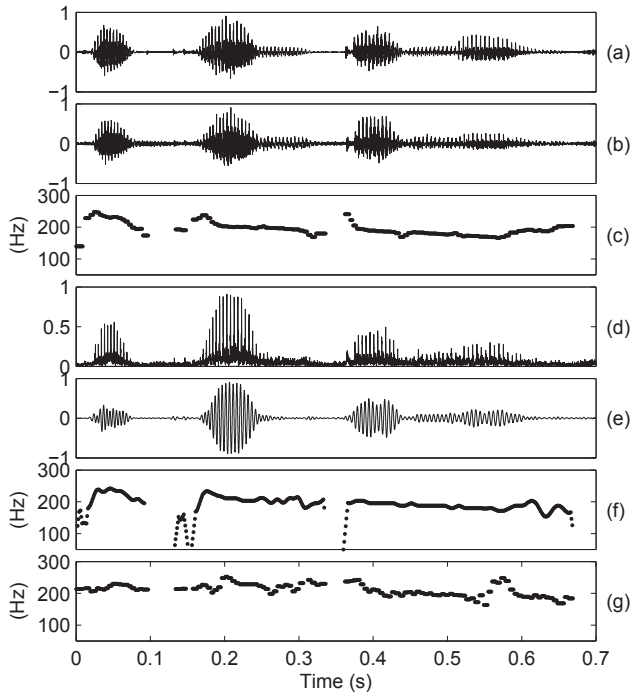

Fig. 2. (a) Close-speaking speech signal. (b) The corresponding distant speech signal. (c) F_0 contour derived from (a). (d) HE of LP residual derived from (b). (e) Zero frequency filtered signal derived from HE. (f) F_0 contour derived from (e) using the proposed method. (g) F_0 contour derived from (b) using ZFR analysis directly on the signal.

Comparison of Figs. 3(d) and (e) shows that the derivation of zero frequency filtered signal from the HE of LP residual is advantageous, compared to its derivation from the reverberant signal itself. This is because of the emphasis of impulse-like excitation in HE of LP residual. The results also show that for reverberant case, the contour will be different from the reference F_0 contour, due to impulse-like degradation caused by reverberation.

Figure 4 illustrates the extraction of F_0 from speech signal collected over telephone channel. It is interesting to note that the proposed method is able to extract F_0 contour (Fig. 4(c)), even though the original signal is bandpass filtered to remove the low-frequency component up to 300 Hz. This is because, the HE of the LP residual shows large amplitude peaks around the glottal closure instants. It may be noted that the F_0 contour obtained by using ZFR analysis directly on the signal (shown in Fig. 4(c)) is less accurate than the F_0 contour obtained using the proposed method. This is due to the fact that spectral components in the neighbourhood of zero frequency are filtered out in the telephone channel speech.

Speech collected at the output of a cellphone channel is affected by the processing during coding and decoding of speech. While spectral characteristics are affected by quantization of parameters, the characteristics of excitation source are robust. Figure 5 illustrates the extraction of F_0 for speech signal collected over cellphone channel. In this case too, the zero frequency filtered signal derived from the HE of LP residual gives more robust pitch information (Fig. 5(c)) compared to the zero frequency filtered signal derived directly from the speech signal (Fig. 5(d)).

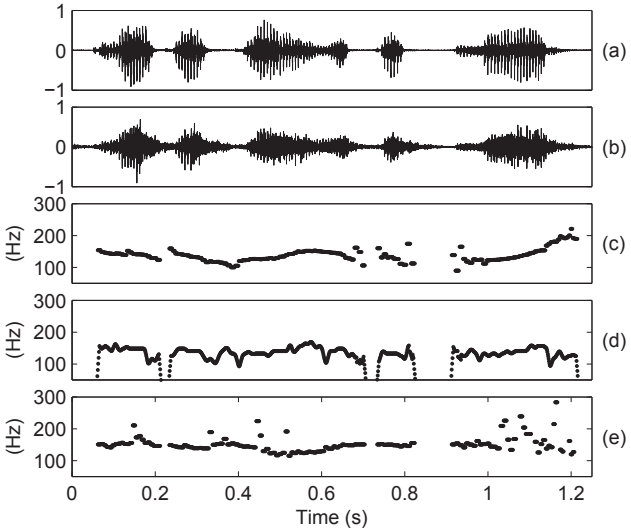

Fig. 3. (a) Close-speaking speech signal. (b) The corresponding distant speech signal. (c) F_0 contour derived from (a). (d) F_0 contour derived from (b) using the proposed method. (g) F_0 contour derived from (b) by using ZFR analysis directly on the signal.

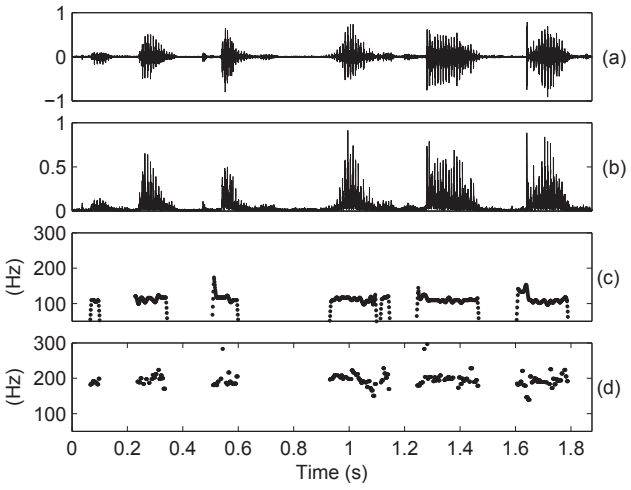

Fig. 4. (a) Telephone channel speech signal. (b) Hilbert envelope of LP residual. (c) F_0 contour derived using proposed method. (d) F_0 contour derived using ZFR analysis directly on the signal.

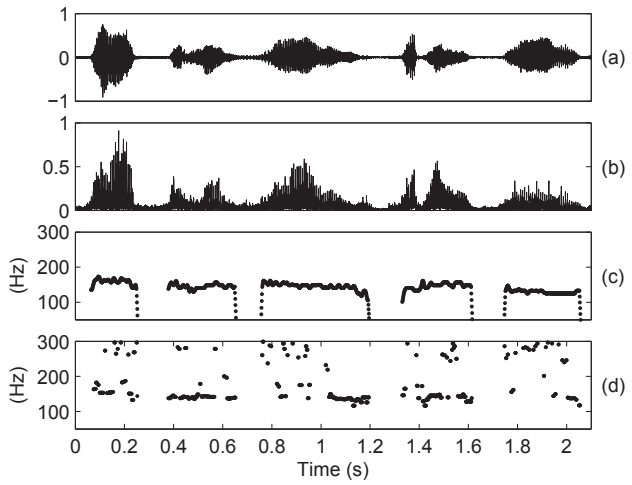

Fig. 5. (a) Cellphone channel speech signal. (b) Hilbert envelope of LP residual. (c) F_0 contour derived using proposed method. (d) F_0 contour derived using ZFR analysis directly on the signal.

The proposed method of F_0 extraction is particularly useful in the context of multispeaker speech collected using two or more microphones. In such cases, separation of speech signals of two speakers on the basis of spectral characteristics is a difficult task. By contrast, the point property of the impulse-like excitation is useful not only for estimating the time delays corresponding to the speakers, but also for separating the excitation components of different speakers. Figures 6(a) and (b) show the signals collected using two spatially separated microphones in a two-speaker case. The HE of LP residual is derived from each of the signals. Time delays corresponding to the two speakers are estimated using crosscorrelation of corresponding segments of HE from the two microphones. Delay-compensated Hilbert envelopes are used for deriving zero frequency filtered signals. Figures 6(c) and (d) show the F_0 contours obtained from the filtered signals, using the proposed method. The delay-compensated Hilbert envelope corresponding to a speaker mostly consists of impulse-like excitations of that speaker, while the impulse-like excitations due to other speakers are deemphasized.

4. SUMMARY AND CONCLUSIONS

In this paper we have shown the robustness of the ZFR based method for deriving the F_0 contour for signals collected in several practical environments. The robustness is due to the use of the Hilbert envelope (HE) of the LP residual, which highlights the impulse-like nature of excitation of voiced speech. The filtered signal obtained from the HE of the LP residual was further processed using short-time spectrum analysis of the filtered signal. Except in the case of reverberant signals, the method gives fairly accurate F_0 contours for all the types of environments studied in this paper. However, processing reverberant speech and multispeaker speech is still a challenge due to the presence of impulse-like degradation in the signals.

5. REFERENCES

[1] B. Yegnanarayana and K. S. R. Murty, "Event-based instantaneous fundamental frequency estimation from speech signals," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17(4), pp. 614–624, May 2009.

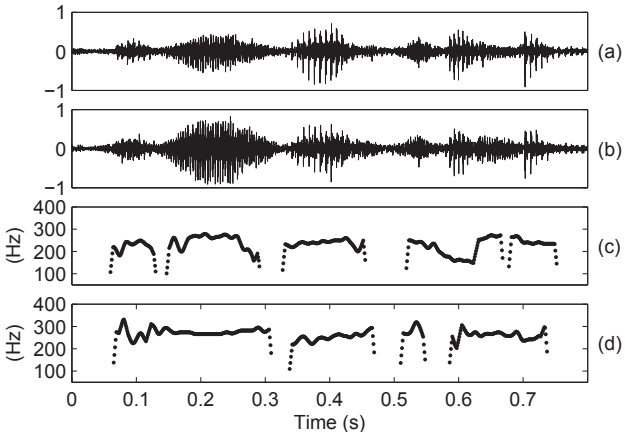

Fig. 6. (a) Two-speaker speech collected using mic-1. (b) Two-speaker speech collected using mic-2. (c) F_0 contour of speaker-1, derived from delay-compensated HE of speaker-1. (d) F_0 contour of speaker-1, derived from delay-compensated HE of speaker-1.

[2] A. M. Noll, "Cepstrum pitch determination," *J. Acoust. Soc. Amer.*, vol. 41(2), pp. 293–309, Feb. 1967.

[3] J. D. Markel, "The SIFT algorithm for fundamental frequency estimation," *IEEE Trans. Audio Electroacoust.*, vol. 20(5), pp. 367–377, Dec. 1972.

[4] D. Krubsack and R. Niederjohn, "An autocorrelation pitch detector and voicing decision with confidence measures developed for noisecorrupted speech," *IEEE Trans. Signal Process.*, vol. 39(2), pp. 319–329, Feb. 1991.

[5] T. Shimamura and H. Kobayashi, "Weighted autocorrelation for pitch extraction of noisy speech," *IEEE Trans. Speech Audio Process.*, vol. 9(7), pp. 727–730, Oct. 2001.

[6] S. R. M. Prasanna and B. Yegnanarayana, "Extraction of pitch in adverse conditions," in *Proc. Int. Conf. Acoust. Speech Signal Process.*, 2004, vol. 1, pp. 109–112.

[7] B. Kotnik, H. Hoge, and Z. Kacic, "Evaluation of pitch detection algorithms in adverse conditions," in *Proc. Third Int. Conf. Speech Prosody*, 2006, pp. 149–152.

[8] I. Luengo, I. Saratxaga, E. Navas, I. Hernaez, J. Sanchez, and I. Sainz, "Evaluation of pitch detection algorithms under real conditions," in *Int. Conf. Acoust. Speech Signal Process.*, 2007, vol. 4, pp. 1057–1060.

[9] B. Yegnanarayana and S. R. M. Prasanna, "Analysis of instantaneous f_0 contours from two speakers mixed signal using zero frequency filtering," in *Int. Conf. Acoust. Speech Signal Process.*, 2010, pp. 5074–5077.

[10] S. Guruprasad and B. Yegnanarayana, "Extraction of fundamental frequency from distant speech signals," *Accepted for publication in IEEE Trans. Audio, Speech, Lang. Process.*, 2010.

[11] K. S. R. Murty and B. Yegnanarayana, "Epoch extraction from speech signals," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16(8), pp. 1602–1613, Nov. 2008.