Analysis of instantaneous F_0 contours from two speakers mixed signal using zero frequency filtering

B. Yegnanarayana[1] and S. R. Mahadeva Prasanna[2]
[1] *International Institute of Information Technology Hyderabad, AP, India,*
[2] *Indian Institute of Technology Guwahati, Assam, India,*
Email: yegna@iiit.ac.in, prasanna@iitg.ernet.in

***Abstract*—Instantaneous fundamental frequency (F_0) in voiced speech can be obtained from the sequence of epochs corresponding to the instants of significant excitation. The epoch sequence can be derived using the recently proposed epoch extraction method based on zero frequency filtering. The epoch extraction method is robust against additive noise degradation. But in a multispeaker mixed signal, the degradation is caused due to overlapping impulse-like excitations of two or more speakers. The feasibility of extracting the instantaneous F_0 contours from the two speaker mixed signal using zero frequency filtering is studied in this paper. The present study is based on deriving speaker-specific Hilbert Envelope (HE) signal which emphasizes peaks due to impulse-like excitation of one speaker and suppresses peaks due to other speaker. The epochs from this speaker-specific signal are obtained using the approach based on zero frequency filtering. The results of the proposed method is demonstrated for three different cases of mixed signals of two speakers data.**

***Index Terms*—Two speakers, instantaneous F_0, epochs, zero frequency filtering**

I. INTRODUCTION

The instantaneous fundamental frequency (F_0) of voiced speech corresponds to the reciprocal of the period of each glottal cycle. Due to quasiperiodic nature of the glottal vibration, the short term average period over a few cycles is perceived as pitch period, or the reciprocal of it as pitch frequency. Thus pitch is a *perception* of the average periodicity of the glottal vibration, whereas the instantaneous F_0 is a *production* feature representing the period of each cycle of glottal vibration [1]. Methods for extraction of pitch and instantaneous F_0 from speech signal differ significantly. For example, most methods for pitch extraction rely on the property of periodicity, which requires similarity of waveforms in successive glottal cycles, and regularity (same period) of successive excitation pulses. Similarity of waveforms is captured using the autocorrelation function of a short segment of speech, where the peak amplitude at the shift (delay) of one period indicates the extent of similarity, and the location of the peak indicates pitch period [2]. These methods assume that the length of the analysis segments is greater than at least two pitch periods. Since the computation is done over segments which are not aligned pitch synchronously, the methods yield only an estimate of the average pitch period within each segment. The Hilbert Envelope (HE) of the Linear Prediction (LP) residual highlights the peaks around the instants of significant excitation in each glottal cycle [3]. The regularity, i.e., equal intervals between successive instants, of the peaks in the HE is captured using the autocorrelation function of the HE of the LP residual [4].

The periodicity property of successive glottal cycles is also affected by degradations in the collected speech signal. The pitch extraction methods fail if the speech is degraded due to the occurrence of background noise and additional impulse-like excitations within each analysis segment. Impulse-like excitations are caused by reverberation and/or multispeaker data. In multispeaker data the speech signal contains segments having speech from two or more speakers simultaneously. Hence pitch extraction from reverberant speech or multi-speaker data is a challenging task. Methods for extraction of the instantaneous F_0 depend on the property that voiced speech is a result of excitation of the vocal tract system by a sequence of impulse-like excitations, with one significant excitation in each glottal cycle [5]. The instant of significant excitation in each glottal cycle corresponds to the instant of glottal closure, also called *epoch* [5]. Due to strong excitation at each epoch, the speech signal strength is also large around the epoch compared to the strength of the signal in the rest of the glottal cycle. Thus high strength property of the signal around the epochs is exploited to derive the epoch sequence from the speech signal directly. A recent method proposes a zero frequency filtering approach for extracting the epochs [6]. Note that the sequence of epochs gives the instantaneous F_0 automatically, and there is no assumption of periodicity in the epoch extraction method. The method works well even for reasonably large additive noise.

Robustness of the epoch extraction method is affected if the degradation in the speech signal introduces additional impulse-like excitation sequences, as it happens in reverberant speech and multispeaker data. One way to deal with such degradations is to collect the speech data from several spatially distributed microphones. In such data the instantaneous F_0 of each speaker is preserved at each microphone. By compensating for the delay of the speech signal of a speaker between two microphones, the epoch sequence due to each speaker can be enhanced relative to the effects due to degrading impulses. The enhanced impulse-like excitation sequences help to extract the instantaneous F_0 contour for each speaker.

In this paper we propose a method for extracting the instantaneous F_0 contours of two speakers from the simultaneously collected speakers data at the two spatially separated microphones. The zero frequency filtering method does not work well on the delay compensated speech signals directly

due to the presence of the degrading impulse-like excitations. We therefore propose the use of delay compensated HE of the LP residual. In Section II the basic pitch extraction method and also the instantaneous F_0 method based on zero frequency filtering are briefly reviewed. The proposed method for the extraction of instantaneous F_0 contours from the two speaker data at two microphones is described in Section III. In Section IV the method is illustrated for some synthetically mixed two speaker data, and also for real two speakers data collected in a live room. Section V gives a summary of the paper and outlines some issues that need further study.

II. Methods for extraction of pitch and instantaneous F_0 from speech signal

A. Pitch extraction by autocorrelation function

The normalized autocorrelation function $r(n)$ of a short (about 30 ms) segment of speech signal $s(n)$ is computed as $r(n) = \sum_m s(m)s(m-n)$. Fig. 1(a) shows $r(n)$ for a segment of voiced speech. The location of the first largest peak after the center peak gives the average pitch period in the segment. Peak picking from the autocorrelation function is affected by the presence of formants as shown by the smaller peaks in $r(n)$. The effects of peaks due to formants is reduced by computing the normalized autocorrelation function $r_e(n)$ of the LP residual $e(n)$. Fig. 1(b) shows the $r_e(n)$ for a segment of the LP residual, computed using 10^{th} order LP analysis, where the peak at the pitch period has better resolution than in Fig. 1(a). Some of the errors due to LP analysis can be further reduced by considering the normalized autocorrelation $r_{he}(n)$of the HE of the LP residual $h(n)$ as shown in Fig. 1(c), where the regularity of the impulse helps in detecting the peak [3].

The HE of the LP residual is computed as follows:

$$h(n) = \sqrt{e^2(n) + e_h^2(n)}, \tag{1}$$

where $e_h(n)$ is the Hilbert transform of $e(n)$, and is given by

$$\begin{aligned} e_h(n) &= \text{IDFT}(-jE(\omega)) \qquad \omega > 0 \quad (2)\\ &= \text{IDFT}(jE(\omega)) \qquad \omega < 0 \quad (3)\\ E(\omega) &= \text{DFT}(e(n)) \quad (4) \end{aligned}$$

Here the DFT and IDFT refer to the discrete Fourier transform and inverse DFT, respectively.

For a segment of two speaker data the normalized autocorrelation function of the speech, LP residual and of the HE are also shown in Fig. 1(d)-1(f), where it is difficult to isolate the peaks corresponding to pitch of each speaker.

B. Instantaneous F_0 extraction by zero frequency filtering

The zero frequency filtered signal is derived by passing the differenced speech signal twice through a second order digital resonator located at 0 Hz [6]. It is equivalent to double integration of the signal. This will reduce the effects of all the formants significantly. The output signal grows/decays as a polynomial function. The signal obtained by removing the trend using a local mean subtraction is called *zero frequency filtered signal*. The epochs are obtained at the negative to positive zero crossings in the filtered signal. The following are the steps in the epoch extraction method using the zero frequency filter:

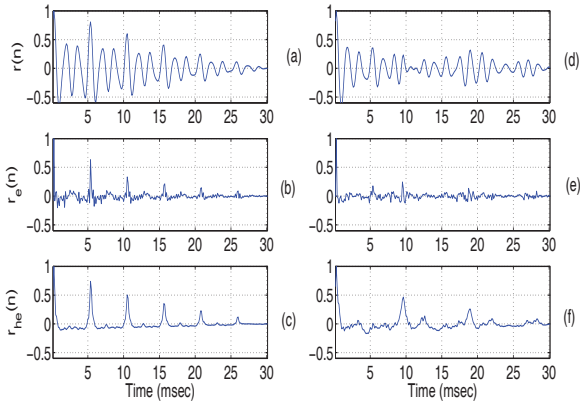

Fig. 1. Autocorrelation function of 30 msec segments of voiced speech, its 10^{th} order LP residual and HE of LP residual for one speaker ((a)-(c)) and two speaker ((d)-(f)) cases, respectively.

1) Difference the signal

$$x(n) = s(n) - s(n-1) \tag{5}$$

2) Compute the output of cascade of two ideal digital resonators at 0 Hz.

$$y(n) = \sum_{k=1}^{4} a_k y(n-k) + x(n) \tag{6}$$

where, $a_1 = -4$, $a_2 = 6$, $a_3 = -4$ and $a_4 = 1$

3) Remove the trend

$$\hat{y}(n) = y(n) - \bar{y}(n) \tag{7}$$

$$\bar{y}(n) = \frac{1}{2N+1} \sum_{n=-N}^{N} y(n). \tag{8}$$

Here $2N+1$ corresponds to the size of the window used for computing the local mean. The choice of the window size is not very critical. Normally the average pitch period computed over a long segment of speech is used as the size of the window.

Fig. 2 shows the signals at various stages of computation of epochs by zero frequency filtering approach. The reciprocal of the interval between successive epochs gives the instantaneous F_0. This method of extraction of epochs is robust against degradation due to additive noise. But when the degradation contains significant impulse-like excitation components as in the case of two speakers data, the epoch locations of both the speakers are affected as shown in Fig. 2(f). In this case the epochs of only the dominating speaker are detected. It is therefore necessary to separate the effects due to interfering impulse-like excitations in order to extract the instantaneous F_0 contours of each speaker. This is accomplished by processing the two speakers data collected at two spatially separated microphones.

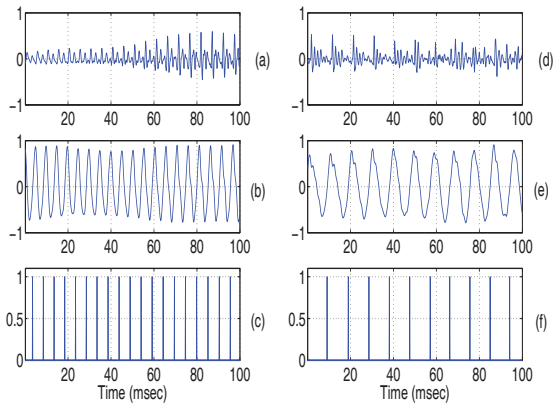

Fig. 2. Illustration of zero frequency filtering approach for epoch extraction for one speaker ((a)-(c)) and two speakers ((d)-(f)) cases, respectively. (a) and (d) speech. (b) and (e) zero frequency filtered signal. (c) and (f) epochs.

III. A Method for Extracting Instantaneous F_0 from Two Speakers Data

To extract the instantaneous F_0 from two speakers data, it is necessary to separate the impulse-like excitation characteristics of each speaker. This is accomplished by using the data collected at two spatially separated microphones. Each microphone collects a mixture of the two speakers data, but the mixing takes place with different delays. By estimating the delay of each speaker at the two microphones, the delay compensated signal is derived, which enhances one speaker's data relative to the other. The delay is estimated using the crosscorrelation function of the HEs of the LP residuals of the signals at the two microphones [7]. The peaks in the crosscorrelation function give the desired delays in samples.

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 can be overcome by a method in which the effect of impulse-like excitations due to the other speaker is reduced significantly in the delay-compensated HEs of the LP residuals. The method of deriving the new delay compensated HEs of the LP residuals is described below.

Let $h_1(n)$ and $h_2(n)$ be the HEs of the LP residuals of the two microphone signals. As can be seen from Fig. 3, there are many peaks in $h_1(n)$ and $h_2(n)$ at irregular intervals corresponding to glottal closure instants of both the speakers. The time delays between the speech signals at the two microphones due to each speaker is obtained by computing the crosscorrelation of 100 ms segments of the HE signals with a shift of 10 ms, and using the percentage of frames for each delay. The locations of the peaks in the time delay histogram correspond to the time delays (τ_1 and τ_2) due to the two speakers [8]. By aligning the Hilbert envelopes $h_1(n)$ and $h_2(n)$ after compensating for the estimated time delay τ_1 of *Speaker-1*, the epochs corresponding to *Speaker-1* will be in coherence, whereas the epochs of *Speaker-2* will be incoherent. By considering $h_{s1}(n) = min(h_1(n), h_2(n-\tau_1))$, only the HE peaks around the epochs of *Speaker-1* are retained in $h_{s1}(n)$. Note that the HE peaks at the epochs of the other speaker are suppressed. Thus we obtain a signal that retains the peaks in the HE specific to *Speaker-1*, as shown in Fig. 3(c). In a similar manner we can derive the signal that retains the peaks in the HE specific to *Speaker-2*, as shown in Fig. 3(e), where $h_{s2} = min(h_1(n), h_2(n-\tau_2))$.

It is to be emphasized at this point that despite emphasizing the speaker-specific peaks in the signals $h_{s1}(n)$ and $h_{s2}(n)$, there are still some traces of the impulse-like excitations of the other speaker in each signal. To reduce these effects further, the following new signals are derived. $f_{s1}(n) = h_{s1}(n) - \alpha h_{s2}(n)$ and $f_{s2}(n) = h_{s2}(n) - \alpha h_{s1}(n)$, where α is a small fraction of the order 0.001. The pitch period of a given speaker can be estimated by measuring the interval between two successive peaks in the speaker-specific HE of that speaker. This requires detecting the peaks in the speaker-specific HEs that have peak amplitudes. This difficult task of peak detection can be avoided by using the zero frequency filtering of the signal to derive the epoch information. Applying the zero frequency filtering operation on these new signals, we obtain the filtered signals shown in Figs. 3(d) and 3(f), corresponding to $f_{s1}(n)$ and $f_{s2}(n)$, respectively. The epochs are weighted by their strength measured by the slope around each epoch as described in [9]. The instantaneous F_0 contours are obtained only for the voiced regions in each case. The voiced regions are detected using a threshold on the strength of excitation impulse at the epochs.

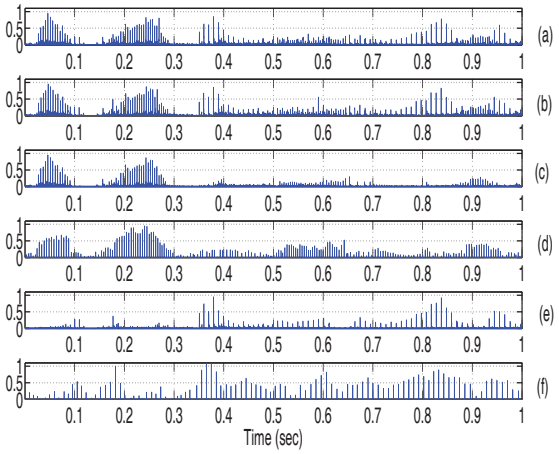

Fig. 3. Illustration of epoch extraction from delay compensated HEs of the LP residual. (a) HE of the LP residual of mixed signal from mic-1. (b) HE of the LP residual of mixed signal at mic-2. (c) Delay compensated HE ($h_{s1}(n)$) for speaker-1. (d) Epochs from $h_{s1}(n)$. (e) Delay compensated HE ($h_{s2}(n)$) for speaker-2. (f) Epochs from $h_{s2}(n)$.

IV. Illustrative Examples

In this section we show the results of applying the method of extracting the instantaneous F_0 contours for three different cases of two speakers two microphone data.

1) **Case A:** The data is a synthetic mixture of two speech signals (one male and one female) taken from TIMIT database combined using different delays. The data of

the two individual speakers is used as a reference. The instantaneous F_0 contours derived from the reference data and from the processed HEs of the LP residuals are shown in Fig. 4. The results show good agreement with the reference contours, indicating the effectiveness of the proposed method of extracting the instantaneous F_0 contours from mixed signal data.

2) **Case B:** This data is collected in a live laboratory environment where the reverberation time is about 0.5 sec. The two male speakers were reading texts simultaneously, and the data was collected using two microphones separated by a distance of 2 feet, and the speakers are about 3.5 feet from the microphones. In this case the average pitch period values of the two speakers are around 8 ms and 5 ms, respectively. The extracted instantaneous F_0 contours are shown in Figs. 5(a) and (b). Note that there is no reference data for comparison. However, the instantaneous F_0 contour is around 150 Hz for one speaker and around 200 Hz for the other speaker.
3) **Case C:** The data was collected in the same live room as in the Case B, but the two speakers are one male and the other female. The extracted instantaneous F_0 contours are shown in Fig. 5(c) and 5(d). The instantaneous F_0 contour is around 110 Hz for one speaker and around 250 Hz for the other speaker. These results demonstrate the effectiveness of proposed method in extracting the instantaneous F_0 contours of each speaker.

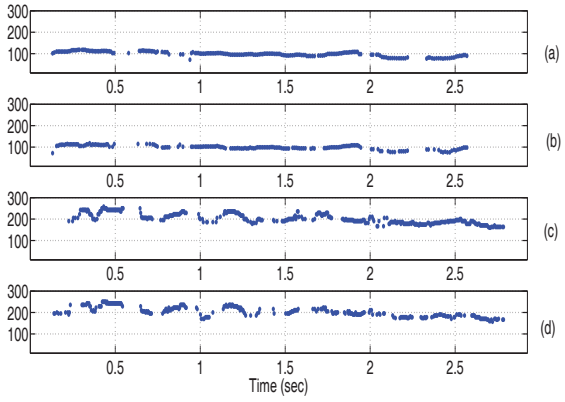

Fig. 4. Illustration of instantaneous F_0 contours derived from (a) the reference signal of Speaker-1, (b) modified delay-compensated HE $f_{s1}(n)$, (c) reference signal of Speaker-2, and (d) modified delay-compensated signal $f_{s2}(n)$.

V. Summary and Conclusions

In this paper we have analyzed and presented a method to extract the instantaneous F_0 contours from two speakers data collected at two spatially separated microphones. The method is based on extracting epoch locations from the impulse-like sequence in the speaker-specific HEs. The epochs are extracted using the filtered signal derived from the output of the 0 Hz resonators. The interval between successive epochs in the voiced regions was used to derive the instantaneous F_0 contour of each speaker. The method was demonstrated for three different cases of two speaker data.

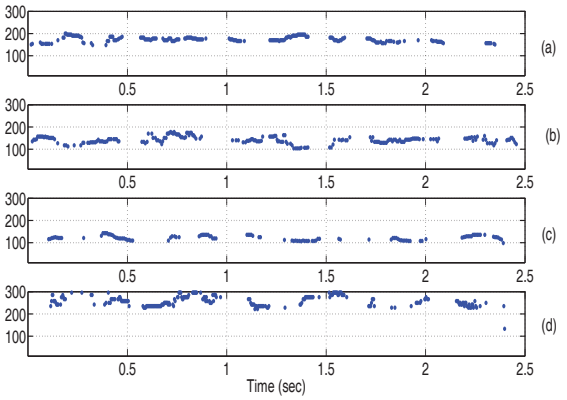

Fig. 5. Illustration of instantaneous F_0 contours derived for CASE B example ((a) and (b)) and CASE C example ((c) and (d)).

This study can be extended to examine the performance of the method when there are more than two speakers speaking simultaneously. While the method should work in principle even for 3 or more speakers, difficulty may arise in deriving the HE signal that contains peaks specific to a single speaker.

Acknowledgments

This work is a part of ongoing UKIERI project titled *Study of source features for speech synthesis and speaker recognition* among CSTR, University of Edinburgh, UK, IIT Guwahati, India and IIIT Hyderabad, India.

References

[1] B. Yegnanarayana and K. S. R. Murty, "Event-based instantaneous fundamental frequency estimation from speech signals," *IEEE Trans. Audio, Speech, Language Processing*, vol. 17(4), pp. 614–624, May 2009.
[2] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*, 1st ed. N. Delhi, India: Pearson Education, 1978.
[3] T. 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.
[4] S.R.M. Prasanna and B. Yegnanarayana, "Extraction of pitch in adverse conditions," in *Proc. IEEE Inter. Conf. on Acoustics, Speech, and Signal Processing*, vol. 1, May 2004, pp. I–109–I–112.
[5] R. Smits and B. Yegnanarayana, "Determination of instants of significant excitation in speech using group delay function," *IEEE Trans. Speech Audio Processing*, vol. 3, pp. 325–333, Sept. 1995.
[6] K. S. R. Murty and B. Yegnanarayana, "Epoch extraction from speech signals," *IEEE Trans. Audio, Speech, Language Processing*, vol. 16(8), pp. 1602–1613, Nov. 2008.
[7] B. Yegnanarayana, S. R. M. Prasanna, R. Duraiswami, and D. Zotkin, "Processing of reverberant speech for time-delay estimation," *IEEE Trans. Speech, Audio Processing*, vol. 13(6), pp. 1110–1118, Nov. 2005.
[8] B. Yegnanarayana, R. K. Swamy, and S. R. M. Prasanna, "Separation of multispeaker speech using excitation information," in *Proc. 3rd Int. Conf. Non-linear speech processing*, Apr. 2005.
[9] K. S. R. Murty, B. Yegnanarayana, and A. Joseph, "Characterization of glottal activity from speech signals," *IEEE Signal Process. Letters*, vol. 16(6), pp. 469–472, June 2009.